Chapter 5

Asymptotic Efficiency

5.1 Motivation

This Chapter is concerned with the concept of *efficiency*. To set the stage, think of the problem of estimating an unknown value θ^*. Were we to choose between two different methods, what criterion should we use? We will provide a definite answer to this question, using a mathematical theory of risk and loss functions.

As an introduction to the problem, let us consider the easiest statistical estimation problem: we seek to estimate $\theta^* = \mathbb{E}(X_n)$, where $X_n; n = 1, 2, \dots$ are iid random variables with finite variance σ^2. Let

$$\theta_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

then by the Central Limit Theorem (CLT), $\mathbb{E}(\theta_n) = \theta^*$ and $\sqrt{n}(\theta_n - \theta^*) \approx \mathcal{N}(0, \sigma^2)$. Such results, which concern the "limit" *distribution* of the sequence of "centralised" random variables (thus, the name CLT), allow us to calculate how many samples should be used to attain a given precision. For instance, if we tolerate an error of $\rho = 0.01$, then we need approximately a sample size N such that

$$z_{1-\alpha/2}\sqrt{\sigma^2/n} \approx 0.01,$$

for a confidence level $1 - \alpha$. That is, with probability at least $1 - \alpha$, choosing $N \geq (z_{1-\alpha/2}/\rho)^2$ attains (approximately) the required precision ρ.

An estimator will be more "efficient" than another one if it attains better precision with the *same number of samples*. This, of course, is the well known motivation for variance reduction in statistical estimation.

The goal when building algorithms to find an "optimal" value θ^* will be to obtain the smallest error at small "cost" N. To generalise the results for general estimation problems, and particularly to stochastic approximations, we will first provide the supporting theorems for Functional CLT's.

5.2 Functional CLT

It should be apparent from the convergence Theorems in Chapters 3 and 4 that the behaviour of the algorithms, seen as a stochastic process, "approximate" a smooth and deterministic behaviour in some limiting sense. In particular, for the constant stepsize algorithm, the interpolation process converges in distribution to a deterministic process when the assumptions of Theorem 4.2 are satisfied.

In this section we explore how the processes "hover" around the limit process. As a preliminary motivation, recall that the Law of Large Numbers (LLN) for iid zero-mean random variables states that

$$S_N = \frac{1}{N}\sum_{k=1}^{N} \xi_k \to 0, a.s.$$

and the Central Limit Theorem (CLT) establishes that, if $\sigma^2 = \mathsf{Var}(\xi_1) < \infty$, then the "standardised" variable $\sqrt{N}S_N/\sigma$ approaches a standard normal random variable, or:

$$\frac{1}{\sqrt{N}}\sum_{k=1}^{N} \xi_k \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0,\sigma^2)$$

for large N. This result is at the basis of error estimation for the approximations (eg, confidence intervals). To prepare for the extension of the CLT to the FCLT, consider the interpolation process of this same partial sum, as follows. For each fixed $\epsilon > 0$, let $m(t) = \lfloor t/\epsilon \rfloor$ and suppose that $\{\xi_k, k \in \mathbb{N}\}$ is an infinite sequence of iid zero-mean random variables. Now let's consider the interpolated process with constant step size ϵ, namely $\vartheta^\epsilon(t) = S_{m(t)}$. By construction, for any fixed t the sequence of random variables $\vartheta^\epsilon(t) \to 0$ as $\epsilon \to 0$. Call now

$$W^\epsilon(t) = \sqrt{\epsilon}\sum_{k=1}^{m(t)} \xi_k = \sqrt{\epsilon\, m(t)}\left(\frac{1}{\sqrt{m(t)}}\sum_{k=1}^{m(t)} \xi_k\right). \tag{5.1}$$

Notice that $t - \epsilon < \epsilon\, m(t) \le t$, so that as $\epsilon \to 0$ $\epsilon\, m(t) \to t$. Fix t and look at the processes $W^\epsilon(t)$ evaluated at that time t. By the usual CLT, as ϵ decreases, the random variables converge: $W^\epsilon(t) \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0,\sigma^2 t)$. We will see now that $W^\epsilon(\cdot)$ converges in distribution, as a sequence of processes, to the Wiener process or *Brownian motion* $W(\cdot)$. We now briefly recall some definitions and results.

Definition 5.1 *A process $\{W(t), t \in \mathbb{R}\}$ on $(\Omega, \{\mathfrak{F}_t\}, \mathbb{P})$ is a vector-valued* Wiener process *(or* Brownian motion*) if there is a matrix Σ, called the* covariance matrix*, such that:*

(a) $W(0) = 0, \mathbb{E}[W(t)\,|\,\mathfrak{F}_t] = 0$*, and $W(\cdot)$ has a.s. continuous paths,*

(b) $W(\cdot) \in \mathbb{R}^k$ *has* independent increments*, that is, for any set of increasing numbers $\{t_i, i \in \mathbb{N}\}$, $\{W(t_{i+1}) - W(t_i)\}$ are independent random variables,*

(c) for any $t \in \mathbb{R}$ and any $s > 0$, the distribution of the increment $W(t+s) - W(t)$ is independent of t, and

(d) for any t, $\mathbb{E}[W(t)^\top W(t)] = \Sigma t$.

When $\Sigma = \mathbb{I}$ is the identity matrix the process is referred to as "standard" Wiener process.

An equivalent definition uses the fact that the increments of a Wiener processes have zero-mean normal distribution (see [21]) in lieu of condition (c) above. We now state some results that will be useful when dealing with the FCLT for stochastic approximation methods. For details we refer to page 283 in [21].

Lemma 5.1 *Let $\{\delta M_n^\epsilon\}$ and $\{U_n^\epsilon\}$ be sequences of $\mathbb{R}^k$-valued random variables on $(\Omega, \mathfrak{F}, \mathbb{P})$ and call $\mathfrak{F}_n^\epsilon$ the minimal σ-algebra generated by $\{(U_i^\epsilon, \delta M_i^\epsilon; i \leq n\}$. For $t \geq 0$, define:*

$$W^\epsilon(t) = \sqrt{\epsilon} \sum_{i=0}^{\lfloor t/\epsilon \rfloor - 1} \delta M_i^\epsilon, \quad U^\epsilon(t) = U_n^\epsilon; \;\; t \in [n\epsilon, (n+1)\epsilon)$$

and suppose that $\mathbb{E}[\delta M_n^\epsilon \,|\, \mathfrak{F}_{n-1}^\epsilon] = 0$ w.p.1 for all n an ϵ. Finally, suppose that there exist a matrix Σ and an integer $p > 0$ such that

$$\sup_{n,\epsilon} \mathbb{E}[(\delta M_n^\epsilon)^{2+p}] < \infty, \quad \text{and } \mathbb{E}\left[\delta M_n^\epsilon (\delta M_n^\epsilon)^\top \,|\, \mathfrak{F}_{n-1}^\epsilon\right] \to \Sigma,$$

in probability, as $n \to \infty, \epsilon \to 0$. Then $W^\epsilon(\cdot)$ converges weakly to a Wiener process with covariance matrix Σ. Suppose that $(U^\epsilon(\cdot), W^\epsilon(\cdot))$ converges in distribution, in the space $D^{2k}[0,\infty)$ to a joint limit $(U(\cdot), W(\cdot))$. Then $W(\cdot)$ is a $\mathfrak{F}_t$-Wiener process, where $\mathfrak{F}_t = \sigma(U(s), W(s); s \leq t)$.

