TECHNICAL RESEARCH REPORT

Convergence of Sample Path Optimal Policies for Stochastic Dynamic Programming

by Michael C. Fu, Xing Jin

TR 2005-84

CONVERGENCE OF SAMPLE PATH OPTIMAL POLICIES FOR STOCHASTIC DYNAMIC PROGRAMMING

MICHAEL C. FU,* *University of Maryland*

XING JIN,** *National University of Singapore*

Abstract

We consider the solution of stochastic dynamic programs using sample path estimates. Applying the theory of large deviations, we derive probability error bounds associated with the convergence of the estimated optimal policy to the true optimal policy, for finite horizon problems. These bounds decay at an exponential rate, in contrast with the usual canonical (inverse) square root rate associated with estimation of the value (cost-to-go) function itself. These results have practical implications for Monte Carlo simulation-based solution approaches to stochastic dynamic programming problems where it is impractical to extract the explicit transition probabilities of the underlying system model.

Keywords: stochastic dynamic programming; Markov decision processes; sample path optimization; large deviations; simulation optimization.

2000 Mathematics Subject Classification: Primary 49L20
Secondary 65C50; 65C05

* Postal address: Robert H. Smith School of Business, University of Maryland College Park, MD 20742-1871

** Postal address: Department of Mathematics, National University of Singapore, Singapore 117543

1. Introduction

Consider a stochastic dynamic programming model (also known as a Markov decision process (MDP), see Arapostathis et al. 1993, Bertsekas 1995, Puterman 1994), in the setting where only sample paths of state transition sequences are available, e.g., when it is impractical to explicitly specify the transition probabilities, but the underlying system can be readily simulated. This is often the case when the system of interest is large and complex, and must therefore be modeled by a stochastic simulation model. One drawback of using sample path estimation is the relatively slow convergence rate for estimation of performance measures (e.g., the value, or cost-to-go, function), which is generally on the order of $O(N^{-0.5})$, where N is the number of sample paths. The focus of this paper is the problem of finding an optimal policy, and we exploit the fact that the policy search involves ordinal comparisons, rather than absolute estimation. In practice, the main idea of this approach is to compare relative orders of performance measures in finding the best action as quickly as possible rather than wasting effort in getting a more precise absolute estimate of the value function associated with each possible action. Under appropriate conditions, we show that the probability of selecting suboptimal actions is bounded by a quantity that decays to zero at an exponential rate.

The overriding purpose of our work is to provide a rigorous theoretical foundation for the sample path approach in finding good policies in stochastic dynamic programming problems. The convergence results obtained here are completely new to this setting. To put our results in some perspective, we touch on the most closely related work. A type of exponential (geometric) convergence rate is well known in the traditional MDP framework (e.g., Puterman 1994), where the convergence is with respect to the *horizon length* for the value iteration procedure in infinite horizon problems with explicitly known transition probabilities. Our finite action setting is included in the book of Bertsekas and Tsitsiklis (1996), where the solution approach goes under the name of neuro-dynamic programming, but the focus there is on *approximating the value function*, and sample path optimal policies are not analyzed. Our results buttress the literature on *ordinal optimization* see Ho et al. (1992, 2000), which focuses on the efficiency of ordinal comparisons rather than absolute estimation. In particular, the exponential convergence rate for *static* stochastic optimization problems is established in Dai (1996) and Dai and Chen (1997). Also, somewhat in the same spirit as our approach is the work of Robinson (1996) and Gürkan, Özge, and Robinson (1999), who consider sample path solution to stochastic variational inequalities, and establish conditions under which the sample path solution converges to the true solution; however, their setting is quite different from ours, in that we consider a dynamic model involving sequential decision making under uncertainty, and we focus on actually quantifying the error incurred in utilizing sample path estimates, going beyond just establishing convergence.

The rest of the paper is organized as follows. Section 2 defines the problem setting. Section 3 establishes the theoretical results on the exponentially decaying probability error bounds for the basic finite horizon discounted cost problems. Section 4 briefly discusses some easy extensions, and the Appendix contain the detailed proof of one of the more technical lemmas.

2. Problem Setting

In this section, we formulate the basic problem of minimizing total expected discounted cost in a setting where the state space and action space are finite, albeit possibly non-stationary. Let $\{X_k, k = 1, 2, ...\}$ denote a Markov decision process with finite state space $\mathcal{S}$ ($|\mathcal{S}| > 1$), where X_1 is the starting state. Let $T \geq 2$ be the time horizon, or number of periods (also known as stages), $\mathcal{S}_k \subseteq \mathcal{S}$ be the state space for the kth period, and $\mathcal{A}_k(x), x \in \mathcal{S}_k$, be the (finite) set of feasible actions in state x and period k. At stage k in state x, the decision maker chooses an action $a \in \mathcal{A}_k(x)$; as a result the following occur:

(i) an immediate (deterministic) cost $c_k(x, a) \geq 0$ is accrued, and

(ii) the process moves to a state $x' \in \mathcal{S}_{k+1}$ with transition probability $p_k(x'|x, a)$, where $p_k(x'|x, a) \geq 0$ and $\sum_{x' \in \mathcal{S}_{k+1}} p_k(x'|x, a) = 1$.

The objective is to find a sequence of *decision rules* $\{\mu_k(\cdot)\}$ comprising a *policy* $\mu = \{\mu_k\}$ that minimizes total expected discounted cost given by

$$E\left[\sum_{k=1}^{T} \alpha^{k-1} c_k(X_k, A_k)\right], \tag{1}$$

where A_k is the action taken in period k — which would be $\mu_k(X_k)$ under policy μ — and $\alpha \in (0, 1)$ is the (constant) discount factor. Here X_{k+1} depends on both X_k and A_k, i.e., given $X_k = x$ and $A_k = a$, we have

$$X_{k+1}(x, a) \sim \{p_k(\cdot|x, a)\}, \tag{2}$$

but such dependence will generally be suppressed for the sake of simplicity. Throughout, we assume a fixed initial state $X_1 = x_1$, but this can easily be generalized to the setting where the initial state is a random variable with an associated probability distribution.

