Thursday 11/17/05

Simulation & Learning

Recall: Last time, Prediction Error Method.

$$Y(n+1) = [H(\theta_0, z^{-1}) U](n) + N(n+1)$$

$$\hat{Y}(n+1|n) = [H(\hat{\theta}(n), z^{-1}) U](n), \quad \text{given estimate.}$$

Given $\hat{\theta} = \theta$ consider MSE. $\Gamma(\theta) = \frac{1}{2} E\left[\left(Y(n+1) - \hat{Y}(n+1|n) \right)^2 \right]$.

Goal: Solve $\boxed{\nabla \Gamma(\theta) = - E\left[(Y(n+1) - \hat{Y}(n+1|n)) \nabla_\theta \hat{Y}(n+1|n) \right] = 0}$

Usual Gradient Algorithm

$$\hat{\theta}(k+1) = \hat{\theta}(k) + a \nabla \Gamma(\hat{\theta}(k)) \qquad \text{Not computable}$$

Stochastic Gradient Algorithm

$$\hat{\theta}(k+1) = \hat{\theta}(k) + a(k) \left[(Y(k+1) - \hat{Y}(k+1|k)) \phi(k) \right]$$

$$\phi(k) = \nabla_\theta \hat{Y}(k+1|k) \Big|_{\theta = \hat{\theta}(k)} = \left[\nabla_\theta H(\theta, z^{-1}) U \right](k) \Big|_{\theta = \hat{\theta}(k)}.$$

Special case of stochastic Approximation

- Robbins + Monro
- Hirsch '89
- Benveniste et. al. '90
- Kushner + Yin '97

$\boxed{\text{Borkar + Meyn '00}}$

General set-up: $\theta^* \in \mathbb{R}^d$ unknown parameter

$f : \mathbb{R}^{d+m} \to \mathbb{R}^d$, $N \in \mathbb{R}^m$ random vector

$$E[f(\theta, N)] = 0 \quad \text{when} \quad \theta = \theta^*.$$

Examples 1) Simulation $E[f(N) - \theta] = 0$, $\theta^* = \eta = \pi(f)$.

2) Optimization $f(\theta, N) = \nabla\Gamma(\theta, N)$

3) Fixed point equations Recall ACOE,

$$\min_u [c(x,u) + P_u h^*(x)] = h^*(x) + \eta^*, \quad x \in \mathsf{X}.$$

DCOE,

$$\min_u [c(x,u) + P_u h_\gamma^*(x)] = (1+\gamma) h_\gamma^*(x)$$

Define $Q(x,u) = (1+\gamma)^{-1} [c(x,u) + P_u h_\gamma^*(x)]$

$$\therefore \quad \min_u Q(x,u) = h_\gamma^*(x)$$

$$\therefore \quad Q(x,u) = (1+\gamma)^{-1} \Big[c(x,u) + \sum_y P_u(x,y) \big[\min_{u'} Q(y,u')\big] \Big]$$

(concave in Q)

Here $\Theta = \{Q(x,u) : x \in \mathsf{X},\ u \in \mathsf{U}(x)\}$, $N = \{N(x,u) : \cdots\}$

$$f(\theta, N) = -Q(x,u) + (1+\gamma)^{-1} [c(x,u) + \min_{u'} Q(N, u')]$$

where $N(x,u) \sim X(k+1)$ when $X(k) = x$, $U(k) = u$.

$\longrightarrow$ SA $\equiv$ Q-learning

General S.A. algorithm

$$\Theta(k+1) = \Theta(k) + a_k f(\Theta(k), N(k+1)), \quad k \geq 0.$$

Restrict to $\{N(k)\}$ iid.

Standard form: $g(\theta) = E[f(\theta, N)]$

$$\Theta(k+1) = \Theta(k) + a_k [g(\Theta(k)) + \Delta(k+1)]$$

$$\Delta(k+1) = f(\Theta(k), N(k+1)) - E[f(\Theta(k), N(k+1)) \mid \underbrace{\Theta_0^k, N_0^k}_{\mathcal{F}_k}]$$

Two canonical settings: Constant step-size $a_k \equiv a$

Tapering/Vanishing step-size

$$\boxed{\sum a_k = \infty, \quad \sum a_k^2 < \infty}$$

Assumptions

(A1) $g: \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous,

$$\| g(x^1) - g(x^2) \| \leq b_g \| x^1 - x^2 \| \quad \text{all } x^1, x^2 \in \mathbb{R}^d.$$

(A2) $E[\| \Delta(n+1) \|^2 \mid \mathcal{F}_n] \leq \sigma_\Delta^2 (1 + \| \Theta(n) \|^2), \quad n \geq 0.$

Issues: Stability & Convergence

Approach: ODE method

$$\dot{x} = g(x) \qquad \text{and other "fluid models".}$$

Simplest example is simulation: Monte-Carlo,

$$\theta(n+1) = \frac{1}{n+1} \sum_0^n f(N(k))$$

$$= \frac{1}{n+1} \{ n\,\theta(n) + f(N(n)) \}$$

$$= \theta(n) + \frac{1}{n+1} \big(f(N(n)) - \theta(n) \big)$$

More generally, take any sequence $\{a(n) \ge 0\}$ satisfying

$$\sum_0^\infty a_n = \infty, \quad \sum_0^\infty a_n^2 < \infty.$$

$$\theta(n+1) = \theta(n) + a_n [f(N(n)) - \theta(n)]$$

If $\{N\}$ is a nice Markov chain, or iid, then

$$\theta(n) \to \theta^* = \pi(f), \quad n \to \infty \quad \text{a.s.}$$

Fixed step-size algorithm: $\theta(n+1) = \theta(n) + a [f(N(n)) - \theta(n)]$

$$= (1-a)\,\theta(n) + a f(N(n))$$

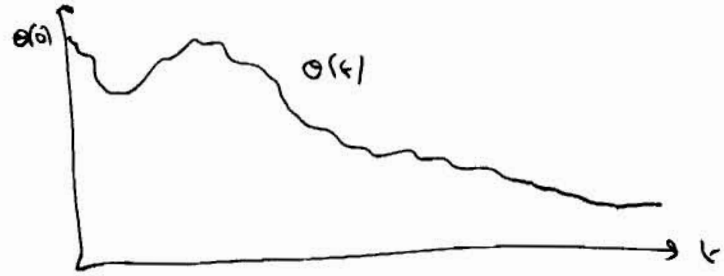
$$= (1-a)^{n+1}\theta(0) + \sum_{i=0}^n a(1-a)^{n-i} f(N(i)).$$

$$\boxed{E[\|\theta(n) - \theta^*\|^2] \le B_1(a) + B_2[1 + \|\theta(0)\|^2] e^{-\varepsilon(a) n}}$$

Variance $O(a)$

This can be generalized to general algorithm.

Stability Considerations



Let $r \geq 1$, $\theta^r(k) = \frac{1}{r}\theta(k)$, with $\theta(0) = rx \in \mathbb{R}^d$.

$$\theta(k+1) = \theta(k) + a_k\left[g(\theta(k)) + \Delta(k+1)\right]$$

$$\theta^r(k+1) = \theta^r(k) + a_k\left[\tfrac{1}{r}g(\theta(k)) + \tfrac{1}{r}\Delta(k+1)\right]$$

Define $g_r(x) = \frac{1}{r}g(rx)$, $x \in \mathbb{R}^d$.

$$\theta^r(k+1) = \theta^r(k) + a_k\left[g_r(\theta^r(k)) + \tfrac{1}{r}\Delta(k+1)\right].$$

Motivates O.D.E. $\boxed{\dot{x} = g_r(x)}$ and $\boxed{\dot{x} = g_\infty(x)}$

provided $g_r(x) \to g_\infty(x)$ pointwise.

<u>Proposition</u> Suppose the origin is locally asymptotically stable for $\dot{x} = g_\infty(x)$. Then it is <u>globally exponentially</u> asymptotically stable

<u>Proof</u> : g_∞ is homogeneous,

$$g_\infty(tx) = \lim_{r\to\infty} \frac{1}{r} g(rtx) = t\, g_\infty(x).$$

Let $\varepsilon > 0$ s.t. $x(t) \to 0$ uniformly for $x(0) \in \overline{B(\varepsilon)} = \{x : \|x\| \leq \varepsilon\}$.

Thus, we can find $T > 0$ s.t. $\|x(T)\| \leq \varepsilon/2$ whenever $\|x(0)\| \leq \varepsilon$.

Consider any solution, and write $\bar{x}(t) = \varepsilon \frac{x(t)}{\|x(0)\|}$.

$\Rightarrow \|\bar{x}(T)\| \leq \varepsilon/2 \quad \Rightarrow \quad \|x(T)\| \leq \frac{1}{2}\|x(0)\|$

[Note: can also project]

Constant step-size algorithm : main ideas from Borkar + Meyn

$\theta(t)$ is a Markov chain.