What the above lemma establishes is that the FCLT still holds when the ξ_n "error terms" are no longer assumed iid, as long as they are zero-mean noise terms, and "well-behaved".

Theorem 5.1 *Consider the algorithm (4.5):*

$$\theta_{n+1}^\epsilon = \theta_n^\epsilon + \epsilon Y_n^\epsilon$$

and assume all the conditions of Theorem 4.2 hold. Call $\vartheta(\cdot)$ the limit and assume that θ^* is the only stable point of the corresponding limit ODE. Let $\delta M_n^\epsilon = Y_n^\epsilon - g(\xi_{n-1}^\epsilon, \theta_n^\epsilon,)$ be the Martingale noise (refer to (4.6)) and assume:

(a6) There is a neighbourhood $B_\rho(\theta^*)$ of θ^* and a symmetric matrix Σ such that

$$\mathbb{E}\left[\delta M_n^\epsilon (\delta M_n^\epsilon)^\top \mathbf{1}_{\{\|\theta_n^\epsilon - \theta^*\| \leq \rho\}} \,|\, \mathfrak{F}_{n-1}^\epsilon\right] \to \Sigma.$$

(a7) The function $g(\cdot, \xi)$ admits a Taylor expansion in θ:

$$g(\theta, \xi) = g(\theta^*, \xi) + \nabla_\theta g(\theta^*, \xi)^T (\theta - \theta^*) + \rho_1(\theta, \xi),$$

where the error term satisfies: $\mathbb{E}[\rho_1(\theta, \xi_n^\epsilon)] = O(\|\theta - \theta^*\|^2)$ as $n \to \infty, \epsilon \to 0$.

(a8) There is a Hurwitz matrix A (i.e. a matrix where all the eigenvalues have a negative real part) such that

$$\lim_{m \to \infty} \frac{1}{m} \sum_{i=n}^{n+m-1} \mathbb{E}\left[\nabla_\theta g(\xi_{n-1}^\epsilon, \theta^*)^\top - A\right] = 0.$$

Define the continuous interpolation cadlag processes:

$$U_n^\epsilon = \frac{\theta_n^\epsilon - \theta^*}{\sqrt{\epsilon}}, \quad U^\epsilon(t) = U_n^\epsilon; \;\; t \in [n\epsilon, (n+1)\epsilon),$$

$$W^\epsilon(t) = \sqrt{\epsilon} \sum_{i=0}^{\lfloor t/\epsilon \rfloor - 1} \delta M_i^\epsilon.$$

Then the sequence $\{(U^\epsilon(\cdot), W^\epsilon(\cdot))\}$ converges weakly in $D^{2k}[0,\infty)$ to a limit $(U(\cdot, W(\cdot))$ satisfying:

$$dU(t) = AU(t)\,dt + dW(t), \tag{5.2}$$

and $W(t)$ is a Wiener process with covariance matrix Σ.

Proof: We will only sketch the steps of the proof, for details the reader is referred to Chapter 10 of [21]. Re-write the recursion as:

$$\theta_{n+1}^\epsilon - \theta_n^\epsilon = \epsilon g(\xi_{n-1}^\epsilon, \theta_n^\epsilon) + \epsilon \delta M_n^\epsilon.$$

The proof now proceeds in three steps: first, we "replace" the random argument θ_n^ϵ by θ^* in the function g. Second, we use the expressions to write the recursion for the sequence U_n^ϵ and express an integral form for it. Third, we characterise the limits using Lemma 5.1.

Expansion around θ^*: Using (a8), we express:

$$\theta_{n+1}^\epsilon - \theta_n^\epsilon = \epsilon\big(g(\xi_{n-1}^\epsilon, \theta^*) + \nabla_\theta g(\xi_{n-1}^\epsilon, \theta^*)^\top (\theta_n^\epsilon - \theta^*)\big) + \epsilon \delta M_n^\epsilon + \epsilon \rho_1(\xi_{n-1}^\epsilon, \theta_n^\epsilon),$$

and notice that the error term satisfies $\mathbb{E}[\rho_1(\xi_{n-1}^\epsilon, \theta)] = O(\mathbb{E}[\|\theta_n^\epsilon - \theta^*\|^2])$. We know that the interpolation process converges weakly $\vartheta^\epsilon(\cdot) \overset{\mathcal{L}}{\Longrightarrow} \vartheta(\cdot)$, where the limit process is deterministic. This implies that $\mathbb{E}\|\theta_n^\epsilon - \theta^*\|^2$ (see exercise).

Integral representation: From the definition of U_n^ϵ, we have:

$$U_{n+1}^\epsilon - U_n^\epsilon = \sqrt{\epsilon} g(\xi_{n-1}^\epsilon, \theta^*) + \sqrt{\epsilon}\delta M_n^\epsilon + \epsilon A U_n^\epsilon + \epsilon\big(\nabla_\theta g(\xi_{n-1}^\epsilon, \theta^*)^\top - A\big) U_n^\epsilon + \sqrt{\epsilon}\rho_1(\theta_n^\epsilon, \xi_n^\epsilon)$$

Adding up the terms in left and right hand sides,

$$U^\epsilon(t+s) - U^\epsilon(t) = \int_t^{t+s} A U^\epsilon(u)\, du + W^\epsilon(t) + \sqrt{\epsilon} \sum_{i=m(t)}^{m(t+s)-1} g(\theta^*, \xi_i^\epsilon) + \epsilon \sum_{i=m(t)}^{m(t+s)-1} \big(\nabla_\theta g(\theta^*, \xi_n^\epsilon)^T - A\big) U_n^\epsilon + \rho(\epsilon),$$

where $m(t) = \lfloor t/\epsilon \rfloor$ also depends on ϵ. Note that the number of terms in the sums is $m(t+s) - m(t) \approx s/\epsilon$. The cumulative error term $\rho(\epsilon)$ converges to zero in mean square error, as $\epsilon \to 0$.

Averaging: The last (and more technical) step carries out the "long term averaging" effects. Indeed, for fixed θ, from the Markovian model assumptions (a1), (a2) and (a3), the quantities $g(\theta, \xi_n)$ converge geometrically fast to their stationary expectation, that is, $G(\theta^*) = 0$. This result is used to establish that

$$\sqrt{\epsilon} \sum_{i=m(t)}^{m(t+s)-1} g(\theta^*, \xi_i^\epsilon)$$

converges to zero in (absolute-value) expectation. As well, assumption (a7) will establish that the contribution of the term

$$\epsilon \sum_{i=m(t)}^{m(t+s)-1} \big(\nabla_\theta g(\theta^*, \xi_n^\epsilon)^\top - A\big) U_n^\epsilon$$

also vanishes in the limit. Applying now Lemma 5.1, the processes $(U^\epsilon(\cdot),^\epsilon(\cdot))$ have a limit $(U(\cdot), W(\cdot))$ where the limit $W(\cdot)$ is a Wiener process with covariance Σ. Because of the averaging of the error terms, the limit must then satisfy:

$$U^\epsilon(t+s) - U^\epsilon(t) = \int_t^{t+s} A U^\epsilon(u)\, du + W^\epsilon(t),$$

which is the same as (5.2). The process $U(\cdot)$ that satisfies (5.2) is called an *Ornstein-Uhlenbeck* process.

