RECURSIVE APPROACHES FOR SINGLE SAMPLE PATH BASED MARKOV REWARD PROCESSES

Hai-Tao Fang, Han-Fu Chen and Xi-Ren Cao

ABSTRACT

In this paper, two single sample path-based recursive approaches for Markov decision problems are proposed. One is based on the simultaneous perturbation approach and can be applied to the general state problem, but its convergence rate is low. In this algorithm, the small perturbation on current parameters is necessary to get another sample path for comparison, but it may worsen the system. Hence, we introduce another approach, which directly estimates the gradient of the performance for optimization by "potential" theory. This algorithm, however, is limited to finite state space systems, but its convergence speed is higher than the first one. The estimate for gradient can be obtained by using the sample path with current parameters without any perturbation. This approach is more acceptable for practical applications.

KeyWords: Markov decision processes, stochastic approximation, potential, recursive approach.

INTRODUCTION

The Markov decision processes (MDP) and the associated dynamic programming (DP) ([1,10]) methodology provide a general framework for posing and analyzing problems of sequential decision making under uncertainty. But there are two main difficulties associated with the standard DP approach:

1. "Curse of dimensionality". For the case that the state space is very large (or infinite), the computational requirements are overwhelming, if not impossible.
2. It requires the exact knowledge of the transition matrix, which may not be available for practical systems.

To solve the problems above, the single sample path based optimization techniques are good way for real systems. Many single sample path-based optimization approaches proposed in literature (e.g. [3,7] and [9]), including those based on perturbation analysis(PA) of discrete event systems, apply mainly to performance optimization with respect to continuous parameters. In this paper, we also concentrate on methods based on policy parameterization and gradient improvement.

Two recursive algorithms in this paper are proposed based on classical stochastic approximation methods: KW algorithm and RM algorithm. Since the dimension of the parameter is always very high, the one-sided randomized differences [12] are used for the KW algorithm. This algorithm can be applied to the general state problem, and the required conditions for its convergence are relatively weak. In this method the difference is applied to estimate differential. But this is not a good estimation and leads to a significant loss in convergence rate (from $O(n^{-1/2})$ to $O(n^{-1/3})$) in comparison with the case where a better estimate for differential is applied. For the denumber Markov chain, the gradient can be estimated directly by using "potential" given in [2]. This method is used and referred as the second algorithm in this paper.

The similar idea can be found in [7], where, however the estimate for the gradient of the performance is given by important sampling, so the transition probability of the Markov chain $p_{ij}(\theta)$ for all θ in decision space must be uniformly bounded away from 0, if it is not 0. This condition seems too restrict for many applications.

We state our main results in Section 2, and their proofs are given in Section 3. Section 4 gives some conclusions and remarks.

Manuscript received August 3, 2000; accepted November 7, 2000.

Hai-Tao Fang and Han-Fu Chen are with Laboratory of Systems and Control, Institute of Systems Science, Academy of Mathematics and Systems Sience, Chinese Academy of Sciences, Beijing, 100080.

Xi-Ren Cao is with Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology Clear Water Bay, Kowloon, Hong Kong.

This research was supported in part by the National Key Project of China and Chinese National Nature Science Foundation, and in part by Hong Kong UGC under grant HKUST.

II. MAIN RESULTS

We consider a controlled Markov chain $\{x_n^\theta, n = 0, 1, \ldots\}$, evolving on state space S, with Borel σ– algebra $B(S)$. The state space S is taken to be general, locally compact and separable metric space. The transition probabilities of the Markov chain $\{x_n^\theta\}$ depend on a parameter vector $\theta \in \mathbb{R}^d$, and denoted by

$$P(\theta, x, A) = P\{x_n^\theta \in A \,|\, x_{n-1}^\theta = x, \theta\}$$

for any $A \in B(S)$ and $x \in S$.

Whenever the state is equal to x, we receive a on-stage reward that also depends on θ, and is denoted by $f(x, \theta)$.