Define the optimal cost-to-go (or value) function from stage k by

$$J_k(x) = \min_{\mu \in \mathcal{U}} E\left[\sum_{i=k}^{T} \alpha^{i-k} c_i(X_i, \mu_i(X_i)) \middle| X_k = x\right], \forall x \in \mathcal{S}_k, \quad k = 1, ..., T, \tag{3}$$

where $\mathcal{U}$ denotes the set of all policies. The value of the MDP is given by $J_1(x_1)$, and an *optimal* policy μ^* — defined as any policy that minimizes (1) — satisfies the following set of equations:

$$\mu_k^*(x) \in \arg\min_{a \in \mathcal{A}_k(x)} \{c_k(x, a) + \alpha E\left[J_{k+1}(X_{k+1}(x, a))\right]\}, \quad k = 1, 2, ..., T, \tag{4}$$

where the expectation is taken with respect to the next state X_{n+1}, which is a function of the current state x and action a, and we follow the convention that $J_{T+1}(\cdot) = 0$. It will be convenient to introduce the *Q-factors* defined by the expectation on the right-hand side (e.g., Bertsekas 1995):

$$Q_k(x, a) = c_k(x, a) + \alpha E\left[J_{k+1}(X_{k+1}(x, a))\right], \quad k = 1, 2, ..., T, \tag{5}$$

representing the expected cost of taking action a from state $X_k = x$ in period k, and then following the optimal policy thereafter. In particular,

$$Q_T(x, a) = c_T(x, a). \tag{6}$$

Thus, we have

$$J_k(x) = \min_{a \in \mathcal{A}_k(x)} Q_k(x, a).$$

For finite horizon dynamic programming with finite space, backward induction can be used via Equation (3) to obtain the optimal value functions $\{J_k(x), x \in \mathcal{S}_k, k = 1, 2, ..., T\}$ and a corresponding optimal policy satisfying (4). For the infinite horizon case, value iteration, policy iteration, or variants on these are used to solve the stationary version of (3) when applicable. When the transition probabilities are explicitly known, these procedures can sometimes be carried out in closed form or by using straightforward numerical procedures to calculate the necessary expectations.

In our setting, based on sample paths of the MDP sequence $X_1, X_2, ...$ for a given policy μ, the expectations in (1), (3), or (4), are estimated by taking sample means. By a *sample path optimal policy*, we mean a policy (possibly only partially specified, if not all states are visited in the sample paths) that optimizes the sample mean of the objective function given in (1). (This is not to be confused with using a single "long" sample path to estimate a stationary optimal policy for infinite horizon problems.) This will be a function of both the sample path length and the number of sample paths. For the finite horizon setting, the sample path length will be equal to the number of periods T, whereas in the infinite horizon case, the optimal policy is approximated by a finite horizon sample path optimal policy. A direct implementation for using sample paths would be to take a "large" number of samples for each value that must be estimated, thus in essence reducing the problem to the traditional setting. In practice, taking a large number of samples may be unnecessarily wasteful, especially when the ultimate objective is to find the optimal policy, not necessarily to precisely estimate the optimal value functions for all states. The underlying philosophy is that one may obtain good policies through ordinal comparison even while the estimate of the value function itself is not that accurate.

3. Sample Path Probability Error Bounds

We now derive probability error bounds for the convergence of sample path optimal policies to a true optimal policy. We focus on searching for the optimal action in the first period, since optimal actions for subsequent periods can be obtained in the same manner. Write the feasible action set for the initial period starting in state x_1 as

$$\mathcal{A}_1(x_1) = \{a_1, a_2, ..., a_m\}.$$

The Q-factor of interest for the first period, as defined by (5), is

$$Q_1(x, a) = c_1(x, a) + \alpha E\left[J_2(X_2(x, a))\right],$$

where J_2 is the cost-to-go function defined by (3) with horizon $T-1$. Since throughout we are focusing on the first period with initial state $X_1 = x_1$, we will simplify notation

by dropping explicit display of the dependence on the period and initial state by defining the unsubscripted Q-factor:

$$Q(a) = Q_1(x_1, a).$$

Without loss of generality, we assume

$$Q(a_1) < Q(a_2) \leq \ldots \leq Q(a_m),$$

i.e., $\mu_1^*(x_1) = a_1$. The case with ties for the best can also be handled in exactly the same way; see Remark 3.2 following Theorem 3.1.

The procedure to estimate the optimal action from a given state in the setting of this section uses sample trees. Specifically, for state x_1, for each action $a_l \in \mathcal{A}_1(x_1)$, n independent sample trees are generated. Each tree begins by taking action a_l in period 1, and then sampling all possible actions in subsequent states visited. Since the state space is finite, there may be common states visited between trees and also within trees. We keep the tree structure by sampling from each node *separately and independently* according to (2), so there will be no "recombining" branches, even if the same state were reached at different nodes of the tree. To be more specific, a sample tree is generated as follows for initial period action a_l:

(i) In period 1, generate one next (period 2) state sample (node) according to $p_1(\cdot|x_1, a_l)$.

(ii) In period 2, generate a next (period 3) state sample (node) according to $p_2(\cdot|x, a)$, for each feasible action $a \in \mathcal{A}_2(x)$, where x is the state generated in step (i).

(iii) Starting from each state x visited in period k of the tree ($k = 3, \ldots, T-1$), generate a next (period k+1) state sample (node) according to $p_k(\cdot|x, a)$ for each feasible action $a \in \mathcal{A}_k(x)$.

As mentioned earlier, all sampling is done independently of other trees and other nodes in the same tree; however, correlation between sampling of different actions from the *same node* in a tree is allowed. Let $\mathcal{S}_k^{(l)} \subseteq \mathcal{S}_k, k = 2, \ldots, T$, denote the set of states actually visited in period k over all n sample trees initiated with action a_l. An example for n=3 is shown in Figure 1. In this example, even if $x_5 = x_6$, i.e., the state reached is the same, the nodes themselves remain distinct, in that separate independent samples would be generated from each for each possible action in $\mathcal{A}_3(x_5) = \mathcal{A}_3(x_6)$.