If g_∞ defines stable vector field then for large r, small a,

$$E\left[\|\theta^r(T)\|^2\right] \leq \tfrac{3}{4}\|\theta^r(0)\|^2 \quad \text{when } \|\theta^r(0)\| \geq 1.$$

Proved by relating stability of g_∞ to stability of g_r ...

$$\Rightarrow \quad E\left[\|\theta(T)\|^2\right] \leq \tfrac{3}{4}\|\theta(0)\|^2 \quad \text{when } \|\theta(0)\| \geq r.$$

Lyapunov drift condition. Provided a density condition holds,

Proposition There exists $a_0 > 0$ s.t. for all $0 < a \leq a_0$

$$E\left[\|\theta(n) - \theta^*\|^2\right] \leq B_1(a) + B_2\left[1 + \|x\|^2\right] e^{-\varepsilon_0(a) n}.$$

where $B_1(a) \to 0$, $\varepsilon_0(a) \to 0$, $a \to 0$.

$$B_1(a) = E_\pi\left[\|\theta(0) - \theta^*\|^2\right].$$

"Mean - Variance trade-off"

Vanishing step-size algorithm

Preliminaries: $\{\Delta(k) : k \geq 1\}$ is a martingale-difference sequence,

$$E[\Delta(k+1) \mid \mathcal{F}_k] = 0, \qquad \underset{\uparrow \atop (A2)}{E[\|\Delta(k+1)\|^2 \mid \mathcal{F}_k] \leq \sigma_\Delta^2 (1 + \|\theta(k)\|^2)}$$

(*) $$\theta(k+1) = \theta(0) + \sum_{i=0}^{k} a_i g(\theta(i)) + M(k+1),$$

where $M(\cdot)$ is a martingale

$$M(k+1) = \sum_{i=0}^{k} a_i \Delta(i+1), \qquad E[M(k+1) \mid \mathcal{F}_k] = M(k).$$

$$E[\|M(k+1)\|^2] = \sum_{i=0}^{k} a_i^2 E[\|\Delta(i+1)\|^2]$$ bounded, if $E[\|\Delta(i)\|^2]$ is bounded.

Under (A2): require $E[\|\theta(k)\|^2]$ bounded.

Boundedness is established under stability of O.D.E

$$\dot{x} = g_\infty(x).$$

Approach to convergence: (*) looks like a discrete approximation to the solution to the ODE

$$\dot{x} = g(x)$$

which we assume has unique a. stable equilibrium θ^*.
Also, Assume: $\theta(k) \in H$ compact for all $k \geq 0$.

<u>Time scale</u>: $t(n) = \sum_{i=0}^{n-1} a(i) \to \infty$, as $n \to \infty$.

Fix $T > 0$ and define $T(0) = 0$,

$$T(n+1) = \min\left(t(j) : t(j) > T(n) + T\right), \quad n \geq 0.$$

We have $T(n+1) - T(n) \geq T$ for each n, and

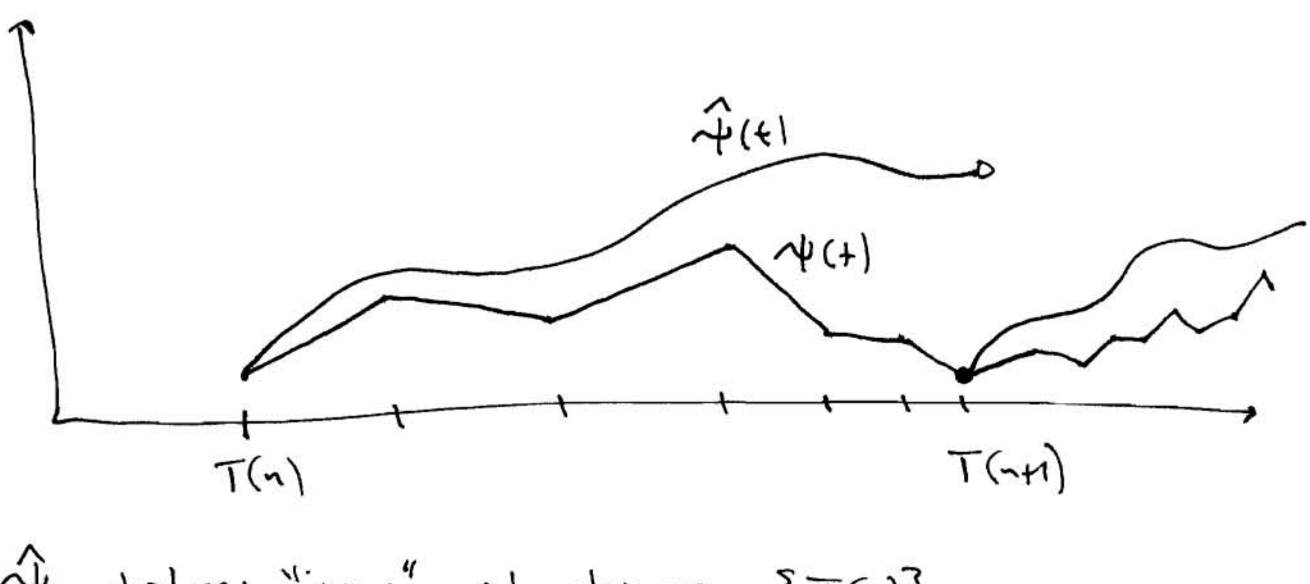
$$T(n+1) - T(n) \longrightarrow T, \quad n \to \infty.$$

<u>Two processes to be compared</u>: Both defined for $t \in \mathbb{R}_+$,

$\psi(t)$: piecewise linear, $\psi(t(n)) = \theta(n)$.

$\hat{\psi}(t)$: piecewise continuous, on each interval $[T(n), T(n+1))$, it is the solution to the ODE

$$\frac{d}{dt}\hat{\psi}(t) = g(\hat{\psi}(t)); \quad \hat{\psi}(T(n)) = \psi(T(n))$$



$\hat{\psi}$ takes "jump" at times $\{T(n)\}$.

For comparison:

(i) Fix $\varepsilon > 0$, and let $B(\varepsilon)$ denote open ball of radius ε, centered at θ^*

(ii) Find $0 < \delta < \varepsilon$ such that $x(t) \in B(\varepsilon)$ for $t \geq 0$ when $x(0) \in B(\delta)$ (for ODE!)

(iii) Find $T > 0$ so large that $x(t) \in B(\delta/2)$ for $t \geq T$ when $x(0) \in H$.

<u>Note</u>: Under (iii) we have $\hat{\varphi}(T(n)-) \in B(\delta/2)$ for each $n \geq 1$.

Next: we bound $\|\varphi(t) - \hat{\varphi}(t)\|$ for $t \geq 0$.

This uses Lipschitz continuity of g, and

<u>Bellman - Gronwall Lemma</u> Suppose $\{A(t) : 0 \leq t \leq T\}$ is non-negative, and satisfies

$$A(t) \leq A(0) + b\int_0^t A(s)\, ds, \qquad 0 \leq s \leq T.$$

Then,

$$A(t) \leq A(0) e^{bt}, \qquad 0 \leq s \leq T.$$

We have for each $t(k) > T(n)$,

$$\psi(t(k)) = \psi(T(n)) + \sum_j a_j g(\psi(t(j))) + M(t(k)) - M(T(n))$$
$$\{j : T(n) \le t(j) < t(k)\}.$$

With some work,

$$\psi(t) = \psi(T(n)) + \int_{T(n)}^{t} g(\psi(t))\,dt + \mathcal{E}(T(n), t)$$

such that $\lim_{n\to\infty}\left(\sup_{T(n)\le t\le T(n+1)} \|\mathcal{E}(T(n),t)\|\right) = 0.$

Also, by definition,

$$\hat{\psi}(t) = \psi(T(n)) + \int_{T(n)}^{t} g(\hat{\psi}(t))\,dt, \quad T(n) \le t < T(n+1).$$

So, with b_0 equal to the Lipschitz constant,

$$\|\psi(t) - \hat{\psi}(t)\| \le b_0 \int_{T(n)}^{t} \|\psi(s) - \hat{\psi}(s)\|\,ds + e(n)$$
$$T(n) \le t < T(n+1),$$

where $e(n) = \sup_{T(n)\le t< T(n+1)} \|\mathcal{E}(T(n), t)\|$.

To place this in the form of the B.G. Lemma define,

$$A(t) = \max\left(\|\psi(t) - \hat{\psi}(t)\|, e(n)\right),$$
$$T(n) \le t < T(n+1).$$

$$A(t) \le \max\left\{ b_0 \int_{T(n)}^{t} \| \psi(s) - \hat{\psi}(s) \| \, ds + e(n), \; e(n) \right\}$$

$$\le b_0 \int_{T(n)}^{t} A(s) \, ds + \underset{\equiv A(0)}{\underset{\uparrow}{e(n)}}$$

$$\therefore \quad \max\left(\| \psi(t) - \hat{\psi}(t) \|, e(n) \right) \le e(n) \, e^{b_0 (t - T(n))}$$

$$T(n) \le t < T(n+1)$$

So, $\displaystyle\sup_{T(n) \le t < T(n+1)} \| \psi(t) - \hat{\psi}(t) \| \longrightarrow 0, \quad n \to \infty.$

In particular, $\psi(T(n)) \in B(\delta)$ for all large n.

$\Longrightarrow \hat{\psi}(t) \in B(\varepsilon)$ for all large t

$\Longrightarrow \displaystyle\limsup_{t \to \infty} \| \psi(t) - \theta^* \| \le \varepsilon.$

Since $\varepsilon > 0$ is arbitrary, this shows that $\theta(n) \to \theta^*$ provided $\{\theta(n)\}$ is bounded.

ECE 555 Control of Stochastic Systems Fall 2005

Handout: ***Reinforcement learning***

In this handout we analyse reinforcement learning algorithms for Markov decision processes. The reader is referred to [2, 10] for a general background of the subject and to other references listed below for further details. This handout is based on [5].

Stochastic approximation In lecture on November 29th we considered the general stochastic approximation recursion,

$$\theta(n+1) = \theta(k) + a_n[g(\theta(n)) + \Delta(n+1)], \qquad n \geq 0,\ \theta(0) \in \mathbb{R}^d. \tag{1}$$

Here we provide a summary of the main results from [5].