Theorem 5.1 is important to obtain limiting estimates for confidence intervals. The stationary Ornstein-Uhlenbeck process $U(t)$ defined by $dU(t) = AU(t)dt + dW(t)$ has a normal distribution with variance V satisfying the implicit equation

$$AV + VA^\top = \Sigma. \tag{5.3}$$

This is sometimes called the continuous Lyapunov equation and common high level languages such as *Mathematica* and *Matlab* have numerical solvers for it, but we need to know A and Σ.

Let us now discuss how we can use this result for the case of a one-dimensional problem. The one-dimensional Ornstein-Uhlenbeck process is of the general form $dU(t) = -a(U(t) - \mu)\, dt + \sigma dW(t)$ (with W a standard Wiener process) and it satisfies:

$$\mathbb{E}[U(t)] = u(0)e^{-at} + \mu(1 - e^{-at}), \quad \mathsf{Cov}(U(t), U(s)) = \frac{\sigma^2}{2a} e^{-a(t+s)} \left(e^{2a(\min(s,t))} - 1 \right).$$

In our case the limit process is centered at $\mu = 0$ and the drift $a = -G'(\theta^*) > 0$ (under he stability condition (a8)), therefore a given precision can be specified for the limit ODE, by choosing t such that e^{-at} is small enough. Accordingly, the process after $N = \lfloor t/\epsilon \rfloor$ iterations will have near stationary values with stationary variance $v = \sigma^2/(2a)$.

Thus Theorem 5.1 can be used to approximate:

$$\frac{(\theta^\epsilon_N - \theta^*)}{\sqrt{\epsilon}} \approx \mathcal{N}(0, v) \implies \theta^\epsilon_N \approx \mathcal{N}(\theta^*, \epsilon\sigma^2/(2a)),$$

and this can be used to estimate the precision of the final result. Alternatively, this can be used to estimate an appropriate value of ϵ given a required precision. Exercise 5.3 illustrates this procedure.

Result: Let $X \sim \mathcal{N}(\mu, V)$ where V is the variance-covariance matrix in d dimensions. Let $Q(\alpha, \chi^2_d)$ be the α-quantile of the Chi distribution with d degrees of freedom. We know that

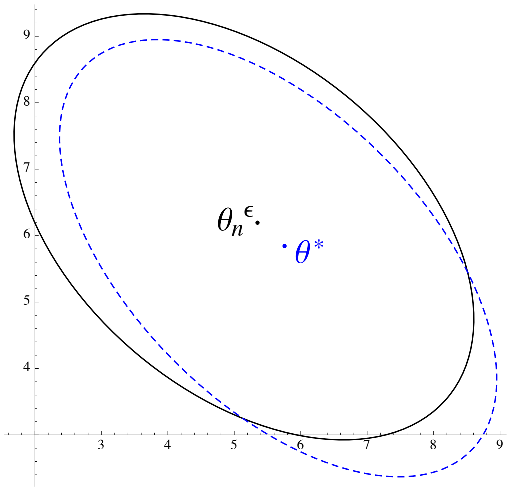
$$(X - \mu)^\top V^{-1}(X - \mu) \sim \chi^2_d.$$

Given an observation X, define the ellipse

$$\mathcal{C} = \left\{ x \in \mathbb{R}^d \colon (X - x)^\top V^{-1}(X - x) = Q(\alpha, \chi^2_d) \right\}.$$

Then by construction, $\mathbb{P}(\mu \in \bar{\mathcal{C}}) = 1 - \alpha$, where $\bar{\mathcal{C}}$ contains the interior of the ellipse.

The easiest way to construct this confidence region is as follows. Call $\rho_i, \mathbf{e}_i$ the eigenvalues and eigenvectors of V, for $i = 1, \ldots, d$. Then the axes of the ellipse are in the directions $\mathbf{e}_i$, and each of the diameters are given by $\pm\mathbf{e}_i \sqrt{\epsilon\, Q(\alpha, \chi^2_d)\lambda_i}$. Figure 5.1 illustrates the resulting confidence interval.



The plot to the left shows an approximate confidence interval centered on the estimate θ_n^ϵ. The width along each of the ellipse's axes is $\sqrt{\epsilon\, Q(\alpha, \chi_d^2)\lambda_i}$. Because this is an approximation, the actual value of θ^* is shown in blue together with the true region for which the probability is exactly $1-\alpha$.

Figure 5.1: Example of a confidence region centered on $\theta_n^\epsilon \in \mathbb{R}^2$.

Example 5.1. Suppose that we use stochastic approximation for a robot that is supposed to track the location of a target (for example in a rescue mission, or deep under the ocean). The target's location is estimated by the robot's sensors via noisy thermal, sonar or optical signals $\{\xi_n\}$, assumed iid, with mean α. The simple target tracking equation:

$$\theta_{n+1}^\epsilon = \theta_n^\epsilon - \epsilon(\theta_n^\epsilon - \xi_n)$$

will satisfy the conditions of Theorem 2.3 under regularity assumptions on the noise, for example $\mathsf{Var}(\xi_n) < \infty$. Here $A = -\mathbb{I}$ is Hurwitz, so that the asymptotic variance is $V = \frac{\sigma^2}{2}\mathbb{I}$. Estimating $\Sigma = \sigma^2\mathbb{I}$ concurrently with the stochastic approximation is straightforward using sample variance of the observations, so that at the end of a large number N of iterates one can provide a confidence interval of the form $\widehat{\alpha(i)} = \theta_N(i) \pm 1.96\sqrt{\frac{\epsilon\widehat{\sigma^2}}{2}}$, for each component $\alpha(i)$ of the target vector location. In this case independence of the noisy observations makes it simpler to estimate V using an estimate for Σ rather than estimating V directly with consecutive values of θ_n. Because each component of U_n follows a decrease rate proportional to e^{-t} then given ϵ one can stop at N such that $e^{-N\epsilon} \le \rho_1$, for a given precision ρ_1. Alternatively, pilot observations $\{\xi_n\}$ can be used to get an estimate $\widehat{\sigma^2}$ and then choose N, ϵ such that $e^{-N\epsilon} \le \rho_1$ and $\epsilon\widehat{\sigma^2} \le \rho_2$ for desired precision values ρ_1, ρ_2.