Sample path estimates for the Q-factors and cost-to-go functions are obtained via backward induction as follows:

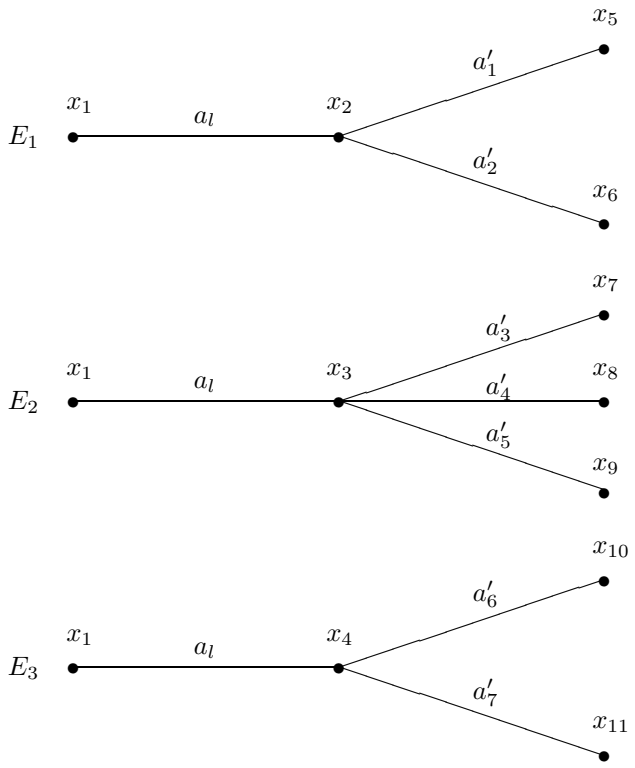
$$\widehat{Q}_T^{(l)}(x, a) = c_T(x, a), \ x \in \mathcal{S}_T^{(l)}, \tag{7}$$

$$\widehat{\mu}_k^{(l)}(x) \in \arg\min_{a \in \mathcal{A}_k(x)} \widehat{Q}_k^{(l)}(x, a), \ x \in \mathcal{S}_k^{(l)}, \ k = 2, \ldots, T, \tag{8}$$

$$\widehat{J}_k^{(l)}(x) = \min_{a \in \mathcal{A}_k(x)} \widehat{Q}_k^{(l)}(x, a) = \widehat{Q}_k^{(l)}(x, \widehat{\mu}_k^{(l)}(x)), \ x \in \mathcal{S}_k^{(l)}, \ k = 2, \ldots, T, \tag{9}$$

$$\widehat{Q}_k^{(l)}(x, a) = c_k(x, a) + |\mathcal{N}_{k+1}^{(l)}(x, a)|^{-1} \alpha \sum_{y \in \mathcal{N}_{k+1}^{(l)}(x,a)} \widehat{J}_{k+1}^{(l)}(y), \ x \in \mathcal{S}_k^{(l)}, \ k=2, \ldots, T-1, \tag{10}$$

where $\mathcal{N}_{k+1}^{(l)}(x, a)$ is the multi-set (i.e., includes states repeated if sampled more than once) of states reached in period $k+1$ from state x with action a in period k ($k = 1, \ldots, T-1$).



$$\mathcal{A}_2(x_2) = \{a'_1, a'_2\}, \quad \mathcal{A}_2(x_3) = \{a'_3, a'_4, a'_5\}, \quad \mathcal{A}_2(x_3) = \{a'_6, a'_7\},$$
$$\mathcal{S}_2^{(l)} = \{x_2, x_3, x_4\}, S_3^{(l)} = \{x_5, x_6, \ldots, x_{11}\}.$$

FIGURE 1: Example of simulated trees for n=3.

Similar to the unsubscripted initial period, initial state, Q-factors defined earlier, we define the following corresponding tree-based estimator:

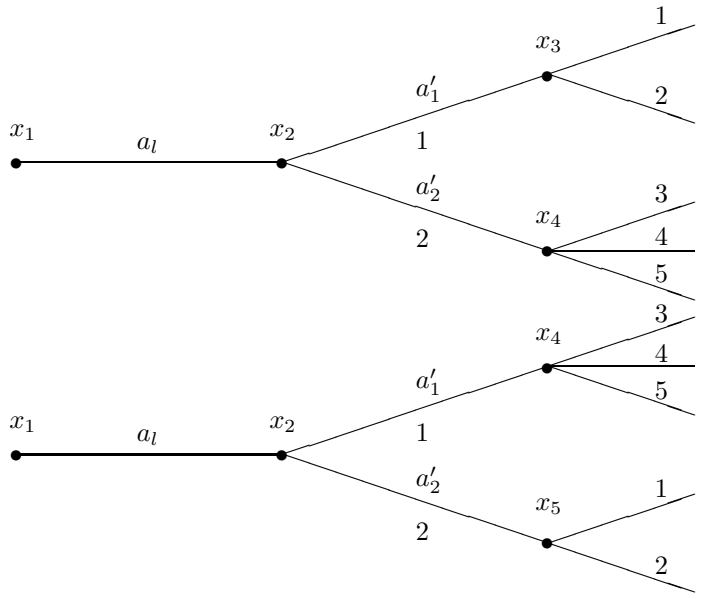
$$\widehat{Q}(a_l) = c_1(x_1, a_l) + \frac{1}{n}\alpha \sum_{y \in \mathcal{N}(a_l)} \widehat{J}_2^{(l)}(y),$$

where we have defined $\mathcal{N}(a_l) = \mathcal{N}_2^{(l)}(x_1, a_l)$ and $|\mathcal{N}(a_l)| = n$. We then estimate the optimal first-period action in the natural way:

$$\hat{a}_1(n) = \arg \min_{a_l \in \mathcal{A}_1(x_1)} \{\widehat{Q}(a_l)\}. \tag{11}$$

Averaging over $\mathcal{N}_{k+1}^{(l)}(x, a)$ in (10) is needed to ensure consistency in defining decision rules via (8), since the same state can be reached more than once in sampling, on the same tree or on different trees. If all n trees for a given a_l are distinct with no common states in any period beyond the initial state, so $|\mathcal{N}_{k+1}^{(l)}(x, a)| = 1$ for $k > 1$, then the DP algorithm simply corresponds to performing (deterministic) backward induction individually on each tree.