Associated with the recursion (1) are two O.D.E.s,

$$\frac{d}{dt}x(t) = g(x(t)) \tag{2}$$

$$\frac{d}{dt}x(t) = g_\infty(x(t)), \tag{3}$$

where $g_\infty : \mathbb{R}^d \to \mathbb{R}^d$ is the scaled function, $\lim_{r\to\infty} r^{-1}g(rx) = g_\infty(x)$, $x \in \mathbb{R}^d$. We assumed in lecture that this limit exists, along with some additional properties,

(A1) The function g is Lipschitz, and the limit $g_\infty(x)$ exists for each $x \in \mathbb{R}^d$. Furthermore, the origin in $\mathbb{R}^d$ is an asymptotically stable equilibrium for the O.D.E. (3).

(A2) The sequence $\{\Delta(n) : n \geq 1\}$ is a martingale difference sequence with respect to $\mathcal{F}_n = \sigma(\theta(i), \Delta(i), i \leq n)$. Moreover, for some $\sigma_\Delta^2 < \infty$ and any initial condition $\theta(0) \in \mathbb{R}^d$,

$$\mathsf{E}[\|\Delta(n+1)\|^2 \mid \mathcal{F}_n] \leq \sigma_\Delta^2(1 + \|\theta(n)\|^2), \qquad n \geq 0.$$

The sequence $\{a_n\}$ is deterministic and is assumed to satisfy one of the following two assumptions. Here TS stands for 'tapering stepsize' and BS for 'bounded stepsize'.

(TS) The sequence $\{a_n\}$ satisfies $0 < a_n \leq 1,\ n \geq 0$, and

$$\sum_n a_n = \infty, \quad \sum_n a_n^2 < \infty.$$

(BS) The sequence $\{a_n\}$ is constant: $a_n \equiv a > 0$ for all n.

Stability of the O.D.E. (3) implies stability of the algorithm:

Theorem 1 *Assume that (A1), (A2) hold. Then, for any initial condition $\theta(0) \in \mathbb{R}^d$,*

(i) *Under (TS),* $\sup_n \|\theta(n)\| < \infty$ *a.s..*

(ii) *Under (BS) there exists $a_0 > 0$, $b_0 < \infty$, such that for any fixed $a \in (0, a_0]$,*

$$\limsup_{n\to\infty} \mathsf{E}[\|\theta(n)\|^2] \leq b_0.$$

□

For the TS model we have convergence when the O.D.E. (2) has a stable equilibrium point:

Theorem 2 *Suppose that (A1), (A2), (TS) hold and that the O.D.E. (2) has a unique globally asymptotically stable equilibrium θ^*. Then $\theta(n) \to \theta^*$ a.s. as $n \to \infty$ for any initial condition $\theta(0) \in \mathbb{R}^d$.*

We can also obtain bounds for the fixed stepsize algorithm. Let $\boldsymbol{e}$ denote the error sequence,

$$e(n) = \|\theta(n) - \theta^*\|, \qquad n \geq 0.$$

Theorem 3 *Assume that (A1), (A2) and (BS) hold, and suppose that (2) has a globally asymptotically stable equilibrium point θ^*. Then, for $a \in (0, a_0]$, and for every initial condition $\theta(0) \in \mathbb{R}^d$,*

(i) *For any $\varepsilon > 0$, there exists $b_1 = b_1(\varepsilon) < \infty$ such that*

$$\limsup_{n\to\infty} \mathsf{P}(e(n) \geq \varepsilon) \leq b_1 a.$$

(ii) *If θ^* is a globally exponentially asymptotically stable equilibrium for the O.D.E. (2), then there exists $b_2 < \infty$ such that,*

$$\limsup_{n\to\infty} \mathsf{E}[e(n)^2] \leq b_2 a.$$

□

Suppose that the increments of the model take the form,

$$g(\theta(n)) + \Delta(n+1) = f(\theta(n), N(n+1)), \qquad n \geq 0, \tag{4}$$

where $\boldsymbol{N}$ is an i.i.d. sequence on $\mathbb{R}^q$. In this case, for the BS model, the stochastic process $\boldsymbol{\theta}$ is a (time-homogeneous) Markov chain. Assumptions (5) and (6) below are required to establish ψ-irreducibility:

There exists a $n^* \in \mathbb{R}^q$ with $f(\theta^*, n^*) = 0$, and a continuous density $p : \mathbb{R}^q \to \mathbb{R}_+$ satisfying $p(n^*) > 0$ and

$$\mathsf{P}(N(1) \in A) \geq \int_A p(z)dz, \qquad A \in \mathcal{B}(\mathbb{R}^q); \tag{5}$$

The pair of matrices (A, B) is controllable with

$$A = \frac{\partial}{\partial x} f(\theta^*, n^*) \quad \textit{and} \quad B = \frac{\partial}{\partial n} f(\theta^*, n^*), \tag{6}$$

Under Assumptions (5) and (6) there exists a neighborhood $B(\epsilon)$ of θ^* that is *small* in the sense that there exists a probability measure ν on $\mathbb{R}^d$ and $\delta > 0$ such that

$$P^d(x, A) := \mathsf{P}\{\theta(r) \in A \mid \theta(0) = x\} \geq \delta\nu(A), \qquad x \in B(\epsilon)$$

Stability of the O.D.E. (2) can be used to show that the resolvent satisfies,

$$R(x, B(\epsilon)) := \sum_{k=0}^{\infty} 2^{-k-1} P^k(x, B(\epsilon)) > 0, \qquad x \in \mathbb{R}^d,$$

which is equivalent to ψ-irreducibility [9].

Theorem 4 *Suppose that (A1), (A2), (5), and (6) hold for the Markov model satisfying (4) with $a \in (0, a_0]$. Then we have the following bounds:*

(i) *There exist positive-valued functions A_0 and ε_0 of a, and a constant A_1 independent of a, such that*

$$\mathsf{P}\{e(n) \geq \varepsilon \mid \theta(0) = x\} \leq A_0(a) + A_1(\|x\|^2 + 1)\exp(-\varepsilon_0(a)n), \qquad n \geq 0,\ a \in (0, a_0].$$

The functions satisfy $A_0(a) \leq b_1 a$ and $\varepsilon_0(a) \to 0$ as $a \downarrow 0$.

(ii) *If in addition the O.D.E. (2) is exponentially asymptotically stable, then the stronger bound holds,*

$$\mathsf{E}[e(n)^2 \mid \theta(0) = x] \leq B_0(a) + B_1(\|x\|^2 + 1)\exp(-\epsilon_0(a)n), \qquad n \geq 0,\ a \in (0, a_0],$$

where $B_0(a) \leq b_2 a$, $\varepsilon_0(a) \to 0$ as $a \downarrow 0$, and B_1 is independent of a.

Markov decision processes We now review general theory for Markov decision processes. It is assumed that the state process $\boldsymbol{X} = \{X(t) : t \in \mathbb{Z}_+\}$ takes values in a finite state space $\mathsf{X} = \{1, 2, \cdots, s\}$, and the control sequence $\boldsymbol{U} = \{U(t) : t \in \mathbb{Z}_+\}$ takes values in a finite action space $\mathsf{U} = \{u_0, \cdots, u_r\}$. The controlled transition probabilities are denoted $P_u(i,j)$ for $i, j \in \mathsf{X}, u \in \mathsf{U}$. We are most interested in stationary policies of the form $U(t) = \phi(X(t))$, where the *feedback law* ϕ is a function $\phi \colon \mathsf{X} \to \mathsf{U}$.

Let $c : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ be the one-step cost function, and consider first the infinite horizon discounted cost control problem of minimizing over all admissible $\boldsymbol{U}$ the total discounted cost

$$h_U(i) = \mathsf{E}\Big[\sum_{t=0}^{\infty}(1+\gamma)^{-t-1}c(X(t), U(t)) \mid X(0) = i\Big],$$

where $\gamma \in (0, \infty)$ is the discount factor. The minimal value function is defined as

$$h^*(i) = \min_U h_U(i),$$

where the minimum is over all admissible control sequences $\boldsymbol{U}$. The function h^* satisfies the dynamic programming equation

$$(1+\gamma)h^*(i) = \min_u \Big[c(i,u) + \sum_j P_u(i,j)h^*(j)\Big], \qquad i \in \mathsf{X},$$

and the optimal control minimizing h is given as the stationary policy defined through the feedback law ϕ^* given as any solution to

$$\phi^*(i) := \arg\min_u \Big[c(i,u) + \sum_j P_u(i,j)h^*(j)\Big], \qquad i \in \mathsf{X}.$$

The *value iteration algorithm* is an iterative procedure to compute the minimal value function. Given an initial function $h_0 \colon \mathsf{X} \to \mathbb{R}_+$ one obtains a sequence of functions $\{h_n\}$ through the recursion

$$h_{n+1}(i) = (1+\gamma)^{-1} \min_u \Big[c(i,u) + \sum_j P_u(i,j)h_n(j)\Big], \qquad i \in \mathsf{X},\ n \geq 0. \tag{7}$$

This recursion is convergent for any initialization $h_0 \geq 0$.