For any $\theta \in \mathbb{R}^d$, we assume that

C1) $\{x_n^\theta\}$ is irreducible.
C2) $\{x_n^\theta\}$ is Harris positive recurrent, i.e., there exists a positive, finite, invariant measure μ_θ such that for any A with $\mu_\theta(A) > 0$ and for all $x \in S$

$$P_x\left\{\sum_{n=1}^{\infty} 1_A(x_n^\theta) = \infty\right\} = 1.$$

C3) μ_θ is a probability measure, i.e., $\mu_\theta(S) = 1$ and $f(x, \theta) \in L_+^1(\mu_\theta)$.
C4) There exists a state $\alpha \in S$ such that for any $\theta \in \mathbb{R}^d$ and $x \in S$, $P_x\{T_\alpha(\theta) < \infty\} = 1$, where $T_\alpha(\theta) = \inf\{t > 0, x_t^\theta = \alpha\}$.

Remark 1. If S is of finite state and $\{x_n^\theta\}$ is irreducible, then any state of S can be taken as α in C4) and C1)-C4) hold.

We have that the average reward of the sample path is

$$\frac{1}{T}\sum_{n=0}^{T-1} f(x_n, \theta) \underset{T\to\infty}{\longrightarrow} E_{\mu_\theta} f(x_n, \theta) \triangleq \eta(\theta), \quad P_\nu - \text{a.s.}$$

for any probability measure ν., i.e., it is the same for almost all paths. Thus we can use $\eta(\theta)$ as the performance measure to compare different policies.

The single sample path based optimization is to find the optimal θ, i.e. $\theta^0 = \arg\min_{\theta \in \mathbb{R}^d} \eta(\theta)$, by using $\{x_n\}$ and $\{f(x_n, \theta_n)\}$, where θ_n is the parameter of the policy used at time n. To solve this problem, we use the stochastic approximation method.

Since $\nabla\eta(\theta_k)$ can not be observed directly from the sample path, it is important to obtain an estimate which can be observed on-line for a recursive algorithm. Let y_k be the kth estimate of $h_k(\theta_k) \triangleq \beta_k \nabla\eta(\theta_k)$ and

$$y_k = -\eta_k(\theta_k) + \epsilon_{k+1},$$

where β_k is positive, and ϵ_{k+1} is the observation noise. Then we can update the estimate of θ^0 recursively based on y_k as follows

$$\theta_{k+1} = (\theta_k + a_k y_{k+1}) 1_{[\|\theta_k + a_k y_{k+1}\| \le M_{\sigma_k}]} + \theta^* 1_{[\|\theta_k + a_k y_{k+1}\| > M_{\sigma_k}]}, \tag{1}$$

$$\sigma_k = \sum_{i=0}^{k-1} 1_{[\|\theta_i + a_i y_{i+1}\| > M_{\sigma_i}]}, \quad \sigma_0 = 0, \tag{2}$$

where θ^* is a point in $\mathbb{R}^d$ given later and $\{M_k\}$ is a sequence of positive increasing numbers diverging to infinity.

The following conditions are used:

H1) $\nabla\eta(\theta)$ is locally Lipschitz continuous;
H2) There exist h_0 and θ^* such that $\inf_{\|\theta\| = h_0} \eta(\theta) > \eta(\theta^*)$ and $\eta(J)$ is nowhere dense, where $J = \{\theta \in \mathbb{R}^d, \nabla\eta(\theta) = 0\}$.

2.1 Simultaneous perturbation approach

For a Markov chain in general state space with accessible state ([8]), we use a simultaneous perturbation gradient approximation method ([12,5]) to estimate $h_k(\theta_k)$.

We assume the sample path starts with $x_0 = \alpha$. Otherwise, we simply discard the initial period from x_0 to the first state $x_k = \alpha$. Let $\{x_n\}$ be a sample path of the corresponding Markov chain. Let t_k be the time of the kth visit to the accessible state α. We refer to the sequence $x_{t_k}, x_{t_k+1}, \ldots, x_{t_{k+1}}$ as the kth renewal cycle.