***

Example 5.2. Consider again the equilibrium problem visited in Exercise 3.6 and Example 4.5. In this example, the flows $T_i(\theta)$ are linear in θ, and

$$Y_n^\epsilon(i) = -\left(T_i(\theta_n) - \frac{1}{3}\sum_k T_k(\theta_n)\right) - \left(\gamma_i - \frac{1}{3}\sum_k \gamma_k\right),$$

where $\gamma_i \sim \Gamma(2, 0.5)$ are iid. Let $\sigma^2 = \mathsf{Var}(\gamma_i)$. Then we have:

$$g(\theta) = \mathbb{E}[Y_n^\epsilon \,|\, \mathfrak{F}_n] \equiv G(\theta) = \begin{pmatrix} T_1(\theta) - \frac{1}{3}\sum_k T_k(\theta) \\ T_2(\theta) - \frac{1}{3}\sum_k T_k(\theta) \\ T_3(\theta) - \frac{1}{3}\sum_k T_k(\theta) \end{pmatrix}$$

only depends on the noise via the dependency on θ_n. The vector field $G(\theta)$ is the one driving the limit ODE (see Example 4.5). We can calculate Σ in Theorem 5.1 as follows:

$$\mathsf{Var}\left(\gamma_i - \frac{1}{3}\sum_k \gamma_k\right) = \frac{2}{3}\sigma^2 = \frac{1}{3},$$

so $\Sigma = \frac{1}{3}\,\mathbb{I}$ is a diagonal matrix. Finally, A is independent of θ due to the linearity of T:

$$A = \nabla G(\theta) = -\frac{1}{3N}\begin{pmatrix} 1 & 1 & -2 \\ -1 & 2 & -1 \\ -2 & 1 & 1 \end{pmatrix}$$

(this matrix is calculated in Exercise 3.6 and we will omit the details here.) This matrix is not Hurwitz, because it has two negative and one null eigenvalues. For this reason, Matlab could not solve the problem directly, and using a small perturbation of A to solve (5.3) the solution to $AV + VA^T = \Sigma$ gives:

$$V = \begin{pmatrix} -1493 & -1490 & -1497 \\ -1490 & -1495 & -1495 \\ -1496 & -1495 & -1503 \end{pmatrix}.$$

The square root of the eigenvalues of V are $(212, 7, 4)$ so that the corresponding ellipse lies basically along the first eigenvector $(.576, .576, .567)$ where all components have equal weight. Use $Q(0.95, \chi^2_3) = 7.81$. The resulting approximate confidence for each component $\theta^*(i)$ region is thus:

$$\theta^\epsilon_n(i) \pm 340\,\sqrt{\epsilon}.$$

Looking back at the plots for Example 4.5, $\theta^* = (25, 50, 25)$ where we used $\epsilon = (.5, .1, .05, .01)$ it is clear that this approximate confidence interval is meaningless. Actually, using various perturbations for the matrix A we found that the eigenvectors and two of the eigenvalues were relatively stable, but the first eigenvalue (which is the dominant value for the confidence interval) increases without bound as the perturbation decreases.

Therefore, for problems like this where A is either unknown or mot Hurwitz, one must estimate the asymptotic variance V directly. A straightforward approach would be to run many replications of the algorithm in parallel to obtain an iid sample of the end point θ_N, and then use the sample variance to build the confidence interval.

⁂

To finalise the study of the convergence behaviour of stochastic approximation algorithms, we now provide the most general results available to date for decreasing step size. The following result is in [22], and we will present it without proof.

Theorem 5.2 *Consider the algorithm (3.4)*

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n$$

for the Martingale noise model. Let $G\colon \mathbb{R}^n \to \mathbb{R}^n$ *be a bounded and continuous function with a unique point* θ^* *such that* $G(\theta^*) = 0$ *and suppose that* $\mathbf{A} = \nabla G(\theta^*)$ *is a Hurwitz matrix eigenvalues* λ_1, λ_d*. Call* $\lambda_{\min} = \min(\Re(\lambda_i))$*. Define:*

$$B_n = \mathbb{E}[Y_n \,|\, \mathfrak{F}_{n-1}] - G(\theta_n) \tag{5.4}$$

$$V_n = \mathbb{E}[(Y_n - \mathbb{E}[Y_n \,|\, \mathfrak{F}_{n-1}])^2], \tag{5.5}$$

and assume that there are constants γ, β, δ *such that:*

- $\gamma + \beta > 1, 2\gamma + \delta > 1$
- $\epsilon_n = n^{-\gamma}, \|B_n\| = \mathcal{O}(n^{-\beta}), V_n = \mathcal{O}(n^{-\delta})$.

Then $\theta_n \to \theta^*$ w.p.1, and if $\gamma < 1$

$$\mathbb{E}[\|\theta_n - \theta^*\|^2] = \mathcal{O}(n^{-\kappa}), \quad \kappa \overset{\text{def}}{=} \min(2\beta, \gamma + \delta). \tag{5.6}$$

If $\gamma = 1$, then (5.6) is satisfied provided that $\lambda_{\min} > \max(\beta, (1+\delta)/2)$.

The above result gives the rate of convergence κ of the mean square error (MSE) and can be used to design the parameters of the algorithm. The following example illustrates how this result can also be used to assess the benefits of variance reduction through increasing sample sizes.

Example 5.3. In Economic Theory, the supply of products is assumed to be an increasing function of the price θ, while de demand is a decreasing function of θ. The price dictated by the market is the value θ^* such that $S(\theta^*) = D(\theta^*)$. Many models exist to describe price dynamics (Cournot models and more general models assuming delays and lags in production like in [**?**]). Here we propose the following model in order to find the price θ^* when the demand function $D(\theta)$ is known but the supply function is unknown. This scenario is in accordance to several views that (particularly for commodities that are very expensive to produce) supplies suffer from unpredictable "shocks", many times due to political and social environments in countries that produce raw materials and/or manage important assembly lines. Suppose that consecutive instances for the supply can be generated through complex simulations and historical data, so that given a price θ, $\{\xi_k(\theta)\}$ are consecutive observations (possibly correlated) of the supply, satisfying that

$$\mathbb{E}[\xi_n(\theta) \,|\, \xi_{n-1}(\theta_{n-1})] = S(\theta).$$

Assume that $\mathsf{Var}[\xi_n(\theta)] = \sigma^2(\theta)$ is a bounded function of θ. Because we assume strict monotonicity of the functions, it follows that $G(\theta) = -(S(\theta) - D(\theta))$ is strictly decreasing, that is, $D'(\theta) - S'(\theta) < 0$. Consider now using the recursion

$$\theta_{n+1} = \theta_n + \epsilon_n(D(\theta_n) - \xi_n(\theta_n)).$$

Because $G' < 0$ for all θ (uniform Hurwitz condition), G is convex and the ODE

$$\frac{dx(t)}{dt} = G(x(t))$$

has a unique limit point (asymptotically stable) $\lim_{t\to\infty} x(t) = \theta^*$. Under this scheme we have $B_n = 0$ (no bias, so $\beta = +\infty$), and $V_n = \mathsf{Var}[\xi_n(\theta_n)] = \sigma^2(\theta_n) = \mathcal{O}(1)$. We may therefore apply Theorem 3.3 to establish that the sequence $\theta_n \to \theta^*$ w.p.1., provided that ϵ_n satisfies the usual conditions (3.2).