The value iteration algorithm is initialized with a function $h_0 \colon \mathsf{X} \to \mathbb{R}_+$. In contrast, the *policy iteration algorithm* is initialized with a feedback law ϕ^0, and generates a sequence of feedback laws $\{\phi^n : n \geq 0\}$. At the nth stage of the algorithm a feedback law ϕ^n is given, and the value function h_n is computed. Interpreted as a column vector in $\mathbb{R}^s$, the vector h_n satisfies the equation

$$((1+\gamma)I - P_n)h_n = c_n \tag{8}$$

where the $s \times s$ matrix P_n is defined by $P_n(i,j) = P_{\phi^n(i)}(i,j)$, $i,j \in \mathsf{X}$, and the column vector c_n is given by $c_n(i) = c(i, \phi^n(i))$, $i \in \mathsf{X}$. Given h_n, the next feedback law ϕ^{n+1} is then computed via

$$\phi^{n+1}(i) = \arg\min_u \Big[c(i,u) + \sum_j P_u(i,j)h_n(j)\Big], \qquad i \in \mathsf{X}. \tag{9}$$

Each step of the policy iteration algorithm is computationally intensive for large state spaces since the computation of h_n requires the inversion of the $s \times s$ matrix $(1+\gamma)I - P_n$ to solve (8). For each n, this can be solved using the 'fixed-policy' version of value iteration,

$$V_{N+1}(i) = (1+\gamma)^{-1}[P_n V_N(i) + c_n], \qquad i \in \mathsf{X},\ N \geq 0, \tag{10}$$

where $V_0 \in \mathbb{R}^s$ is given as an initial condition. Then $V_N \to h_n$, the solution to (8), at a geometric rate as $N \to \infty$.

In the average cost optimization problem one seeks to minimize over all admissible $\boldsymbol{U}$,

$$\eta_U(x) := \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathsf{E}_x[c(X(t), U(t))]. \tag{11}$$

The policy iteration and value iteration algorithms to solve this optimization problem remain unchanged with a few exceptions. One is that the constant γ must be set equal to zero in equations (7) and (10). Secondly, in the policy iteration algorithm the value function h_n is replaced by a solution to Poisson's equation

$$P_n h_n = h_n - c_n + \eta_n, \tag{12}$$

where η_n is the steady state cost under the policy ϕ^n. The computation of h_n and η_n again involves matrix inversions via

$$\pi_n(I - P_n + ee') = e', \quad \eta_n = \pi_n c_n, \quad (I - P_n + ee')h_n = c_n,$$

where $e \in \mathbb{R}^s$ is the column vector consisting of all ones, and the row vector π_n is the invariant probability for P_n. The introduction of the outer product ensures that the matrix $(I - P_n + ee')$ is invertible, provided that the invariant probability π_n is unique.

$\boldsymbol{Q}$-learning If we define *Q-values* via

$$Q^*(i,u) = c(i,u) + \sum_j P_u(i,j)h^*(j), \qquad i \in \mathsf{X}, u \in \mathsf{U}, \tag{13}$$

then $h^*(i) = \min_u Q^*(i,u)$ and the matrix Q^* satisfies

$$Q^*(i,u) = c(i,u) + (1+\gamma)^{-1} \sum_j P_u(i,j) \min_v Q^*(j,v), \qquad i \in \mathsf{X}, u \in \mathsf{U}.$$

The matrix Q^* can be computed using the equivalent formulation of value iteration,

$$Q_{n+1}(i,u) = c(i,u) + (1+\gamma)^{-1} \sum_j P_u(i,j)\big(\min_v Q_n(j,v)\big), \qquad i \in \mathsf{X}, u \in \mathsf{U}, n \geq 0, \tag{14}$$

where $Q_0 \geq 0$ is arbitrary.

If transition probabilities are unknown so that value iteration is not directly applicable, one may apply a stochastic approximation variant known as the *Q-learning algorithm* of Watkins [11, 12]. This is defined through the recursion

$$Q_{n+1}(i,u) = Q_n(i,u) + a_n\Big[(1+\gamma)^{-1} \min_v Q_n(\Xi_{n+1}(i,u),v) + c(i,u) - Q_n(i,u)\Big], \qquad i \in \mathsf{X}, u \in \mathsf{U},$$

where $\Xi_{n+1}(i,u)$ is an independently simulated X-valued random variable with law $P_u(i,\cdot)$.

Making the appropriate correspondences with the stochastic approximation theory surrounding (1), we have $\theta(n) = Q_n \in \mathbb{R}^{s\times(r+1)}$ and the function $g\colon \mathbb{R}^{s\times(r+1)} \to \mathbb{R}^{s\times(r+1)}$ is defined as follows. Define $F\colon \mathbb{R}^{s\times(r+1)} \to \mathbb{R}^{s\times(r+1)}$ as $F(Q) = [F_{iu}(Q)]_{i,u}$ via,

$$F_{iu}(Q) = (1+\gamma)^{-1} \sum_j P_u(i,j) \min_v Q(j,v) \;+\; c(i,u).$$

Then $g(Q) = F(Q) - Q$ and the associated O.D.E. is

$$\tfrac{d}{dt}Q = F(Q) - Q := g(Q). \tag{15}$$

The map $F : \mathbb{R}^{s\times(r+1)} \to \mathbb{R}^{s\times(r+1)}$ is a contraction w.r.t. the max norm $\|\cdot\|_\infty$,

$$\|F(Q^1) - F(Q^2)\|_\infty \leq (1+\gamma)^{-1}\|Q^1 - Q^2\|_\infty, \qquad Q^1, Q^2 \in \mathbb{R}^{s\times(r+1)}.$$

Consequently, one can show that with $\widetilde{Q} = Q - Q^*$,

$$\tfrac{d}{dt}\|\widetilde{Q}\|_\infty \leq -\gamma(1+\gamma)^{-1}\|\widetilde{Q}\|_\infty,$$

which establishes global asymptotic stability of its unique equilibrium point θ^* [7]. Assumption (A1) holds, with the (i,u)-th component of $g_\infty(Q)$ given by

$$(1+\gamma)^{-1} \sum_j P_u(i,j) \min_v Q(j,v) - Q(i,u), \qquad i \in \mathsf{X}, u \in \mathsf{U}.$$

This also is of the form $g_\infty(Q) = F_\infty(Q) - Q$ where $F_\infty(\,\cdot\,)$ is an $\|\cdot\|_\infty$- contraction, and thus the origin is asymptotically stable for the O.D.E. (3).

We conclude that Theorems 1–4 hold for the Q-learning model.

Adaptive critic algorithm Next we consider the *adaptive critic algorithm*, which may be considered as the reinforcement learning analog of policy iteration. There are several variants of this, one of which, taken from [8], is as follows. The algorithm generates a sequence of approximations to h^* denoted $\{h_n : n \geq 0\}$, interpreted as a sequence of s-dimensional vectors. Simultaneously, it generates a sequence of randomized policies denoted $\{\phi^n\}$.

At each time n the following random variables are constructed independently of the past:

(i) For each $i \in \mathsf{X}$, $\Omega_n(i)$ is a U-valued random variable independently simulated with law $\phi^n(i)$;

(ii) For each $i \in \mathsf{X}$, $u \in \mathsf{U}$, $\Xi_n^a(i,u)$ and $\Xi_n^b(i,u)$ are independent X-valued random variables with law $P_u(i,\,\cdot\,)$.

For $1 \le \ell \le r$ we let e^ℓ is the unit r-vector in the ℓ-th coordinate direction. We let $\Gamma(\,\cdot\,)$ denote the projection onto the simplex $\{x \in \mathbb{R}^r_+ : \sum_i x_i \le 1\}$.

For $i \in \mathsf{X}$ the algorithm is defined by the pair of equations,

$$\begin{aligned} h_{n+1}(i) &= h_n(i) + b_n\big[(1+\gamma)^{-1}[c(i,\Omega_n(i)) + h_n(\Xi_n^a(i,\Omega_n(i)))] - h_n(i)\big], & (16) \\ \widehat{\phi}^{n+1}(i) &= \Gamma\Big\{\widehat{\phi}^n(i) + a_n \sum_{\ell=1}^r \Big([c(i,u_0) + h_n(\Xi_n^b(i,u_0))] - [c(i,u_\ell) + h_n(\Xi_n^b(i,u_\ell))]\Big)\mathsf{e}^\ell\Big\}. & (17) \end{aligned}$$

For each i, n, $\phi^n(i) = \phi^n(i,\,\cdot\,)$ is a probability vector on U defined in terms of $\widehat{\phi}^n(i) = [\widehat{\phi}^n(i,1),\dots,\widehat{\phi}^n(i,r)]$ as follows,

$$\phi^n(i,u_\ell) = \begin{cases} \widehat{\phi}^n(i,\ell) & \ell \ne 0; \\ 1 - \sum_{j\ne 0}\widehat{\phi}^n(i,j) & \ell = 0. \end{cases}$$

This is an example of a *two time-scale* algorithm: The sequences $\{a_n\}, \{b_n\}$ are assumed to satisfy

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 0,$$

as well as the usual conditions for vanishing gain algorithms,

$$\sum_n a_n = \sum_n b_n = \infty, \quad \sum_n (a_n^2 + b_n^2) < \infty.$$

To see why this is based on policy iteration, recall that policy iteration alternates between two steps: One step solves the linear system of equation (8) to compute the fixed-policy value function corresponding to the current policy. We have seen that solving (8) can be accomplished by performing the fixed-policy version of value iteration given in (10). The first step (16) in the above iteration is indeed the 'learning' or 'simulation-based stochastic approximation' analog of this fixed-policy value iteration. The second step in policy iteration updates the current policy by performing an appropriate minimization. The second iteration (17) is a particular search algorithm for computing this minimum over the simplex of probability measures on U.