Figure 2 shows a simple example, which we use to illustrate how Equations (7)-(11) are applied and why the averaging is necessary. There are two trees (n=2), and both reach the same state x_2 in period 2, hence the multi-set $\mathcal{N}(a_l) = \{x_2, x_2\}$. Assume



$$\mathcal{N}_2(x_1, a_l) = \{x_2, x_2\}, \mathcal{N}_3(x_2, a_1') = \{x_3, x_4\}, \mathcal{N}_3(x_2, a_2') = \{x_4, x_5\},$$

$$\mathcal{S}_2^{(l)} = \{x_2\}, \mathcal{S}_3^{(l)} = \{x_3, x_4, x_5\}.$$

FIGURE 2: Example of two simulated trees with common states (costs shown on branches). Dynamic programming performed separately on each tree would lead to a different action from state x_2 in period 2 on the two trees (a_1' in the upper tree, a_2' in the lower tree). Averaging appropriately over the corresponding nodes in two trees leads to $\widehat{\mu}_2(x_2) = a_1'$.

for simplicity that the discount factor is one ($\alpha = 1$). Applying the DP algorithm separately to each tree, we obtain (suppressing superscripted (l) for notational convenience) $\widehat{J}_3(x_3) = 1, \widehat{J}_3(x_4) = 3, \widehat{J}_3(x_5) = 1$; for the upper tree, $\widehat{J}_2(x_2) = 2$ and $\widehat{\mu}_2(x_2) = a_1'$, whereas for the lower tree, $\widehat{J}_2(x_2) = 3$ and $\widehat{\mu}_2(x_2) = a_2'$, leading to a conflict in specifying the decision rule (action for state x_2). On the other hand, with the averaging (over just the two trees, i.e., $n = 2$), $\widehat{Q}_2(x_2, a_1') = 1 + (1 + 3)/2 = 3$ and $\widehat{Q}_2(x_2, a_2') = 2 + (3 + 1)/2 = 4$, which gives $\widehat{J}_2(x_2) = \min\{\widehat{Q}_2(x_2, a_1'), \widehat{Q}_2(x_2, a_2')\} = 3$ and $\widehat{\mu}_2(x_2) = \arg\min_{a_i'}\{\widehat{Q}_2(x_2, a_i')\} = a_1'$, hence $\widehat{Q}(a_l) = c_1(x_1, a_l) + 3$. This would be repeated for all other actions in $\mathcal{A}_1(x_1)$, and then (one of) the action(s) with the lowest value of $\widehat{Q}(\cdot)$ would be selected to be the estimated optimal action in state x_1.

Our results use the large deviations principle (cf. Dembo and Zeitouni 1998), which yields exponentially decaying probability error bounds under appropriate conditions.

Lemma 3.1: Consider a sequence of i.i.d. random variables $\{Y_n, n \geq 1\}$ with moment generating function $M(\lambda) = E[\exp(\lambda Y_1)]$. Let $S_n = \sum_{i=1}^n Y_i$. If $M(\lambda)$ exists in a neighborhood $(-\varepsilon, \varepsilon)$ of $\lambda = 0$ for some $\varepsilon > 0$, then

$$P(S_n/n \geq x) \leq \exp(-n\Lambda_+^*(x)), \quad \forall x,$$

and

$$P(S_n/n \leq x) \leq \exp(-n\Lambda_-^*(x)), \quad \forall x,$$

where

$$\Lambda_+^*(x) = \sup_{0 \le \lambda < \varepsilon} (\lambda x - \log M(\lambda))$$

and

$$\Lambda_-^*(x) = \sup_{-\varepsilon < \lambda \le 0} (\lambda x - \log M(\lambda)).$$

Furthermore, if $|Y_1| \le M$ for some constant $M < \infty$, then $\Lambda_+^*(x) > 0$ for $x > E[Y_1]$, and $\Lambda_-^*(x) > 0$ for for $x < E[Y_1]$.

Proof. The first part follows directly from Xie (1997). For the second part, we show only the $x > E[Y_1]$ case, since the $x < E[Y_1]$ case is similar.

Using a Taylor series expansion around $\lambda \ge 0$, there exists $\xi \in [0, \lambda)$ such that

$$\begin{aligned} \Lambda(\lambda) &= \log E[\exp[\lambda Y_1]] \\ &= \Lambda(0) + \Lambda'(0)\lambda + \frac{1}{2}\Lambda''(\xi)\lambda^2 = \lambda E[Y_1] + \frac{1}{2}\Lambda''(\xi)\lambda^2, \end{aligned}$$

the last equality following from $\Lambda(0) = 0$ and $\Lambda'(0) = E[Y_1]$.

We now turn to evaluating $\Lambda''(\xi)$. Since, $|Y_1| \le M$,

$$\begin{aligned} \Lambda''(\xi) &= \frac{E[Y_1^2 \exp(\xi Y_1)]E[\exp(\xi Y_1)] - (E[Y_1 \exp(\xi Y_1)])^2}{(E[\exp(\xi Y_1)])^2} \\ &\le \frac{E[Y_1^2 \exp(\xi Y_1)]}{E[\exp(\xi Y_1)]} \le M^2. \end{aligned}$$

Consequently, for $x > E[Y_1]$,

$$\begin{aligned} \Lambda_+^*(x) &= \sup_{\lambda \ge 0}\{\lambda x - \log E[\exp[\lambda Y_1]]\} \\ &\ge \sup_{\lambda \ge 0}\left\{\lambda(x - E[Y_1]) - \frac{M^2\lambda^2}{2}\right\} > 0, \end{aligned} \tag{12}$$

completing the proof. □

Remark 3.1: For a finite-horizon MDP with finite action and state spaces, the total discounted cost $\sum_{k=1}^{T} \alpha^{k-1} c_k(X_k, \mu_k(X_k))$ has finite moment generating function on $(-\infty, \infty)$ for any policy $\mu \in \mathcal{U}$. Define

$$\widetilde{c}_0 = \max_{k \in \{1,2,\ldots,T\}} \max_{x \in \mathcal{S}_k, a \in \mathcal{A}_k(x)} c_k(x, a),$$

and

$$\widetilde{J}_0 = \sum_{k=1}^{T} \alpha^{k-1} \widetilde{c}_0.$$

From the backward induction DP algorithm, it is easy to show that for any l, k, and x,

$$\widehat{J}_k^{(l)}(x) \le \widetilde{J}_0,$$

and, from the definition of $J_k(x)$, it is easy to see

$$J_k(x) \le \widetilde{J}_0.$$

Set

$$|\mathcal{A}| = \max_{k\in\{1,2,\ldots,T\}} \max_{x\in\mathcal{S}_k} |\mathcal{A}_k(x)|.$$