Now suppose that we wish to use Theorem 5.2 to improve the rate of convergence by getting more accurate estimates of the supply. Specifically, assume that for each iteration, we use a number $T_n = n^\delta$ of observations of $\xi_n(\theta_n)$ to build the feedback estimator Y_n. Call $\nu(n) = \sum_{k\le n} T(k)$. Then

$$\begin{aligned} V_n = \mathbb{E}[(Y_n - G(\theta_n))^2] &= \mathbb{E}\left[\left(\left(D(\theta_n) - \frac{1}{T_n}\sum_{k=\nu(n-1)+1}^{\nu(n)} \xi_k(\theta_n)\right) - (D(\theta_n) - S(\theta_n))\right)^2\right] \\ &= \mathsf{Var}\left[\frac{1}{T_n}\sum_{k=\nu(n-1)+1}^{\nu(n)} \xi_k(\theta_n)\right] = \mathcal{O}(T_n^{-1}) = \mathcal{O}(n^{-\delta}). \end{aligned}$$

According to the result in Theorem 5.2, $\kappa = \gamma + \delta$ so in principle we can make δ as large as we wish to have "infinite" convergence rate. What's wrong with this analysis? Well, it's missing the fact that we are not really using $\|\theta_n - \theta^*\|$ for the analysis, but each iteration requires $\nu(n)$ steps, therefore what Theorem 5.2 ascertains is that

$$\mathbb{E}\|\theta_{\nu(n)} - \theta^*\| = \mathcal{O}(n^\kappa).$$

In this example, $\nu(n) = \sum_{k\le n} k^\delta$ is the generalized Harmonic number and it satisfies $n^\delta \le \nu(n) \le n^{\delta+1}$, so $\nu(n) = \mathcal{O}(n^{\delta+1})$. Thus

$$\mathbb{E}\|\theta_{\nu(n)} - \theta^*\| = \mathcal{O}(n^\kappa) = \mathcal{O}(\nu(n)^{\kappa/(\delta+1)}),$$

where $\kappa = \gamma + \delta$ is maximized at $\gamma = 1$. Therefore the actual convergence rate $\kappa' = \kappa/(\delta+1)$ will be maximized at $\kappa' = 1$ regardless of the value chosen for δ. Recall that $\delta = 0$ is the case when we use only one observation for the feedback Y_n. So we conclude that, from the point of view of asymptotic convergence rate, there is no gain in averaging the partial observations. We will come back to the analysis of computational effort and convergence rate in our next section.

***

To motivate the following result we will refer to the usual CLT in the simplest statistical estimation context. Suppose that $\{X_n\}$ is a sequence such that $X_n \to 0$, and that $\mathbb{E}(\|X_n\|^2) = \mathcal{O}(n^{-\kappa})$. In the case of partial averages where $X_n = S_n = n^{-1}\sum_{i=1}^n \xi_i$, and $\{\xi_n\}$ are zero-mean iid variables, $V_n = \mathsf{Var}[X_n] = n^{-1}\sigma^2$, so with $\delta = 1$ we have $V_n = \mathcal{O}(n^{-\delta})$. Then the standardized sequence $n^{\delta/2}X_n$ will have an approximate normal distribution with variance σ^2. These limiting results are of major importance when estimating the errors in our approximations, confidence intervals and predictions.

EXAMPLE 5.4. Suppose that a sequence of independent estimators $X_1, X_2, \ldots$ with $X_n \to 0$ w.p.1 has decreasing bias, $\mathbb{E}[X_n] = b_n \to 0$, then the limit distribution depends on how fast b_n itself decreases. Let $b_n = \mathcal{O}(n^{-\beta})$, and $\mathsf{Var}(X_n) = n^{-1}\sigma^2 = \mathcal{O}(n^{-\delta}) =$, for $\delta = 1$.

Case 1: If $\beta > \delta/2$ then the bias terms decrease *very fast* (no bias corresponds to $\beta = +\infty$) because $n^{\delta/2}b_n = \mathcal{O}(n^{\delta/2-\beta})$ and we have:

$$\mathbb{E}[n^{\delta/2}X_n] \to 0, \quad \text{and } \mathsf{Var}(n^{\delta/2}X_n) \to \sigma^2,$$

so here $\kappa = 2$ gives the convergence rate for the MSE. Therefore in this case the variance dominates the limit behaviour: the scaled sequence $n^\delta X_n$ has fluctuations around zero with limiting variance σ^2. Te most common situation is when X_n has an approximate normal distribution $\mathcal{N}(0,\sigma^2)$.

Case 2: If $\beta < \delta/2$ then $n^\beta X_n$ has a limiting bias, that is,

$$\bar{B} \overset{\text{def}}{=} \lim_{n\to\infty} n^\beta X_n$$

exists, and of course, here $\mathbb{E}(n^{\delta/2}X_n) \to \infty$. However, scaling with n^β we see that the limiting behaviour is controlled now by the bias, because $\mathsf{Var}(n^\beta X_n) \approx n^{-(\delta-2\beta)}\sigma^2 \to 0$. This is the case where the limiting distribution of the scaled sequence is degenerate, concentrated on $\bar{B}$.

Case 3: The case $\beta = \delta/2$ is a mixture of the previous cases, where the scaled sequence has both a limiting bias and a limiting variance. Under normality assumptions, we have $n^{\delta/2}X_n \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(\bar{B}, \sigma^2)$.

***

Theorem 5.3 *Consider the algorithm (3.4) for the Martingale model. Assume that:*

(A1) $\mathbb{E}(Y_n \,|\, \mathfrak{F}_{n-1}) = G(\theta_n) + b_n$, *where* $G \in \mathcal{C}^1$ *has a unique stationary point* θ^* *such that* $G(\theta^*) = 0$ *and suppose that* $\mathbb{A} = \nabla G(\theta^*)$ *is a Hurwitz matrix with eigenvalues* $\lambda_1, \dots, \lambda_d$*. Call* $\lambda_{\min} = \min(-\Re(\lambda_i))$*.*

(A2) There are constants γ, β, δ *such that* $\gamma + \beta > 1$, $2\gamma + \delta > 1$ *and* [1]

$$\epsilon_n = n^{-\gamma}, \quad \|b_n\| = \mathcal{O}(n^{-\beta}), \quad V_n \overset{\text{def}}{=} \mathbb{E}[(\delta M_n)(\delta M_n)^T] = \mathcal{O}(n^{-\delta}),$$

(A3) If $\gamma = 1$*, then suppose that* $\lambda_{\min} > \max(\beta, (1+\delta)/2)$*.*