The different choices of stepsize schedules for the two iterations (16), (17) induces the 'two time-scale' effect discussed in [6]. Thus the first iteration sees the policy computed by the second as nearly static, thus justifying viewing it as a fixed-policy iteration. In turn, the second sees the first as almost equilibrated, justifying the search sheme for minimization over U.

The boundedness of $\{\widehat{\phi}^n\}$ is guaranteed by the projection $\Gamma(\,\cdot\,)$. For $\{h_n\}$, the fact that $b_n = o(a_n)$ allows one to treat $\widehat{\phi}^n(i)$ as constant, say $\bar{\phi}(i)$ [8]. The appropriate O.D.E. then turns out to be

$$\tfrac{d}{dt}x = F(x) - x := g(x) \tag{18}$$

where $F : \mathbb{R}^s \to \mathbb{R}^s$ is defined by:

$$F_i(x) = (1+\gamma)^{-1} \sum_\ell \bar{\phi}(i,u_\ell)\Big[\sum_j P_{u_\ell}(i,j)x_j + c(i,u_\ell)\Big], \qquad i \in \mathsf{X}.$$

Once again, $F(\,\cdot\,)$ is an $\|\cdot\|_\infty$-contraction and it follows that (18) is globally asymptotically stable. The limiting function $g_\infty(x)$ is again of the form $g_\infty(x) = F_\infty(x) - x$ with $F_\infty(x)$ defined so that its i-th component is

$$(1+\gamma)^{-1}\sum_\ell \bar\phi(i,u_\ell)\sum_j P_{u_\ell}(i,j)x_j.$$

We see that F_∞ is also a $\|\cdot\|_\infty$- contraction and the global asymptoyic stability of the origin for the corresponding limiting O.D.E. follows [7].

Average cost optimal control For the average cost control problem we impose the additional restriction that the chain $\boldsymbol{X}$ has a *unique* invariant probability measure under any stationary policy so that the steady state cost (11) is independent of the initial condition.

For the average cost optimal control problem the Q-learning algorithm is given by the recursion

$$Q_{n+1}(i,u) = Q_n(i,u) + a_n\Big(\min_v Q_n(\Xi_n^a(i,u),v) + c(i,u) - Q_n(i,u) - Q_n(i_0,u_0)\Big),$$

where $i_0 \in \mathsf{X}$, $a_0 \in \mathsf{U}$ are fixed a-priori. The appropriate O.D.E. now is (15) with $F(\,\cdot\,)$ redefined as $F_{iu}(Q) = \sum_j P_u(i,j)\min_v Q(j,v) + c(i,u) - Q(i_0,u_0)$. The global asymptotic stability for the unique equilibrium point for this O.D.E. has been established in [1]. Once again this fits our framework with $g_\infty(x) = F_\infty(x) - x$ for F_∞ defined the same way as F, except for the terms $c(\cdot,\cdot)$ which are dropped. We conclude that (A1) and (A2) are satisfied for this version of the Q-learning algorithm.

In [8], three variants of the adaptive critic algorithm for the average cost problem are discussed, differing only in the $\{\widehat{\phi}^n\}$ iteration. The iteration for $\{h_n\}$ is common to all and is given by

$$h_{n+1}(i) = h_n(i) + b_n[c(i,\Omega_n(i)) + h_n(\Xi_n^a(i,\Omega_n,(i))) - h_n(i) - h_n(i_0)], \qquad i \in \mathsf{X}$$

where $i_0 \in \mathsf{X}$ is a prescribed fixed state. This leads to the O.D.E. (18) with F redefined as

$$F_i(x) = \sum_\ell \bar\phi(i,u_\ell)\Big(\sum_j p_{u_\ell}(i,j)x_j + c(i,u_\ell)\Big) - x_{i_0}, \qquad i \in \mathsf{X}.$$

The global asymptotic stability of the unique equilibrium point of this O.D.E. has been established in [3, 4]. Once more, this fits our framework with $g_\infty(x) = F_\infty(x) - x$ for F_∞ defined just like F, but without the $c(\cdot,\cdot)$ terms.

References

[1] J. Abounadi, D. Bertsekas, and V. S. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM J. Control Optim.*, 40(3):681–698 (electronic), 2001.

[2] D.P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming.* Atena Scientific, Cambridge, Mass, 1996.

[3] V. S. Borkar. Recursive self-tuning control of finite Markov chains. *Appl. Math. (Warsaw)*, 24(2):169–188, 1996.

[4] V. S. Borkar. Correction to: "Recursive self-tuning control of finite Markov chains". *Appl. Math. (Warsaw)*, 24(3):355, 1997.

[5] V. S. Borkar and S. P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2):447–469, 2000. (also presented at the *IEEE CDC*, December, 1998).

[6] Vivek S. Borkar. Stochastic approximation with two time scales. *Systems Control Lett.*, 29(5):291–294, 1997.

[7] Vivek S. Borkar and K. Soumyanath. An analog scheme for fixed point computation. I. Theory. *IEEE Trans. Circuits Systems I Fund. Theory Appl.*, 44(4):351–355, 1997.

[8] Vijaymohan R. Konda and Vivek S. Borkar. Actor-critic-type learning algorithms for Markov decision processes. *SIAM J. Control Optim.*, 38(1):94–123 (electronic), 1999.

[9] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993. online: `http://black.csl.uiuc.edu/~meyn/pages/book.html`.

[10] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press (and on-line, http://www.cs.ualberta.ca/%7Esutton/book/ebook/the-book.html), 1998.

[11] J.N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

[12] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

ECE 555 Control of Stochastic Systems Fall 2005

Handout: ***More on reinforcement learning: Value function approximation***

In this handout we introduce methods to approximate the value function for a given policy for application reinforcement learning algorithms for Markov decision processes. The reader is again referred to [2, 8] for a general background. More detailed treatments of temporal difference (TD) learning and value function approximation can be found in [9, 7, 6, 1].

Throughout this handout we let $\boldsymbol{X}$ denote a Markov chain without control on a state space X with transition matrix P, and unique invariant distribution π. A cost function $c\colon \mathsf{X} \to \mathbb{R}$ is given, and our goal is to estimate the solution to the DP equation,

$$Ph^* - (1+\gamma)h^* + c = 0. \tag{1}$$

We restrict to the finite state space case with $\mathsf{X} = \{1, \dots, s\}$ to simplify notation.

The issues addressed in this handout are summarized in the following remark from [5] in discussing practical issues in the implementation of MDP methods:

> A large state space presents two major issues. The most obvious one is the storage problem, as it becomes impractical to store the value function (or optimal action) explicitly for each state. The other is the generalization problem, assuming that limited experience does not provide sufficient data for each and every state.

These issues are each addressed by constructing an approximate solution to (1) over a parameterized set of functions.

Linear approximations Suppose that $\{\psi_i : 1 \le i \le q\}$ are functions on X. We seek a best fit among a set of parameterized approximations,

$$h^r(x) := r^{\mathrm{T}}\psi(x) = \sum_{i=1}^{q} r_i \psi_i(x), \qquad x \in \mathsf{X}.$$

To choose r we first define a particular metric to describe the distance between h^r and h^*. There are many ways to do this - an L_2 setting leads to an elegant solution. In this finite state space setting we view h^r and h^* as column vectors in $\mathbb{R}^s$. For a given $s \times s$ matrix M we define the L_2 error $\|h^r - h^*\|_M^2 = (h^r - h^*)^{\mathrm{T}} M (h^r - h^*)$. We will focus on the special case $M = \mathrm{diag}(\pi(1), \dots, \pi(s))$ so that the L_2 error can be expressed,

$$\|h^r - h^*\|_M^2 = \mathsf{E}_\pi[(h^r(X(k)) - h^*(X(k)))^2] = \mathsf{E}_\pi[(r^{\mathrm{T}}\psi(X(k)) - h^*(X(k)))^2].$$

The derivative with respect to r has the probabilistic interpretation,

$$\nabla_r \|h^r - h^*\|_M^2 = 2\mathsf{E}_\pi[(r^{\mathrm{T}}\psi(X(k)) - h^*(X(k)))\psi(X(k))], \tag{2}$$

and setting this equal to zero gives the optimal value,

$$r^* = A^{-1}b, \qquad \text{where } A = \mathsf{E}_\pi[\psi(X)\psi(X)^{\mathrm{T}}], \quad b = \mathsf{E}_\pi[h^*(X)\psi(X)].$$

We assume henceforth that $A > 0$.

The steepest descent algorithm to compute r^* is given by,

$$\tfrac{d}{dr} r(t) = -a\nabla_r \|h^r - h^*\|^2_M = -a[Ar + b], \qquad t \geq 0, \tag{3}$$

where $a > 0$ is a gain. Although this leads to a natural stochastic approximation algorithm, the function h^* appearing in the definition of b is not known. Given the representation $h^* = R_\gamma c := [(1+\gamma)I - P]^{-1}c$, we could resort to the pair of O.D.E.s,

$$\begin{aligned} &\tfrac{d}{dt} r = -Ar + \pi(h\psi) \\ &\tfrac{d}{dt} h = Ph - (1+\gamma)h + c \end{aligned} \tag{4}$$

This is exponentially asymptotically stable when $M > 0$. Since $M = \mathrm{diag}(\pi)$, this amounts to irreducibility of $\boldsymbol{X}$. The following S.A. recursion follows naturally

$$\begin{aligned} r(k+1) - r(k) &= a_k \sum_{i=1}^{s} \mathbb{I}\{X(k) = i\}[h(i;k) - \psi^{\mathrm{T}}(i) r(k)]\psi(i) \\ h(i;k+1) - h(i;k) &= a_k \mathbb{I}\{X(k) = i\}[h(X(k+1);k) - (1+\gamma)h(i;k) + c(i)], \qquad i \in \mathsf{X}, \end{aligned} \tag{5}$$

where $\{a_k\}$ is a vanishing gain sequence. The corresponding ODE is almost (4) except that the h equation is modified,

$$\tfrac{d}{dt} h(i;t) = \pi(i)[Ph\,(t;i) - (1+\gamma)h(t;i) + c(i)], \qquad i \in \mathsf{X}.$$

Since this evolves autonomously and is linear, analysis of the two coupled ODEs is straightforward.