Lemma 3.2: If $\gamma \in (0,1)$ and $\delta > 0$ satisfy

$$2\Delta_{\gamma,\delta} < Q(a_2) - Q(a_1), \tag{13}$$

$$\text{where } \Delta_{\gamma,\delta} = (\sum_{i=1}^{T-1} \alpha^{i-1})\delta + (\sum_{i=1}^{T-2} \alpha^{i-1})|\mathcal{S}|\gamma\widetilde{J}_0, \tag{14}$$

then

$$P\left(\bigcup_{l=1}^{m}\{\widehat{Q}(a_l) \geq Q(a_l) + \Delta_{\gamma,\delta}\}\right) \leq |\mathcal{A}|\left(\sum_{i=0}^{T-2} |\mathcal{S}|^i \exp(-n\gamma^i\delta')\right), \tag{15}$$

$$P\left(\bigcup_{l=1}^{m}\{\widehat{Q}(a_l) \leq Q(a_l) - \Delta_{\gamma,\delta}\}\right) \leq |\mathcal{A}|\left(\sum_{i=0}^{T-2} |\mathcal{S}|^i \exp(-n\gamma^i\delta')\right), \tag{16}$$

where

$$\delta' = \sup_{\lambda\geq 0}\left\{\lambda\delta - \frac{(\alpha\widetilde{J}_0)^2\lambda^2}{2}\right\} > 0. \tag{17}$$

Proof. See the Appendix. □

We are now in a position to present and prove the main result of this section. In words, the theorem states that the sample path first-period optimal action(s) contained in the set $\hat{a}_1(n)$ converges in probability to the true optimal action a_1 for the finite horizon problem at an exponentially decaying rate with respect to the number of sample paths (trees).

Theorem 3.1:

$$P\left(\hat{a}_1(n) \neq \{a_1\}\right) \leq 2|\mathcal{A}|\left(\sum_{i=0}^{T-2} |\mathcal{S}|^i \exp(-n\gamma^i\delta')\right),$$

where γ and δ satisfy the conditions of Lemma 3.2 and δ' is given by (17).

Remark 3.2: If $Q(a_1) = Q(a_2) = \ldots = Q(a_k) < Q(a_{k+1}) \leq \ldots \leq Q(a_m)$, then the left-hand side just becomes $P\left(\hat{a}_1(n) \not\subseteq \{a_1, \ldots, a_k\}\right)$.

Proof. Suppose that $\hat{a}_1(n) \neq \{a_1\}$. Then, $\exists\ l \neq 1$ such that $\widehat{Q}(a_l) \leq \widehat{Q}(a_1)$, i.e.,

$$\begin{aligned} P\left(\hat{a}_1(n) \neq \{a_1\}\right) &= P\left(\bigcup_{l\neq 1}\{\widehat{Q}(a_l) \leq \widehat{Q}(a_1)\}\right) \\ &\leq P\left(\bigcup_{l\neq 1}\{\widehat{Q}(a_l) \leq \widehat{Q}(a_1)\}, \widehat{Q}(a_1) \leq Q(a_1) + \Delta_{\gamma,\delta}\right) \\ &\quad + P\left(\widehat{Q}(a_1) > Q(a_1) + \Delta_{\gamma,\delta}\right). \end{aligned}$$

Since $Q(a_2) \leq Q(a_l)$ for any $a_l (\neq a_1)$, condition (13) gives

$$Q(a_1) < Q(a_2) - 2\Delta_{\gamma,\delta} \leq Q(a_l) - 2\Delta_{\gamma,\delta},$$

or

$$Q(a_1) + \Delta_{\gamma,\delta} \leq Q(a_l) - \Delta_{\gamma,\delta},$$

so we have

$$\begin{aligned} P\left(\hat{a}_1(n) \neq \{a_1\}\right) &\leq P\left(\bigcup_{l\neq 1}\{\widehat{Q}(a_l) \leq Q(a_l) - \Delta_{\gamma,\delta}\}\right) + |\mathcal{A}|\left(\sum_{i=0}^{T-2} |\mathcal{S}|^i \exp(-n\gamma^i\delta')\right) \\ &\leq 2|\mathcal{A}|\left(\sum_{i=0}^{T-2} |\mathcal{S}|^i \exp(-n\gamma^i\delta')\right), \end{aligned}$$

where Lemma 3.2 has been applied twice. □

4. Extensions

The results can be extended to the following cases with essentially the same framework:

- random costs;
- stochastic and non-stationary discount factor, by replacing α^k throughout by $\prod_{j=1}^{k} \alpha_j$, where α_j is the discount rate for period j.

Convergence of the same algorithm for infinite state spaces is not a problem, but the current method of proof for the convergence *rate* result will not carry through. Extension to infinite action spaces is also not straightforward, as the current algorithm is not even applicable. These extensions are topics of ongoing research.

Appendix A. Proof of Lemma 3.2

We show (15) only, as the proof for (16) proceeds analogously. First, we first establish three preliminary results.

Lemma A1: Let $Z_i \sim p_k(\cdot|x,a)$ i.i.d. for fixed $x \in \mathcal{S}_k, a \in \mathcal{A}_k(x)$. For any $N \geq 0$ and $\delta > 0$,

$$P\left(\frac{\alpha}{N}\sum_{i=1}^{N} J_{k+1}(Z_i) \geq \alpha E[J_{k+1}(Z_i)] + \delta\right) \leq \exp(-N\delta'),$$

$k = 1, ..., T-1$, where δ' is given by (17).

Proof. The proof follows directly from Lemma 3.1, with $Y_i = \alpha(J_{k+1}(Z_i) - E[J_{k+1}(Z_i)])$, so $E[Y_i] = 0$ and Y_i has finite moment generating function (cf. Remark 3.1). Applying the first part of Lemma 3.1 leads to

$$P\left(\frac{1}{N}\sum_{i=1}^{N} \alpha J_{k+1}(Z_i) \geq \alpha E[J_{k+1}(Z_i)] + \delta\right) \leq \exp(-N\Lambda_+^*(\delta)),$$

where

$$\Lambda_+^*(\delta) = \sup_{\lambda\geq 0}(\lambda\delta - \log E[\exp[\lambda Y_i]]).$$

Since $|Y_i| \leq \alpha\widetilde{J}_0 = M$, the second part of Lemma 3.1 can be applied:

$$\sup_{\lambda\geq 0}(\lambda\delta - \log E[\exp[\lambda Y_i]]) \geq \sup_{\lambda\geq 0}\left\{\lambda\delta - \frac{(\alpha\widetilde{J}_0)^2\lambda^2}{2}\right\} \equiv \delta' > 0,$$

with δ' derived using (12). □

Lemma A1′: Under the same conditions as Lemma A1, let $\mathcal{N}$ be a non-negative integer-valued random variable independent of $\{Z_i\}$. Then,

$$P\left(\frac{\alpha}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}} J_{k+1}(Z_i) \geq \alpha E[J_{k+1}(Z_i)] + \delta, \mathcal{N} \geq N\right) \leq \exp(-N\delta'),$$

$k = 1, ..., T-1$, where δ' is given by (17).