(A4) There is a symmetric positive definite matrix Σ *such that*

$$n^\delta \mathbb{E}[(\delta M_n)(\delta M_n)^T] \to \Sigma.$$

Then calling

$$\kappa = \min(2\beta, \gamma + \delta),$$

the scaled sequence $n^{\kappa/2}(\theta_n - \theta^*)$ *converges in distribution as follows:*

$$n^{\kappa/2}(\theta_n - \theta^*) \overset{\mathcal{L}}{\Longrightarrow} \begin{cases} \mathcal{N}(0, V) & \gamma + \delta < 2\beta \\ H^{-1}\bar{B} & \gamma + \delta > 2\beta \\ \mathcal{N}(H^{-1}\bar{B}, V) & \gamma + \delta = 2\beta, \end{cases}$$

where $\bar{B} \overset{\text{def}}{=} \lim_{n\to\infty} n^\beta b_n$, $H = -\mathbb{A} - \beta\mathbb{I}$ *and* V *satisfies:*

$$V = \int_0^\infty e^{-\tilde{H}u} \Sigma e^{-\tilde{H}^T u} dt,$$

for $\tilde{H} = -\mathbb{A} - \frac{(1+\delta)}{2}\mathbb{I}$*. An equivalent expression for* V *is given by the implicit equation* $\tilde{H}V + V\tilde{H}^T = \Sigma$*.*

Example 5.5. Consider again the case of unconstrained optimisation where finite differences are used, and refer to Example 3.6, where we established that $c \in (0, .5)$ was necessary for convergence to the optimal value when the noise from different observations are independent. We will now apply Theorem 5.3 to this scheme in order to determine the "best" parameters, that is, those that provide faster convergence rate to the optimal point.

Suppose that

$$Y_n = \frac{\xi_n(\theta_n + c_n) - \xi_n(\theta_n - c_n)}{2c_n},$$

where $\mathbb{E}[\xi_n \,|\, \mathfrak{F}_{n-1}] = J(\theta_n)$, and let $c_n = n^{-c}$ for some algorithm parameter c to be determined. As in Example 3.6,

$$\mathbb{E}[Y_n \,|\, \mathfrak{F}_{n-1}] = J'(\theta_n) + \mathcal{O}(c_n^2),$$

so that $\beta = 2c$. For the calculation of the conditional variance V_n we will now consider two cases:

[1] Check the condition in blue because I had the rhs divided by 2!

Case 1: The noise processes are independent, that is $\xi_n(\theta_n + c_n)$ is independent of $\xi_n(\theta_n - c_n)$. In this case, as was calculated in Example 3.6, $V_n = \mathcal{O}(c_n^{-2}) = \mathcal{O}(n^{+2c})$. Thus in this case $\delta = -2c$.

Theorem 5.3 establishes that the convergence rate of the algorithm is $\kappa = \min(2\beta, \gamma + \delta)$, where $\epsilon_n = n^{-\gamma}$ for $\gamma \in (0, 1]$. The minimum is attained when $2\beta = \gamma + \delta$.

For Case 1, this leads to:

$$4c = \gamma - 2c, \implies c = \gamma/6,$$

so the maximum rate is achieved when $\gamma = 1$, and it is $\kappa = 2/3$. Using Taylor expansion as in Example 3.6 we calculate the limit bias:

$$\bar{B} = \lim_{n\to\infty} n^{\beta} b_n = \lim_{n\to\infty} \left(\frac{J(\theta + c_n) - J(\theta - c_n)}{2c_n} \right) = \frac{J'''(\theta^*)}{6},$$

and the limiting variance is calculated calling $\mathsf{Var}[\xi_n(\theta)] = \sigma^2(\theta)$, as follows:

$$V_n = \frac{\sigma^2(\theta_n + c_n) + \sigma^2(\theta_n - c_n)}{4c_n},$$

so that using the fact that as $\lim_{n\to\infty} \theta_n = \theta^*$ a.s.,

$$\Sigma = \lim_{n\to\infty} n^{\delta} V_n = \frac{\sigma^2(\theta^*)}{2}.$$

This is a one dimensional problem and $(1 + \delta/2) = \beta = -1/3$ so here the asymptotic variance is $V = \Sigma/(2(J''(\theta^*) + 1/3))$. With these values for the various parameters, Theorem 5.2 states that

$$n^{1/3}(\theta_n - \theta^*) \overset{\mathcal{L}}{=} \mathcal{N}\left(\frac{\bar{B}}{J''(\theta^*) + 1/3}, \frac{\sigma^2(\theta^*)}{4(J''(\theta^*) + 1/3)} \right).$$

It is worth emphasizing that there is an asymptotic bias for the scaled error.

Case 2: Consider now the particular case where $\xi_n(\omega, \theta) = \xi(\omega; \theta)$ is continuous and differentiable in θ for every given ω. Assume that $\sup_\theta \mathsf{E}[\xi'(\theta)^2] < \infty$. Using *common random numbers* we assume here that the two observations are correlated, that is, we use the same "noise", yielding:

$$\xi_n(\omega; \theta_n + c_n) - \xi_n(\omega; \theta_n - c_n) = \xi_n'(\omega; \theta_n) + \kappa_n(\omega),$$

where κ_n is a bounded random variable, and $\mathbb{E}[\kappa_n] = \mathcal{O}(c_n^2)$, which now gives $V_n = \mathcal{O}(1)$ and $\delta = 0$. This yields (equating $2\beta = \gamma + \delta$)

$$4c = \gamma \implies c = \gamma/4,$$

so the maximum rate is now at $c = 1/4$, and it is $\kappa = 1$, which is the same convergence rate as if we had an unbiased estimator for the derivative. The calculation of the asymptotic normal distribution is left to the reader.

***

5.3 Estimating Confidence Intervals

Perhaps here we should summarize general techniques for giving the final answer. Distinguish between constant and decreasing step size, and perhaps also mention the open problems for adaptive step sizes.

Much work would be required for this, although I think it is worth including.

5.4 Asymptotic Efficiency

This section presents the main theory introduced in [14], and we refer to that reference for al proofs. Let $\{\theta_n\}$ be a stochastic process on the probability space $(\Omega, \{\mathfrak{F}_n\}, \mathbb{P})$, and suppose that the sequence $\{\theta_n\}$ is used to estimate an unknown parameter θ^*. Call $C(n)$ the computational time required to calculate $\theta_1, \ldots, \theta_n$, which is a $\mathfrak{F}_n$ measurable random variable. We will be mostly concerned with the *Loss function*:

$$L(\theta) = \|\theta - \theta^*\|^2,$$

representing the square error at value θ. However, other real valued convex functions $L(\cdot)$ can be used as well. Define:

$$N(c) = \sup(n \in \mathbb{N} \colon C(n) \le c) \tag{5.7}$$

as the (random) number of estimates that can be calculated given a *computational budget* c. The corresponding final estimator is $\theta_{N(c)}$.

Definition 5.2 *Given a computational budget c, the* Risk function *is:*

$$R(c) = \mathbb{E}[L(\theta_{N(c)})]. \tag{5.8}$$

The Risk function measures the expected mean square error of the final estimator as a function of the computational budget and it is often impossible to evaluate analytically.