The algorithm (5) may remain too complex for application in large problems. Observe that it is necessary to maintain estimates of $h^*(i)$ for each $i \in \mathsf{X}$, which means that the memory requirements are linear in the size of X. A simple remedy can be found through a closer look at the derivative equation (2).

L_2 **theory** The right hand side of (2) can be written, $\frac{d}{dr}\|h^r - h^*\|^2_M = 2\pi(fg)$, with $f = h^r - h^*$ and $g = \psi$. The resolvent $R_\gamma c$ will be transformed in the representation $h^* = R_\gamma c$ using some duality theory.

Consider the Hilbert space $L_2(\pi)$ consisting of real-valued functions on X whose second moment under π is finite. This simply means the function is finite-valued in the finite state space case. For $f, g \in L_2(\pi)$ we define the inner product,

$$\langle f, g\rangle = \pi(fg).$$

The *adjoint* $\widetilde{R}_\gamma$ of the resolvent is characterized by the defining set of equations,

$$\langle R_\gamma f, g\rangle = \langle f, \widetilde{R}_\gamma g\rangle, \qquad f, g \in L_2(\pi).$$

To construct the adjoint we obtain a sample path representation for $\langle R_\gamma f, g\rangle$. Let $\boldsymbol{X}$ denote a stationary version of the Markov chain on the two sided interval. We have,

$$\langle R_\gamma f, g\rangle = \mathsf{E}\Big[\Big(\sum_{t=0}^{\infty} (1+\gamma)^{-t-1} P^t f\,(X(0))\Big) g(X(0))\Big]$$

We have by the smoothing property of the conditional expectation,

$$\mathsf{E}[P^t f\,(X(0)) g(X(0))] = \mathsf{E}\big[\mathsf{E}[f(X(t)) \mid X(0)] g(X(0))\big] = \mathsf{E}[f(X(t)) g(X(0))]$$

and then applying stationarity of $\boldsymbol{X}$ and the smoothing property once more,

$$\mathsf{E}[P^t f\,(X(0))g(X(0))] = \mathsf{E}[f(X(0))g(X(-t))] = \sum \pi(x)f(x)\mathsf{E}[g(X(-t)) \mid X(0)=x].$$

Consequently,

$$\langle R_\gamma f, g\rangle = \sum_{t=0}^{\infty}(1+\gamma)^{-t-1}\mathsf{E}[f(X(0))g(X(-t))] = \langle f, \widetilde{R}_\gamma g\rangle,$$

where the adjoint is expressed,

$$\widetilde{R}_\gamma g\,(x) = \sum_{t=0}^{\infty}(1+\gamma)^{-t-1}\mathsf{E}[g(X(-t)) \mid X(0)=x]. \tag{6}$$

Applying the adjoint equation to the definition of b given below (2) gives,

$$b = \mathsf{E}_\pi[h^*(X(k))\psi(X(k))] = \mathsf{E}_\pi[R_\gamma c\,(X(k))\psi(X(k))] = \mathsf{E}_\pi[c(X(k))\widetilde{R}_\gamma\,\psi(X(k))]$$

Based on (6) we obtain,

$$b = \sum_{t=0}^{\infty}(1+\gamma)^{-t-1}\mathsf{E}_\pi[c(X(k))\psi(X(k-t))]. \tag{7}$$

This final representation (7) is the basis of TD learning.

Temporal difference learning Returning to (2) we have,

$$\nabla_r\|h^r - h^*\|_M^2 = 2\langle h^r - h^*, \psi\rangle = 2\langle h^r - R_\gamma c, \psi\rangle$$

and writing $h^r - R_\gamma c = R_\gamma[(1+\gamma)h^r - Ph^r - c]$ we obtain from the adjoint equation,

$$\nabla_r\|h^r - h^*\|_M^2 = 2\big\langle(1+\gamma)h^r - Ph^r - c, \widetilde{R}_\gamma\psi\big\rangle \tag{8}$$

Written as an expectation we obtain

$$\nabla_r\|h^r - h^*\|_M^2 = 2\mathsf{E}\big[[(1+\gamma)h^r(X(k)) - h^r(X(k+1)) - c(X(k))][\widetilde{R}_\gamma\psi\,(X(k))]\big] \tag{9}$$

We now have sufficient motivation to construct the TD learning algorithm based on the O.D.E. (3). The algorithm constructs recursively a sequence of estimates $\{r(k)\}$ based on the following,

(i) The *temporal differences* in TD learning are defined by,

$$d(k) := -\left[(1+\gamma)h^{r(k)}(X(k)) - h^{r(k)}(X(k+1)) - c(X(k))\right]. \tag{10}$$

(ii) *Eligibility vectors* are the sequence of q-dimensional vectors,

$$z(k) = \sum_{t=0}^{k}(1+\gamma)^{-t-1}\psi(X(k-t))\,, \qquad k \ge 1,$$

expressed recursively via,

$$z(k+1) = (1+\gamma)^{-1}[z(k) + \psi(X(k+1))], \qquad k \ge 0,\ z(0)=0.$$

Since $\boldsymbol{X}$ is ergodic we have for any $g\colon \mathsf{X} \to \mathbb{R}$,

$$\lim_{k\to\infty} \mathsf{E}[g(X(k))z(k)] = \langle g, \widetilde{R}_\gamma \psi \rangle.$$

Based on (9), for large k we obtain the approximation,

$$\mathsf{E}[d(k)z(k+1)] \approx -\tfrac{1}{2}\nabla_r \|h^r - h^*\|^2_M, \qquad r = r(k).$$

The TD algorithm is the stochastic approximation algorithm associated with the O.D.E. (3),

$$r(k+1) - r(k) = a_k d(k) z(k+1), \qquad k \ge 0. \tag{11}$$

The O.D.E. (3) is linear and exponentially asymptotically stable under the assumption that $A = \mathsf{E}_\pi[\psi(X)\psi(X)^{\mathrm{T}}] > 0$. Based on this fact, one can show that the sequence of estimates $\{r(k)\}$ obtained from the TD algorithm (11) is convergent for the vanishing step-size algorithm.

Extensions *Where to begin?* There is the issue of constructing the basis functions $\{\psi_i\}$ [5]. One can also extend these methods to construct an approximation based on a family of non-linearly parameterized functions $\{h^r\}$ [2, 8]. Below are a few extensions in the case of linear approximations.

- The most common extension found in the literature is to redefine the definition of $\{z(k)\}$. Fix any $\lambda \in [0,1]$ and consider the new definition,

$$z(k+1) = (1+\gamma)^{-1}[\lambda z(k) + \psi(X(k+1))], \qquad z(0) = 0.$$

The resulting algorithm (11) is called TD(λ), where the definition of the temporal differences remain unchanged. In particular, TD(0) takes the form,

$$r(k+1) - r(k) = a_k d(k) \psi(X(k+1)), \qquad k \ge 0. \tag{12}$$

The purpose of this modification is to speed convergence. The algorithm remains convergent to some $r(\infty) \in \mathbb{R}^q$, but it is no longer consistent. Bounds on the error $\|r(\infty) - r^*\|_M$ are obtained in [9, 4].

- One can change the error criterion. For example, consider instead the minimization of the mean-square "Bellman error",

$$\min_r \mathsf{E}_\pi[(Ph^r\,(X) - (1+\gamma)h^r(X) + c(X))^2]$$

Or, one might ask, why focus exclusively on this L_2 norm? The L_1 error may be more easily justified

$$\min_r \mathsf{E}_\pi[|Ph^r\,(X) - (1+\gamma)h^r(X) + c(X)|],$$

where in each case again $h^r(X) = r^{\mathrm{T}}\psi(X)$.

On differentiating we obtain a fixed point equation that can be solved using S.A. In the first the optimal parameter r^* satisfies,

$$\mathsf{E}_\pi[(r^{\mathrm{T}}(P\psi\,(X) - (1+\gamma)\psi(X) + c(X))(P\psi\,(X) - (1+\gamma)\psi(X))] = 0,$$

and in the second

$$\mathsf{E}_\pi[\mathrm{sign}\,[r^{\mathrm{T}}(P\psi\,(X) - (1+\gamma)\psi(X) + c(X)](P\psi\,(X) - (1+\gamma)\psi(X))] = 0.$$

The associated S.A. recursion appears to be complex since one must estimate $P\psi$.