Let $\{\Delta_k^i, i = 1, \ldots, d, k \in \mathbb{N}\}$ be i.i.d. r.v. sequences with $|\Delta_k^i| < a$, $E(1/\Delta_k^i) = 0$ and $|1/\Delta_k^i| < b$, where $a, b > 0$. Denote by $\Delta_k = [\Delta_k^1 \ \ldots \ \Delta_k^d]^\tau$,

$$g_k = \left[\frac{1}{\Delta_k^1} \ \cdots \ \frac{1}{\Delta_k^d}\right]^\tau. \tag{3}$$

Driving the Markov chain from t_{2k} to t_{2k+1} under the parameter θ_k, and from t_{2k+1} to t_{2k+2} under the parameter $\theta_k + c_k\Delta_k$, we obtain the following two observations

$$F_k^0(\theta_k) = u_{2k+2} \sum_{i=t_{2k}+1}^{t_{2k+1}} f(x_i, \theta_k), \tag{4}$$

$$F_k^+(\theta_k) = u_{2k+1} \sum_{i=t_{2k+1}+1}^{t_{2k+2}} f(x_i, \theta_k + c_k\Delta_k),$$

$$u_{k+1} = t_{k+1} - t_k. \tag{5}$$

Let

$$y_{k+1} = \frac{F_k^0(\theta_k) - F_k^+(\theta_k)}{2c_k} g_k. \tag{6}$$

Then, we have the following result:

Theorem 1. If C1)-C4), H1)-H2) hold, and $a_k > 0$, $c_k > 0$, $c_k \to 0$, $\Sigma_k a_k = \infty$ and $\Sigma_k \frac{a_k^2}{c_k^2} < +\infty$, then

$$d(\theta_k, J) \to 0,$$

where θ_k is defined by (1)(2) with y_{k+1} given in (6).

2.2 Potential based recursive method

In the algorithms given in section 2.1, the differences are used to approximate differentials, and this will influence the convergence rate of the algorithms as mentioned before. By potential theory ([2]), when the Markov chain $X = \{x_0, x_1, \ldots\}$ is in a finite state space $S = \{1, 2, \ldots, M\}$, we can on-line construct an observation of the gradient of the performance. Let i^* be the initial state, i.e. assume $x_0 = i^*$. The sample path is then divided into "basic periods" by the successive occurrence of i^*'s on the path. We denote the sample path as $x_{t_0}, \ldots, x_{t_1}, \ldots, x_{t_k}, \ldots, x_{t_{k+1}}, \ldots$, where $t_0 = 0$, $x_0 = i^*$, $t_{k+1} = \min\{n : n > t_k, x_n = i^*\}$, $k \geq 0$. The kth basic period is $x_{t_k}, \ldots, x_{t_{k+1}-1}$.

Define the estimate y_{k+1} for $h_k(\theta_k)$ to be used in the algorithms (1) (2) as

$$y_{k+1} = -\sum_{n=t_k}^{t_{k+1}-1} \sum_{j=1}^{M} \nabla p_{x_n j}(\theta_k) \hat{d}_k(j), \tag{7}$$

where

$$\hat{d}_k(j) = \begin{cases} d_k(j), & \text{if } t_k(j) < t_k, \\ \hat{d}_{k-1}(j), & \text{if } \sigma_{k-1} \neq \sigma_k,\ t_k(j) \geq t_k, \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$d_k(j) = \sum_{l=t_k^j}^{t_k} [u_k f(x_l, \theta_{k-1}) - \tilde{\eta}_k] \tag{9}$$

$$\tilde{\eta}_{k+1} = \sum_{l=t_{k-1}}^{t_k - 1} f(x_l, \theta_{k-1}), \tag{10}$$

$$t_k(j) = \inf\{n > t_{k-1}, x_n = j\}. \tag{11}$$

Theorem 2. If H1), H2) hold, and $a_k > 0$, $\Sigma_k a_k = \infty$, $a_{k+1} - a_k = o(a_k)$ and $\Sigma_k a_k^2 < +\infty$, then

$$d(\theta_k, J) \underset{k\to\infty}{\to} 0,$$

where J is given in H2) and θ_k is given by Algorithms(1) (2) with y_{k+1} given in (7).

III. PROOFS

Proof of Theorem 1. Note that

$$E(F^+_{k+1}(\theta) \mid \theta_k = \theta) = C_k(\theta)\eta(\theta + c_k\Delta_k),$$

$$E(F^0_{k+1}(\theta) \mid \theta_k = \theta) = C_k(\theta)\eta(\theta),$$

where $C_k(\theta) = E_\theta(u_{2k+1})E_{\theta + c_k\Delta_k}(u_{2k+2})$.