Definition 5.3 *Suppose that there exists real numbers* $0 < v < \infty$ *and* $0 < \mathcal{E} < \infty$ *such that*

$$\lim_{c\to\infty} c^v\, R(c) = \frac{1}{\mathcal{E}}. \tag{5.9}$$

Then we call $\mathcal{E}$ *the* asymptotic efficiency *of the estimation, and the number* v *its corresponding* asymptotic convergence rate*.*

EXAMPLE 5.6. For the simple estimation problem of a sample average of one-dimensional iid observations with variance σ^2, $\theta_n \approx \mathcal{N}(\theta^*, \sigma^2)$ for large n and

$$R(c) = \mathbb{E}[(\theta_{N(c)} - \theta^*)^2] \approx \frac{\sigma^2}{N(c)}.$$

If each sample has constant cost λ units (of CPU time, or sampling efforts, or money) then $C(n) = \lambda\, n$, and $N(c) = \lfloor c/\lambda \rfloor$, which implies:

$$\lim_{c\to\infty} \frac{N(c)}{c} = \lambda,$$

therefore, in this case, called the *canonical* estimation case, $v = 1$, because $c^{-1}R(c) \to \sigma^2\,\lambda$, yielding

$$\mathcal{E} = \frac{1}{\sigma^2\,\lambda}.$$

The basic principle is that the (asymptotic) efficiency of the canonical estimator is inversely proportional to the product of variance and computational effort per sample. This measure of efficiency agrees with the notion of "loss" incurred in estimation, and allows a quantitative comparison between possible estimation or approximation methods.

⁂

Theorem 5.4 *Assume the following conditions:*

(a1) There is an integer $\kappa \in \mathbb{N}$ such that for each $\epsilon > 0$,

$$U^\epsilon \stackrel{\text{def}}{=} \epsilon^{-\kappa/2}(\theta_{\lfloor t/\epsilon \rfloor}) - \theta^*)$$

converges in distribution as $\epsilon \to 0$ to a random element $U \in \mathcal{D}$, the set of real-valued cadlag processes with the Skorokhod topology[2], and assume that $U(\cdot)$ is a.s. continuous.

(a2) There exist $\tau > 0, \text{CPU} > 0$ such that w.p.1

$$\lim_{n\to\infty} n^{-\tau} C(n) = \text{CPU}^{\,\tau}.$$

Then

(i) The following result (FCLT) holds $c^{-\kappa/2\tau}(\theta_{N(ct)} - \theta^*) \xrightarrow{\mathcal{L}} U(t^{1/\tau}/\text{CPU})$ as $c \to \infty$,

(ii) The associated CLT result holds, namely: $c^{-\kappa/2\tau}(\theta_{N(c)} - \theta^*) \xrightarrow{\mathcal{L}} U(1/\lambda) = \text{CPU}^{\,\kappa/2} U(1)$

(iii) The corresponding Weak LLN is satisfied: $\theta(N(c)) \to \theta^*$ in probability, as $c \to \infty$.

Lemma 5.2 *Consider a general loss function $L(\cdot) \in \mathcal{C}^2$, convex, non negative and satisfying $L(\theta^*) = L'(\theta^*) = 0; L''(\theta^*) > 0$. Under the conditions of Theorem 5.4, suppose that $\{c^{\kappa/\tau} L(\theta_{N(c)}; c > 1\}$ is uniformly integrable. Then the asymptotic convergence rate and efficiency are given by:*

$$v = \frac{\kappa}{\tau}, \qquad \mathcal{E} = \frac{2}{L''(\theta^*)}\left(\frac{1}{\text{CPU}^{\,\kappa}\mathbb{E}(U(1)U(1)^\top)}\right).$$

In particular, when $L(\theta) = (\theta - \theta^*)^2$, $L''(\theta^*) = 2$, and the efficiency is again the inverse of a asymptotic variance times the asymptotic cost per sample. This criterion of efficiency provides a way to trade off precision and speed of numerical solutions.

EXAMPLE 5.7. Let $J\colon \mathbb{R} \to \mathbb{R}, J \in \mathcal{C}^2$ be a convex cost function that we wish to minimise, call θ^* the unique minimum and suppose that $J''(\theta^*) > 1$. Assume that there is a random variable $Z(\theta)$ such that $\mathbb{E}[Z(\theta)] = J'(\theta)$, with $\sup_\theta \mathsf{Var}(Z(\theta)) < \infty$, and that there are random variables Y_n such that

$$\mathbb{P}(Y_n \in B \,|\, \mathfrak{F}_{n-1}) = \mathbb{P}(Z(\theta_n) \in B).$$

Use a stochastic approximation procedure of the form:

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n,$$

where $\epsilon_n = n^{-\gamma}$. Using Theorem 5.3, here $\delta = 0$, and $\beta = +\infty$ so that

$$n^{\gamma/2}(\theta_n - \theta^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V),$$

where $2aV = \Sigma$, $a = J''(\theta^*) - 1/2, \Sigma = \mathsf{Var} Z(\theta^*)$. For the one dimensional case this yields:

$$V = \frac{\mathsf{Var}(Z(\theta^*))}{2J''(\theta^*) - 1}.$$

[2] For most of the cases that we are interested in, the limit process will be a Wiener or an Orstein-Ülenbeck process.

Assume that the computational cost for the n-th sample is a $\mathfrak{F}_{n-1}$-measurable random variable t_n with conditional distribution

$$\mathbb{P}(t_n \leq t \,|\, \mathfrak{F}_{n-1}) = F_{\theta_n}(t).$$

Let $t(\theta) \sim F_\theta(\cdot)$ and assume that $\sup_{\theta \in \mathbb{R}} \mathsf{Var}(t(\theta)) < \infty$. Assume that $\mathbb{E}[t(\theta)]$ is continuous in the neighborhood of θ^*. By definition, the computational cost up to iteration n is:

$$C(n) = \sum_{i=1}^{n} t_i.$$

Using a.s. convergence of $\theta_n \to \theta^*$, continuity of $\mathbb{E}[t(\theta)]$ boundedness of the variance, we obtain:

$$n^{-1} C(n) = \frac{1}{n} \sum_{i=1}^{n} t_i \to \mathbb{E}[t(\theta^*)] \overset{\text{def}}{=} \text{CPU} \quad w.p.1,$$

so that in this example, $\tau = 1$. Using Lemma 5.2 with loss function $L(\theta) = (\theta - \theta^*)^2$, the Martingale noise decreasing step size stochastic approximation with no bias has asymptotic rate and efficiency given by:

$$v = \frac{\gamma}{2}, \quad \mathcal{E} = \frac{1}{\text{CPU}^{\;2\gamma}\, V}$$

which supports the known result that the "best" decreasing step size rate is obtained setting $\gamma = 1$.

***

EXAMPLE 5.8. Work out the example for the mollifiers in our paper for the probability constraints.