Proof. Using Lemma A1, note that the conditional probability

$$\begin{aligned} & P\left[\frac{\alpha}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}} J_{k+1}(Z_i) \geq \alpha E[J_{k+1}(Z_i)] + \gamma \middle| \mathcal{N} = N_0\right] \\ &= P\left[\frac{\alpha}{N_0}\sum_{i=1}^{N_0} J_{k+1}(Z_i) \geq \alpha E[J_{k+1}(Z_i)] + \gamma \middle| \mathcal{N} = N_0\right] \leq \exp(-N_0\delta'). \end{aligned}$$

Unconditioning yields the desired result. □

Note that $|\mathcal{N}_t^{(l)}(x,a)|$ is constant over $\mathcal{A}_{t-1}(x)$, i.e., $|\mathcal{N}_t^{(l)}(x,a)| = |\mathcal{N}_t^{(l)}(x,a')|$, for all $a' \in \mathcal{A}_{t-1}(x)$, so we simplify notation by dropping the dependence on the action in writing $|\mathcal{N}_t^{(l)}(x)|$ for $|\mathcal{N}_t^{(l)}(x,a)|$.

Lemma A2: For $x \in \mathcal{S}_{T-1}$,

$$P\Big(\widehat{J}_{T-1}^{(l)}(x) \geq J_{T-1}(x) + \delta, |\mathcal{N}_T^{(l)}(x)| \geq N\Big) \leq |\mathcal{A}| \exp(-N\delta'), \tag{18}$$

Proof. For $x \in \mathcal{S}_T$,

$$\widehat{J}_T^{(l)}(x) = \min_{a\in\mathcal{A}_T(x)} Q_T(x,a) = \min_{a\in\mathcal{A}_T(x)} c_T(x,a) = J_T(x), \tag{19}$$

so

$$\widehat{J}_{T-1}^{(l)}(x) = \min_{a\in\mathcal{A}_{T-1}(x)} \Big\{c_{T-1}(x,a) + |\mathcal{N}_T^{(l)}(x)|^{-1}\alpha \sum_{y\in\mathcal{N}_T^{(l)}(x,a)} J_T(y)\Big\}.$$

Note that

$$\begin{aligned} & \Big\{\widehat{J}_{T-1}^{(l)}(x) \geq J_{T-1}(x) + \delta\Big\} \\ &= \Big\{\min_{a\in\mathcal{A}_{T-1}(x)} \Big\{c_{T-1}(x,a) + |\mathcal{N}_T^{(l)}(x)|^{-1}\alpha \sum_{y\in\mathcal{N}_T^{(l)}(x,a)} J_T(y)\Big\} \\ &\qquad \geq \min_{a\in\mathcal{A}_{T-1}(x)} \Big\{c_{T-1}(x,a) + \alpha E[J_T(X_T(x,a))] + \delta\Big\}\Big\} \\ &\subseteq \bigcup_{a\in\mathcal{A}_{T-1}(x)} \Big\{|\mathcal{N}_T^{(l)}(x)|^{-1}\alpha \sum_{y\in\mathcal{N}_T^{(l)}(x,a)} J_T(y) \geq \alpha E[J_T(X_T(x,a))] + \delta\Big\}. \end{aligned}$$

Thus,

$$\begin{aligned} & P\left(\widehat{J}_{T-1}^{(l)}(x) \geq J_{T-1}(x) + \delta, |\mathcal{N}_T^{(l)}(x)| \geq N\right) \\ &\leq P\Big(\bigcup_{a\in\mathcal{A}_{T-1}(x)} \Big\{|\mathcal{N}_T^{(l)}(x)|^{-1}\alpha \sum_{y\in\mathcal{N}_T^{(l)}(x,a)} J_T(y) \geq \alpha E[J_T(X_T(x,a))] + \delta, |\mathcal{N}_T^{(l)}(x)| \geq N\Big\}\Big) \\ &\leq |\mathcal{A}| \exp(-N\delta'), \end{aligned}$$

the last inequality following from Lemma A1′, proving (18). □

Lemma A3: For $k \in \{2,3,...,T-1\}, x \in \mathcal{S}_k$, and $a \in \mathcal{A}_k(x)$,

$$P\Big(\widehat{J}_k^{(l)}(x) \geq J_k(x) + C_k, |\mathcal{N}_{k+1}^{(l)}(x)| \geq n\gamma^{k-1}\Big) \leq D_k, \tag{20}$$

where $C_k = (\sum_{i=1}^{T-k} \alpha^{i-1})\delta + (\sum_{i=1}^{T-k-1} \alpha^{i-1})|\mathcal{S}|\gamma\widetilde{J}_0$
and $D_k = |\mathcal{A}|(\sum_{i=0}^{T-1-k} |\mathcal{S}|^i \exp(-n\gamma^{k-1+i}\delta'))$.
In particular, $C_1 = \Delta_{\gamma,\delta}$ and $C_{T-1} = \delta$.

Proof. We establish the result via backward induction. By (18) in Lemma A2, (20) holds when $k = T-1$. Assuming that (20) is true when $k = t,\ t \in \{3,...,T-1\}$, we want to show that it holds when $k = t-1$.

Recall that $N_k^{(l)}(y)$ denotes the number of times state y is reached in period k over all n sampled trees initiated by a_l, and define the set

$$\mathcal{R}_k^{(l)} \equiv \mathcal{R}_k^{(l)}(\gamma) = \{y \in \mathcal{S}_k^{(l)} : N_k^{(l)}(y) \geq n\gamma^{k-1}\},$$

where explicit dependence on γ is omitted for notational simplification, since it is fixed. If $y \in \mathcal{R}_k^{(l)}$, then state y was visited at least $n\gamma^{k-1}$ times in period k.