• A simplification is obtained on eliminating the conditional expectation. Consider for simplicity the L_2 setting with,

$$\min_r \mathsf{E}_\pi[(h^r(X(k+1)) - (1+\gamma)h^r(X(k)) + c(X(k)))^2] \tag{13}$$

The minimization (13) is easily solved using S.A. since we don't have to estimate $P\psi$: The optimal parameter r^* satisfies,

$$\mathsf{E}_\pi[(r^{\mathrm{T}}(\psi(X(k+1)) - (1+\gamma)\psi(X(k)) + c(X(k)))(\psi(X(k+1)) - (1+\gamma)\psi(X(k)))] = 0.$$

This can be computed by simulating the deterministic O.D.E.,

$$\begin{aligned}\tfrac{d}{dr}r(t) &= -a\nabla_r\mathsf{E}_\pi[(h^r(X(k+1)) - (1+\gamma)h^r(X(k)) + c(X(k)))^2] \\ &= -a\mathsf{E}_\pi[(r^{\mathrm{T}}(t)(\psi(X(k+1)) - (1+\gamma)\psi(X(k)) + c(X(k)))(\psi(X(k+1)) - (1+\gamma)\psi(X(k)))].\end{aligned}$$

The associated discrete-time algorithm is similar to TD(λ),

$$r(k+1) - r(k) = a_k d(k) z(k+1), \qquad k \ge 0,$$

with $d(k)$ again defined in (10), and $z(k+1) := (1+\gamma)h^{r(k)}(X(k)) - h^{r(k)}(X(k+1))$.

• Finally, with an appropriate notion of distance, one can compute an optimal approximation h^{r^*} using a linear program (LP), or a simulation-based approximate LP [3].

References

[1] D. P. Bertsekas, V. Borkar, and A. Nedic. Improved temporal difference methods with linear function approximation. In J. Si, A. Barto, W. Powell, and D. Wunsch, editors, *Handbook of Learning and Approximate Dynamic Programming*, pages 690–705. Wiley-IEEE Press, Piscataway, NJ., 2004.

[2] D.P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Atena Scientific, Cambridge, Mass, 1996.

[3] D. P. de Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Res.*, 51(6):850–865, 2003.

[4] M. Kearns and S. Singh. Bias-variance error bounds for temporal difference updates. In *Proceedings of the 13th Annual Conference on Computational Learning Theory*, pages 142–147, 2000.

[5] S. Mannor, I. Menache, and N. Shimkin. Basis function adaptation in temporal difference reinforcement learning. *Annals of Operations Res.*, 134(2):215–238, 2005.

[6] A. Nedic and D.P. Bertsekas. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems: Theory and Applications*, 13(1-2):79–110, 2003.

[7] B. Van Roy. Neuro-dynamic programming: Overview and recent trends. In E. Feinberg and A. Shwartz, editors, *Markov Decision Processes: Models, Methods, Directions, and Open Problems*, pages 43–82. Kluwer, Holland, 2001.

[8] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press (and on-line, http://www.cs.ualberta.ca/%7Esutton/book/ebook/the-book.html), 1998.

[9] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Control*, 42(5):674–690, 1997.

ECE 555 Control of Stochastic Systems Fall 2005

Handout: ***Control Variates in Simulation***

In the past few lectures we have considered the general stochastic approximation recursion,

$$\theta(k+1) = \theta(k) + a_k[g(\theta(k)) + \Delta(k+1)], \qquad k \ge 0.$$

Under general conditions, verified by considering various ODEs, it is known that $\{\theta(k)\}$ converges to the set of zeros of g.

The remaining problem is that *convergence can be very slow.* These notes summarize the control variate method for speeding convergence in simulation. It is highly likely that this technique can be generalized to other recursive algorithms.

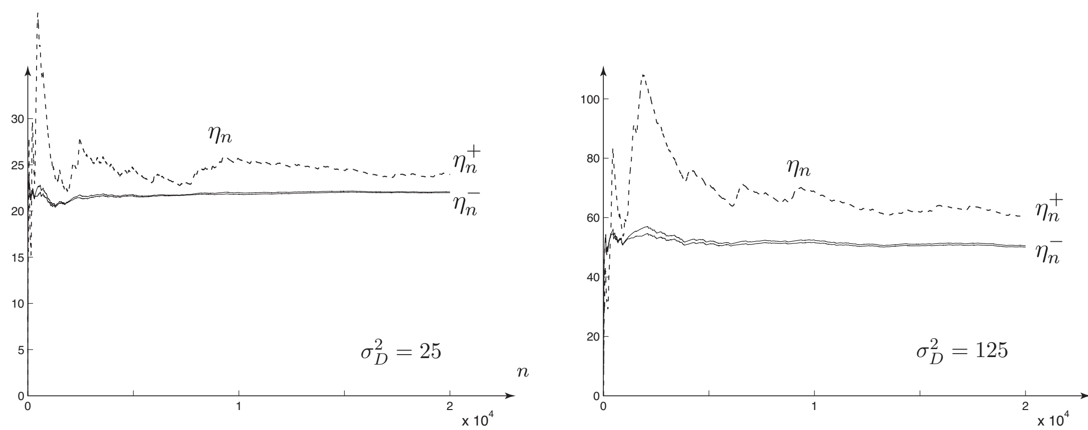


Figure 1: Simulation using the standard estimator, and the two controlled estimators. The plot at left shows results with $\sigma_D^2 = 25$, and at right the variance is increased to $\sigma_D^2 = 125$. In each case the estimates obtained from the standard Monte-Carlo estimator are significantly larger than those obtained using the controlled estimator, and the bound $\eta_n^- < \eta_n^+$ holds for all large n.

Simulating a Markov Chain Suppose that $\boldsymbol{X}$ is a Markov chain on a state space X with invariant distribution π. For background see [8] (as well as [10, 3, 8, 4].)

For a given function $F\colon \mathsf{X} \to \mathbb{R}$ we denote,

$$L_n(F) := \frac{1}{n}\sum_{k=0}^{n-1} F(X(k)) \qquad n \ge 1\,.$$

One can hope to establish the following limit theorems,

The Strong Law of Large Numbers, or SLLN: For each initial condition,

$$L_n(F) \to \pi(F), \qquad a.s.,\ n \to \infty. \tag{1}$$

The Central Limit Theorem, or CLT: For some $\sigma \ge 0$ and each initial condition,

$$\sqrt{n}[L_n(F) - \eta] \xrightarrow{\mathrm{w}} \sigma W, \qquad n \to \infty, \tag{2}$$

where W is a standard normal random variable, and the convergence is in distribution.

It is assumed here that the chain is *ergodic*, which means that the SLLN holds for any bounded function $F\colon \mathsf{X} \to \mathbb{R}$.

Suppose that $F\colon \mathsf{X} \to \mathbb{R}$ is a π-integrable function. Under ergodicity the SLLN can be generalized to any such function. Our interest is to efficiently estimate the finite mean $\eta = \pi(F)$. The standard estimator is the sample path average,

$$\eta_n = L_n(F) \qquad n \geq 1\,. \tag{3}$$

Its performance is typically gauged by the associated asymptotic variance σ^2 used in (2). Below are two well known representations in terms of the centered function $\widetilde{F} := F - \eta$.

Limiting variance:

$$\sigma^2 = \lim_{n\to\infty} n\mathsf{Var}_x(L_n(F)) := \lim_{n\to\infty} \mathsf{E}_x\big[L_n(\widetilde{F})^2\big] \tag{4}$$

Sum of the correlation function:

$$\sigma^2 = \sum_{k=-\infty}^{\infty} \mathsf{E}_\pi\big[\widetilde{F}(X(k))\widetilde{F}(X(0))\big] \tag{5}$$

The following operator-theoretic representation holds more generally. Let Z denote a version of the *fundamental kernel*, defined so that $\widehat{F} = ZF$ solves Poisson's equation for some class of functions F,

$$P\widehat{F} = \widehat{F} - F + \eta. \tag{6}$$

It will be convenient to apply the following bilinear and quadratic forms, defined for measurable functions $F, G\colon \mathsf{X} \to \mathbb{R}$,

$$\langle\!\langle F, G\rangle\!\rangle := P(FG) - (PF)(PG), \qquad \mathcal{Q}(F) := \langle\!\langle F, F\rangle\!\rangle.$$

Using this notation we have the following representation for the asymptotic variance,

$$\sigma^2(F) = \pi(\mathcal{Q}(\widehat{F})). \tag{7}$$

Recall that the resolvent is expressed $R := \sum_0^\infty 2^{-n-1}P^n$. The function $s : \mathsf{X} \to (0,1]$ and the probability measure ν are called *small* if the *minorization condition* holds,

$$R(x, A) \geq s(x)\nu(A), \qquad x \in \mathsf{X},\ A \in \mathcal{B}(\mathsf{X}).$$

The following is the general state space version of Condition (V3):

For functions $V : \mathsf{X} \to (0,\infty]$, $f : \mathsf{X} \to [1,\infty)$, a small function s, a small measure ν, and a constant $b < \infty$, $\qquad \mathcal{D}V \leq -f + bs \qquad$ **(V3)**

The following result is taken from [8, 6]:

Proposition. Suppose that $\boldsymbol{X}$ satisfies (V3) with $\pi(V^2) < \infty$. Then, the SLLN and CLT hold for any $F \in L_\infty^f$, and the asymptotic variance $\sigma^2(F)$ exists, and can be expressed as (4), (5), or (7) above. $\square$

Control-variates The purpose of the control-variate method is to reduce the variance of the standard estimator (3). See [7, 9, 2, 1] for background on the general control-variate method.