***

5.5 Exercises

EXERCISE 5.1. Show that the real-valued processes $\{W_n(\cdot); n \in \mathbb{N}\}$ in (5.1) are tight in $D^k[0, \infty)$, using the fact that $W_n(t+s) - W_n(t) = (1/\sqrt{n}) \sum_{q(t)}^{q(t+s)-1} \xi_n$ and $\mathsf{Var}(\xi_n) = \sigma^2 < \infty$. Use the compactness argument to show that the limit process is a Wiener process. What is Σ here?

EXERCISE 5.2. Usually, weak convergence does not imply other types of convergence. This exercise deals with a particular case where the limit of a random sequence is deterministic. Let $\{X_n\}$ be a sequence of random numbers that converges in distribution to a fixed number $a \in \mathbb{R}$. Show that $X_n \to a$ in mean square error and in probability. Is it also true that $X_n \to a$ a.s.?

EXERCISE 5.3. Consider the supply/demand problem of Example 5.3. The demand function $D(\theta) = \theta^{-d}, d > 0$ is known. However the supply function $S(\theta)$ is only known to be an increasing function of θ that is analytic (infinitely continuously differentiable). Instead, a complex simulation model is used to produce statistically independent unbiased estimates ξ_n such that for any continuous and bounded function f

$$\mathbb{E}[f(\xi_n) \,|\, \theta_n] = f(S(\theta_n)); \quad \mathsf{Var}[\xi_n] = 1. \tag{5.10}$$

(a) Show that

$$\theta_{n+1} = \theta_n + \epsilon\, Y_n, \quad Y_n = D(\theta_n) - \xi_n$$

satisfies the assumptions of Theorem 5.1.

(b) Use $d = 5$ for the demand function. Your economics guru has estimated that $\theta^* \approx 1$ and $S'(\theta^*) \approx 4.5$. With this information, apply Theorem 5.1 to identify the values of a, σ^2 for the (approximate) limit Orstein Uhlenbeck process $U(t)$, and find T such that $e^{-aT} \approx 0.0001$.

(c) Show that $\epsilon \approx 0.0005$ yields a precision of 0.01 (half width of the approximate confidence interval after T/ϵ iterations, with confidence level $\alpha = 0.05$).

(d) In this part of the problem you will generate the random observations $\{\xi_n\}$ and run the stochastic approximation. Conditional on θ_n, let ξ_n ∼LN(m, v^2) have a lognormal distribution. First find the parameters for the m and v such that (5.10) holds, with $S(\theta) = \theta^s, s = 4.3$. Next, run the algorithm and discuss your results.

EXERCISE 5.4. Use Theorem 3.2 to prove the statement in Theorem 5.2 that θ_n converges almost surely to θ^*.

EXERCISE 5.5. Show Lemma 5.2. Use a Taylor expansion of $L(\cdot)$ around θ^* to show first that

$$\lim_{c\to\infty} c^{\kappa/\beta} R(c) = (1/2) L''(\theta^*) \lambda^\kappa \mathbb{E}[U(1)^2].$$

EXERCISE 5.6. An investor wishes to divide her capital in two assets $\{S_i(t), i = 1, 2\}$, which she will access at maturity time T. Assume that her total capital is 1 (by changing monetary units if necessary). Although not known precisely, the corresponding means (μ_1, μ_2) and variances (σ_1^2 and σ_2^2) satisfy $\mu_1 > \mu_2$ and $\sigma_1^2 >> \sigma_2^2$. In order to balance her profit and risk, the investor wishes to find the proportion θ that solves the problem:

$$\max_\theta \quad \mathbb{E}[X(\theta)]$$
$$\text{s.t. } \mathbb{E}[X(\theta)^2] < B,$$

where $\mathbb{E}[S_1(T)^2] > B > \mathbb{E}[S_2(T)^2]$, where $X(\theta) = \theta S_1(T) + (1-\theta)S_2(T)$. Although the exact distributions of the two assets are not known, it is possible to use historical observations or simulations to produce consecutive samples $\xi_n = (\xi_{n,1}, \xi_{n,2}) \overset{\mathcal{L}}{=} (S_1(T), S_2(T))$.

(a) Argue that the optimal value θ^* of the above problem must satisfy $0 < \theta^* < 1$, and that the constraint will be active at the optimum.

(b) For a particular value $x = (x_1, x_2)$, let $\phi(x, \theta) = -\theta x_1 - (1-\theta) x_2$ so that $J(\theta) = \mathbb{E}[\phi(S_1(T), S_2(T); \theta)]$ is the function that we wish to minimise. Write the Langrangian of the problem and show that it is a convex NLP, so that the Arrow-Hurwicz algorithm for the deterministic problem converges to the optimal solution. Specify the vector field $G(\theta, \lambda)$ $(\lambda > 0)$ for the corresponding limit ODE.

(c) Consider the stochastic version of the Arrow-Hurwicz algorithm using finite differences: let $\xi = (\xi_n(1), \xi_n(2)) \overset{\mathcal{L}}{=} \xi' = (\xi_n'(1), \xi_n'(2))$ and suppose that these samples are statistically independent, that is, $\xi_n \perp \xi_n'$ for all n. The algorithm is:

$$\begin{aligned} \theta_{n+1} &= \theta_n - \frac{\epsilon_n}{2c_n}\Big(\phi(\xi_n, \theta_n + c_n) - \phi(\xi_n', \theta_n - c_n) + \lambda_n \left(\phi^2(\xi_n, \theta_n) - B\right)\Big) \\ \lambda_{n+1} &= \big(\lambda_n + \epsilon_n(\phi^2(\xi_n, \theta_n) - B\big)_+ \end{aligned} \tag{5.11}$$

with $\epsilon_n = \mathcal{O}(n^{-\gamma}), c_n = \mathcal{O}(n^{-c})$. Use Theorem 5.2 to establish that the fastest convergence is achieved at $c = 1/6$ and gives $\kappa = 2/3 < 1$.

(d) Let $\Delta_n \overset{\text{def}}{=} \xi_{n,1} - \xi_{n,2}$. Show that $J'(\theta) = -\mathbb{E}[\Delta_n]$ and that $g'(\theta) = 2\mathbb{E}[\theta\Delta_n^2 + \xi_{n,2}\Delta_n]$. Use Theorem 3.3 to show that the stochastic Arrow-Hurwitz algorithm

$$\begin{aligned} \theta_{n+1} &= \theta_n - \epsilon_n \left(-\Delta_n + 2\lambda_n(\Delta_n^2 + \xi_{n,2}\Delta_n)\right) \\ \lambda_{n+1} &= \left(\lambda_n + \epsilon_n(\theta_n\Delta_n + \xi_{n,2}^2)\right)_+ \end{aligned}$$

converges to the solution of the optimization problem θ^*. What do you need to assume on the step size sequence $\{\epsilon_n\}$? Use Theorem 5.2 to find the convergence rate κ assuming that $\epsilon = n^{-\gamma}$. Specify the values of β and δ. Next, find the asymptotic efficiency and the corresponding rate, using Theorem 5.4.

(e) Suppose that instead of using one observation ξ_n, each iteration uses the running average $\bar{\xi_n} = \frac{1}{n}\sum_{k=1}^n \xi_k$. Repeat the tasks in (d) above.