From the definition of $\widehat{J}_t^{(l)}$ given by (9) and (10), we have the following decomposition for $x \in \mathcal{S}_{t-1}$:

$$\begin{aligned}
\widehat{J}_{t-1}^{(l)}(x) &= \min_{a \in \mathcal{A}_{t-1}(x)} \Big\{ c_{t-1}(x,a) + |\mathcal{N}_t^{(l)}(x)|^{-1}\alpha \sum_{y \in \mathcal{N}_t^{(l)}(x,a)} \widehat{J}_t^{(l)}(y) \Big\} \\
&= \min_{a \in \mathcal{A}_{t-1}(x)} \Big\{ c_{t-1}(x,a) + |\mathcal{N}_t^{(l)}(x)|^{-1}\alpha \sum_{y \in \mathcal{N}_t^{(l)}(x) \cap \mathcal{R}_t^{(l)}} \widehat{J}_t^{(l)}(y) \\
&\quad + |\mathcal{N}_t^{(l)}(x)|^{-1}\alpha \sum_{y \in \mathcal{N}_t^{(l)}(x,a) \cap \bar{\mathcal{R}}_t^{(l)}} \widehat{J}_t^{(l)}(y) \Big\},
\end{aligned} \tag{21}$$

where the set complement is denoted using the overbar, and the intersection of a multi-set and an ordinary set is assumed to be given by a corresponding multi-set. For example, $\{1,1,1,2,3\} \cap \{1,3,5\} = \{1,1,1,3\}$. We now find bounds for each of the last two terms in the decomposition given by (21).

By definition of $\mathcal{R}_k^{(l)}$, we have the following bound:

$$|\mathcal{N}_k^{(l)}(x,a) \cap \bar{\mathcal{R}}_k^{(l)}| = \sum_{y \in \mathcal{S}} N_k^{(l)}(y)\mathbf{1}\{N_k^{(l)}(y) < n\gamma^{k-1}\} < \sum_{y \in \mathcal{S}} n\gamma^{k-1} = |\mathcal{S}|n\gamma^{k-1}.$$

Thus, for $x \in \mathcal{S}_{k-1}$ such that $|\mathcal{N}_t^{(l)}(x)| \geq n\gamma^{t-1}$, we have

$$\begin{aligned}
|\mathcal{N}_t^{(l)}(x)|^{-1}\alpha \sum_{y \in \mathcal{N}_t^{(l)}(x,a) \cap \bar{\mathcal{R}}_t^{(l)}} \widehat{J}_t^{(l)}(y) &\leq \frac{|\mathcal{N}_t^{(l)}(x,a) \cap \bar{\mathcal{R}}_t^{(l)}|}{|\mathcal{N}_t^{(l)}(x)|}\alpha\widetilde{J}_0 \qquad (\text{since } \widehat{J}_t^{(l)}(\cdot) \leq \widetilde{J}_0) \\
&\leq \alpha\widetilde{J}_0|\mathcal{S}|n\gamma^t/(n\gamma^{t-1}) \leq |\mathcal{S}|\gamma\widetilde{J}_0.
\end{aligned} \tag{22}$$

Note that for $a \in \mathcal{A}_{t-1}(x), y \in \mathcal{N}_t^{(l)}(x,a) \cap \mathcal{R}_t^{(l)}$, we have $|\mathcal{N}_{t+1}^{(l)}(y)| \geq n\gamma^t$, and by the induction assumption, (20) holds when $k = t$, so

$$\begin{aligned}
&P(\widehat{J}_t^{(l)}(y) \geq J_t(y) + C_t, y \in \mathcal{N}_t^{(l)}(x,a) \cap \mathcal{R}_t^{(l)}) \\
&\quad \leq P\Big(\widehat{J}_t^{(l)}(y) \geq J_t(y) + C_t, |\mathcal{N}_{t+1}^{(l)}(y)| \geq n\gamma^t\Big) \leq D_t,
\end{aligned}$$

implying that

$$P\left(\bigcup_{a\in\mathcal{A}_{t-1}(x)}\bigcup_{y\in\mathcal{N}_t^{(l)}(x,a)\cap\mathcal{R}_t^{(l)}}\{\widehat{J}_t^{(l)}(y)\geq J_t(y)+C_t\}\right)$$
$$\leq\quad P\left(\bigcup_{i=1}^{|\mathcal{S}|}\{\widehat{J}_t^{(l)}(s_i)\geq J_t(s_i)+C_t,|\mathcal{N}_{t+1}^{(l)}(s_i)|\geq n\gamma^t\}\right)\leq|\mathcal{S}|D_t, \tag{23}$$

where we have enumerated all possible states as $\mathcal{S}=\{s_1,\ldots,s_{|\mathcal{S}|}\}$.

Hence, similar to the proof of Lemma A2, by combining (21), (22) and (23), we have

$$\begin{aligned}
&P\Big(\widehat{J}_{t-1}^{(l)}(x)\geq J_{t-1}(x)+C_{t-1},|\mathcal{N}_t^{(l)}(x)|\geq n\gamma^{t-1}\Big)\\
\leq\;& P\Big(\bigcup_{a\in\mathcal{A}_{t-1}(x)}\Big\{|\mathcal{N}_t^{(l)}(x)|^{-1}\alpha\sum_{y\in\mathcal{N}_t^{(l)}(x,a)}\widehat{J}_t^{(l)}(y)\geq\alpha E[J_t(X_t(x,a))]+C_{t-1}\Big\},|\mathcal{N}_t^{(l)}(x)|\geq n\gamma^{t-1}\Big)\\
=\;& P\Big(\bigcup_{a\in\mathcal{A}_{t-1}(x)}\Big\{|\mathcal{N}_t^{(l)}(x)|^{-1}\alpha\Big[\sum_{y\in\mathcal{N}_t^{(l)}(x,a)\cap\bar{\mathcal{R}}_t^{(l)}}\widehat{J}_t^{(l)}(y)+\sum_{y\in\mathcal{N}_t^{(l)}(x,a)\cap\mathcal{R}_t^{(l)}}\widehat{J}_t^{(l)}(y)\Big]\\
&\qquad\geq\alpha E[J_t(X_t(x,a))]+C_{t-1}\Big\},|\mathcal{N}_t^{(l)}(x)|\geq n\gamma^{t-1}\Big)\quad\text{by (21)}\\
\leq\;& P\Big(\bigcup_{a\in\mathcal{A}_{t-1}(x)}\Big\{|\mathcal{S}|\gamma\widetilde{J}_0+|\mathcal{N}_t^{(l)}(x)|^{-1}\alpha\sum_{y\in\mathcal{N}_t^{(l)}(x,a)\cap\mathcal{R}_t^{(l)}}\widehat{J}_t^{(l)}(y)\geq\alpha E[J_t(X_t(x,a))]+C_{t-1}\Big\},\\
&\qquad|\mathcal{N}_t^{(l)}(x)|\geq n\gamma^{t-1}\Big)\quad\text{by (22)}\\
\leq\;& P\Big(\bigcup_{a\in\mathcal{A}_{t-1}(x)}\Big\{|\mathcal{S}|\gamma\widetilde{J}_0+|\mathcal{N}_t^{(l)}(x)|^{-1}\alpha\Big[\sum_{y\in\mathcal{N}_t^{(l)}(x,a)\cap\mathcal{R}_t^{(l)}}(J_t(y)+C_t)\Big]\\
&\qquad\geq\alpha E[J_t(X_t(x,a))]+C_{t-1}\Big\},|\mathcal{N}_t^{(l)}(x)|\geq n\gamma^{t-1}\Big)+|\mathcal{S}|D_t\quad\text{by (23)}\\
=\;& P\Big(\bigcup_{a\in\mathcal{A}_{t-1}(x)}\Big\{|\mathcal{S}|\gamma\widetilde{J}_0+\frac{|\mathcal{N}_t^{(l)}(x,a)\cap\mathcal{R}_t^{(l)}|}{|\mathcal{N}_t^{(l)}(x)|}\alpha C_t+|\mathcal{N}_t^{(l)}(x)|^{-1}\alpha\sum_{y\in\mathcal{N}_t^{(l)}(x,a)\cap\mathcal{R}_t^{(l)}}J_t(y)\Big]\\
&\qquad\geq\alpha E[J_t(X_t(x,a))]+C_{t-1}\Big\},|\mathcal{N}_t^{(l)}(x)|\geq n\gamma^{t-1}\Big)+|\mathcal{S}|D_t\\
\leq\;& P\Big(\bigcup_{a\in\mathcal{A}_{t-1}(x)}\Big\{|\mathcal{N}_t^{(l)}(x)|^{-1}\alpha\sum_{y\in\mathcal{N}_t^{(l)}(x,a)}J_t(y)\geq\alpha E[J_t(X_t(x,a))]+\delta\Big\},\\
&\qquad|\mathcal{N}_t^{(l)}(x)|\geq n\gamma^{t-1}\Big)+|\mathcal{S}|D_t\quad(\text{since }\delta=C_{t-1}-\alpha C_t-|\mathcal{S}|\gamma\widetilde{J}_0)\\
\leq\;& |\mathcal{A}|\exp(-n\gamma^{t-1}\delta')+|\mathcal{S}|D_t=D_{t-1}\quad\text{by Lemma A1}',
\end{aligned}\tag{24}$$