Suppose that $H\colon \mathsf{X} \to \mathbb{R}$ is a π-integrable function with known mean, and finite asymptotic variance. By normalization we can assume that $\pi(H) = 0$. Then, for a given $\vartheta \in \mathbb{R}$ and with $F_\vartheta := F - \vartheta H$, the sequence $\{L_n(F_\vartheta)\}$ provides an asymptotically unbiased estimator of $\pi(F)$. The asymptotic variance of the controlled estimator is given by

$$\sigma^2(F_\vartheta) = \mathcal{Q}(\widehat{F}_\vartheta) = \pi\big(\langle\!\langle ZF, ZF\rangle\!\rangle - 2\vartheta\langle\!\langle ZF, ZH\rangle\!\rangle + \vartheta^2\langle\!\langle ZH, ZH\rangle\!\rangle\big).$$

Minimizing over $\vartheta \in \mathbb{R}$ gives the estimator with minimal asymptotic variance,

$$\vartheta^* = \frac{\pi(\langle\!\langle ZF, ZH\rangle\!\rangle)}{\pi(\langle\!\langle ZH, ZH\rangle\!\rangle)}.$$

For a Markov chain it is easy to construct a function with zero mean: consider $H = J - PJ$ where J is known to have finite mean. Our goal then is to choose J so that it approximates the solution to Poisson's equation (6): The idea is that if $J = \widehat{F}$, then the resulting controlled estimator with $\vartheta = 1$ has *zero* asymptotic variance. This approach has been successfully applied in queueing models by taking J equal to the associated fluid value function described in lecture.

Consider the simple reflected random walk on $\mathbb{R}_+$, defined by the recursion

$$X(k+1) = [X(k) + D(k+1)]_+, \qquad k \ge 0, \tag{8}$$

with $[x]_+ = \max(x, 0)$ for $x \in \mathbb{R}$, and $\boldsymbol{D}$ i.i.d.. The fluid model is given by,

$$q(t) = [q(0) - \delta]_+, \qquad t \ge 0,$$

where $-\delta = \mathsf{E}[D(k)]$ is assumed to be negative. The fluid value function is the quadratic,

$$J(x) = \int_0^\infty q(t)\, dt = \tfrac{1}{2}\delta^{-1}x^2, \qquad x = q(0) \in \mathbb{R}_+.$$

Consider the special case in which $\boldsymbol{D}$ has common marginal distribution,

$$D(k) = \begin{cases} 1 & \text{with probability } \alpha; \\ -1 & \text{with probability } 1 - \alpha. \end{cases}$$

The Markov chain $\boldsymbol{X}$ is then a discrete-time model of the M/M/1 queue with state space $\mathsf{X} = \mathbb{Z}_+$. When $F(x) \equiv x$ we have seen that $\widehat{F}(x) = \frac{1}{2}\delta^{-1}(x^2 + x)$, so that the error $\widehat{F} - J$ is linear in x. Moreover, the representation (7) can be written,

$$\sigma^2(F) = \pi(\mathcal{Q}(\widehat{F})) = 2\pi(\widetilde{F}\widehat{F}) - \pi(\widetilde{F}^2) = \mathsf{E}[\tfrac{1}{2}\delta^{-1}\widetilde{X}^3 - \widetilde{X}^2]$$

which grows like δ^{-4} as $\delta \downarrow 0$ (equivalently, $\rho \uparrow 1$.)

Returning to the random walk (8), consider the following special case in which the sequence $\boldsymbol{D}$ is of the form $D(k) = A(k) - S(k)$, where $\boldsymbol{A}$ and $\boldsymbol{S}$ are mutually independent, i.i.d. sequences, with mean α, μ respectively. We let $\kappa > 0$ denote a variability parameter, and define

$$\mathsf{P}\{S(k) = (1+\kappa)\mu\} = 1 - \mathsf{P}\{S(k) = 0\} = (1+\kappa)^{-1}$$
$$\mathsf{P}\{A(k) = (1+\kappa)\alpha\} = 1 - \mathsf{P}\{A(k) = 0\} = (1+\kappa)^{-1}$$

Consequently, we have $-\delta = \mathsf{E}[A(k)] - \mathsf{E}[S(k)] = -(\mu - \alpha)$, and $\sigma_D^2 = \sigma_A^2 + \sigma_S^2 = (\mu^2 + \alpha^2)\kappa$.

The simulation results shown use $\mu = 4$ and $\alpha = 3$, so that $\delta = 1$. Two estimators $\{\eta_n^-, \eta_n^+\}$ were constructed based on the parameter values $\vartheta_- = 1.05$ and $\vartheta_+ = 1$. The plot at left in Figure 1 illustrates the resulting performance with $\kappa = 2$ ($\sigma_D^2 = 25$), and the plot at right shows the controlled and uncontrolled estimators with $\kappa = 5$, and hence $\sigma_D^2 = 125$.

Note that the bounds $\eta_n^- < \eta_n^+ < \eta_n$ hold for all large n, even though all three estimators are asymptotically unbiased.

A network model The *Kumar-Seidman-Rybko-Stolyar* (KSRS) network shown in Figure 2 is described in Chapter 1 of the course notes.

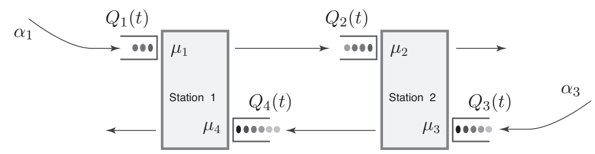


Figure 2: The Kumar-Seidman-Rybko-Stolyar (KSRS) network.

Consider the following policy based on a vector $\bar{w} \in \mathbb{R}_+^2$ of *safety-stock* values: Serve $Q_1 \geq 1$ at Station I if and only if $Q_4 = 0$, or

$$\mu_2^{-1} Q_2 + \mu_3^{-1} Q_3 \leq \bar{w}_2. \tag{9}$$

An analogous condition holds at Station II.

A simulation experiment was conducted to estimate the steady-state mean customer population. So, with $\mathsf{X} = \mathbb{Z}_+^4$, we let $F\colon \mathsf{X} \to \mathbb{R}_+$ denote the ℓ_1 norm on $\mathbb{R}^4$. A CRW network model was constructed in which the elements of $(\boldsymbol{A}, \boldsymbol{S})$ were taken Bernoulli (see course lecture notes.) Details can be found in [5].

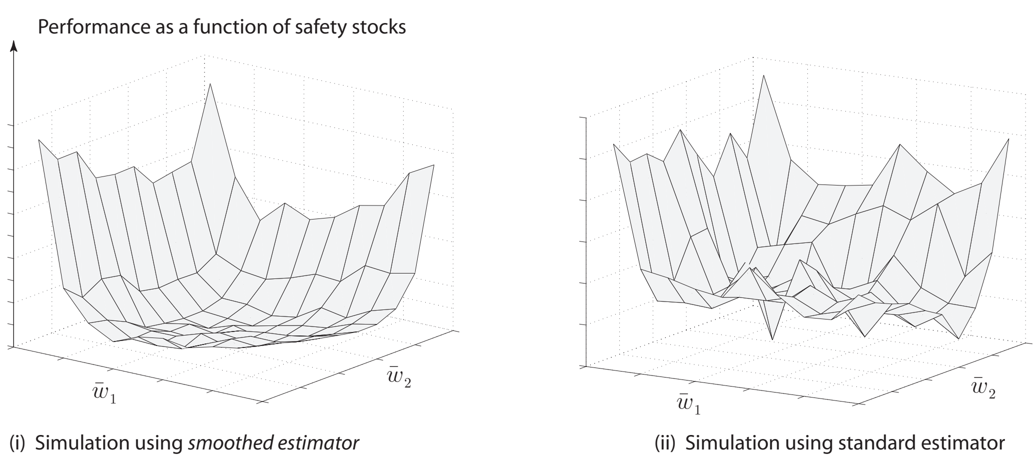


Figure 3: Estimates of the steady-state customer population in the KSRS model as a function of 100 different safety-stock levels using the policy (9). Two simulation experiments are shown, where in each case the simulation runlength consisted of $N = 200,000$ steps. The left hand side shows the results obtained using the smoothed estimator; the right hand side shows results with the standard estimator.

Shown in Figure 3 are estimates of the steady-state customer population in Case I for the family of policies (9), indexed by the safety-stock level $\bar{w} \in \mathbb{R}_+^2$. Shown at left are estimates obtained using the "smoothed estimator" based on a fluid value function. The plot at right shows estimates obtained using the standard estimator.

References

[1] P. Glasserman. *Monte Carlo methods in financial engineering.* Applications of Mathematics, 53. Springer, York, NY, 2004.

[2] P. Glynn and R. Szechtman. Some new perspectives on the method of control variates. In K.T. Fang, F.J. Hickernell, and H. Niederreiter, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2000: Proceedings of a Conference held at Hong Kong Baptist University, Hong Kong SAR, China*, pages 27–49, Berlin, 2002. Springer-Verlag.

[3] P. W. Glynn and S. P. Meyn. A Liapounov bound for solutions of the Poisson equation. *Ann. Probab.*, 24(2):916–931, 1996.

[4] P. W. Glynn and W. Whitt. Necessary conditions for limit theorems in cumulative processes. *Stoch. Proc. Applns.*, 98:199–209, 2002.

[5] S. G. Henderson, S. P. Meyn, and V. B. Tadić. Performance evaluation and policy selection in multiclass networks. *Discrete Event Dynamic Systems: Theory and Applications*, 13(1-2):149–189, 2003. Special issue on learning, optimization and decision making (invited).

[6] I. Kontoyiannis and S. P. Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab.*, 13:304–362, 2003. Presented at the INFORMS Applied Probability Conference, NYC, July, 2001.

[7] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis.* McGraw-Hill, New York, 3rd edition, 2000.

[8] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1993. online: `http://black.csl.uiuc.edu/~meyn/pages/book.html`.

[9] B. L. Nelson. Control-variate remedies. *Operations Res.*, 38(4):974–992, 1990.

[10] E. Nummelin. *General Irreducible Markov Chains and Nonnegative Operators.* Cambridge University Press, Cambridge, 1984.