completing the induction. □

Similar to the proofs of Lemmas A2 and A3, we finish the proof of Lemma 3.2 by

establishing (15), recalling that $\mathcal{N}(a_l) = \mathcal{N}_2^{(l)}(x_1, a_l)$ and $|\mathcal{N}(a_l)| = n$:

$$
\begin{aligned}
& P\left(\bigcup_{l=1}^{m}\{\widehat{Q}(a_l) \geq Q(a_l) + \Delta_{\gamma,\delta}\}\right) \\
&= P\Big(\bigcup_{l=1}^{m}\Big\{\frac{1}{n}\sum_{y\in\mathcal{N}(a_l)}\alpha\widehat{J}_2^{(l)}(y) \geq \alpha E[J_2(X_2(x_1,a_l))] + C_1\Big\}\Big) \\
&\leq P\Big(\bigcup_{l=1}^{m}\Big\{\frac{1}{n}\sum_{y\in\mathcal{N}(a_l)}\alpha J_2(y) + \alpha C_2 + |\mathcal{S}|\gamma\widetilde{J}_0 \geq \alpha E[J_2(X_2(x_1,a_l))] + C_1\Big\}\Big) + |\mathcal{S}|D_2 \\
&\qquad \text{by (21), (22), (23)} \\
&= P\Big(\bigcup_{l=1}^{m}\Big\{\frac{1}{n}\sum_{y\in\mathcal{N}(a_l)}\alpha J_2(y) \geq \alpha E[J_2(X_2(x_1,a_l))] + \delta\}\Big\}\Big) + |\mathcal{S}|D_2 \\
&\leq |\mathcal{A}|\exp(-n\delta') + |\mathcal{S}|D_2 = D_1 = |\mathcal{A}|\left(\sum_{i=0}^{T-2}|\mathcal{S}|^i\exp(-n\gamma^i\delta')\right) \quad \text{using Lemma A1.} \quad \square
\end{aligned}
$$

Acknowledgements

This research was supported in part by the National Science Foundation under Grants DMI-9713720 and DMI-9988867, and by the Air Force Office of Scientific Research under Grant F496200110161. Xing Jin also acknowledges the support of National University of Singapore under Grant R-146-000-045-101.

References

[1] Arapostathis, A., V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh and S.I. Marcus, "Discrete-Time Controlled Markov Processes with Average Cost Criterion: A Survey," *SIAM Journal on Control and Optimization*, **31**, 282-344, 1993.

[2] Bertsekas, D.P., *Dynamic Programming and Optimal Control, Vol. 1 & 2*, Athena Scientific, 1995.

[3] Bertsekas, D.P., and J.N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996.

[4] Dai, L., "Convergence Properties of Ordinal Comparison in the Simulation of Discrete Event Dynamic Systems," *Journal of Optimization Theory and Applications,* **91**, 363-388, 1996.

[5] Dai, L., and C. Chen, "Rate of Convergence for Ordinal Comparison of Dependent Simulations in Discrete Event Dynamic Systems," *Journal of Optimization Theory and Applications,* **94**, 29-54, 1997.

[6] Dembo, A., and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd edition, Springer-Verlag, 1998.

[7] Gürkan, G., A.Y. Özge, and S.M. Robinson, "Sample-path Solution of Stochastic Variational Inequalities," *Mathematical Programming*, **84**, 313-333, 1999.

[8] Ho, Y.C., C.G. Cassandras, C.H. Chen, and L.Y. Dai, "Ordinal Optimization and Simulation," *Journal of Operations Research Society*, **51**, 490-500, 2000.

[9] Ho, Y.C., R. Sreenivas, and P. Vakili, "Ordinal Optimization of DEDS," *Discrete Event Dynamic Systems: Theory and Applications*, **2**, 61-88, 1992.

[10] Puterman, M.L., *Markov Decision Processes*, John Wiley & Sons, New York, 1994.

[11] Robinson, S.M., "Analysis of Sample Path Optimization," *Mathematics of Operations Research*, **21**, 513-528, 1996.

[12] Xie, X., "Dynamics and Convergence Rate of Ordinal Comparison of Stochastic Discrete-Event Systems," *IEEE Transactions on Automatic Control,* **42**, No. 4, 586-590, 1997.