Denote

$$\xi^+_{k+1}(\theta, \omega) = F^+_{k+1}(\theta) - E(F^+_{k+1}(\theta_k) \mid \theta_k = \theta),$$

$$\xi^0_{k+1} = F^0_{k+1}(\theta) - E(F^0_{k+1}(\theta_k) \mid \theta_k = \theta),$$

By (6) it follows that

$$y_{k+1} = -C_k(\theta_k)\frac{\eta(\theta_k + c_k\Delta_k) - \eta(\theta_k)}{c_k} g_k + \frac{\xi^+_{k+1} - \xi^0_{k+1}}{c_k} g_k.$$

By the Taylor expansion and the way similar to that used in [5] and [4], we can show that

$$y_{k+1} = -C_k(\theta_k)\nabla\eta(\theta_k) + \varepsilon_{k+1},$$

and that

$$\limsup_{k\to\infty} \left\| \sum_{i=n_k}^{m(n_k, t)} a_i \varepsilon_{i+1} 1_{\{\|\theta_k\| < N\}} \right\| = o(T)$$

for any $N > 0$ and $t \in [0, T]$, where $m(n, T) = \inf\{k > n, \Sigma_{i=n}^k > T\}$.

Using Theorem 1 in [4], we complete the proof of the theorem. ■

To prove Theorem 2, we prove the following lemma first.

Lemma 1. For any compact set K, there exist constants C_K and ρ_K such that

$$P_\theta\{u_m = l\} \leq C_K \rho_K^l.$$

In particular, $E_\theta(u_m)$ and $E_\theta(u_m^2)$ are bounded functions in bounded domain of θ.

Proof. Let $x_n(\omega)\Big|_{t_k}^{t_{k+1}-1}$ be a path of the Markov Chain in finite state from t_k to $t_{k+1} - 1$. Then, by the cycle decomposition [11], the Markov chain can uniquely be decomposed into several cycles. In each cycle, no state is repeated. We separate all cycles with finite number of states into two groups A and B such that each cycle in A includes state i^* while no cycle in B includes state i^*, Clearly, for any compact set K

$$\max_{\theta \in K} P_\theta\{\text{cycle } C \in B\} < 1.$$

Since if it were not true, then there would exist a $\theta \in K$ such that $P_\theta\{\text{cycle } C \in B\} = 0$. This implies that i^* is a transit state, which is impossible by assumption. Thus,

$$\epsilon_K \triangleq \max_{\theta \in K} P_\theta\{\text{cycle } C \in B\} < 1,$$

and we have

$$P_\theta\{T = l\} \le P_\theta\{\text{There exist at least } \left[\frac{l}{N}\right] + 1 \text{ cycles and no cycle belongs to A}\}$$

$$< \sum_{k=\left[\frac{l}{N}\right]+1}^{\infty} \epsilon_K^k = \frac{\epsilon_K^{\left[\frac{l}{N}\right]+1}}{1-\epsilon_K} < C_K \rho_K^l$$

by taking $\rho_K = \epsilon_K^N$. ■

Define

$$\lambda(\theta) = E_\theta\{1_{\{t(j) < t(i^*)\}} \mid x_0 = i^*\},$$

where $t(j) = \inf\{n > 0, x_n = j\}$.

Lemma 2. $\lambda(\theta)$ is locally Lipschitz continuous.

Proof. Note that

$$\lambda(\theta) = P_\theta(t(j) < t(i^*) \mid x_0 = i^*\},$$

is a taboo transition probability [6]. By the properties of taboo transition probability, $\lambda(\theta)$ can be expressed by

$$\lambda(\theta) = \frac{m_{i^*i^*}(\theta)}{m_{ji^*}(\theta) + m_{i^*j}(\theta)}, \qquad (12)$$

where $m_{i^*i^*}(\theta) = E^\theta t(i^*)$, $m_{ji^*}(\theta) = E^\theta\{t(j) \mid x_0 = i^*\}$ and $m_{i^*j}(\theta) = E^\theta\{t(i^*) \mid x_0 = j\}$. Define $h(j) = [1, \ldots, 1, 0, 1, \ldots, 1]\tau$. Then for fixed j, $\dfrac{m_{ji^*}(\theta)}{\pi_j(\theta)}$ is the solution to $(I - P^\theta + e\pi^\theta)g = h(j)$, where $e = (1, 1, \ldots, 1)\tau$ is an M-dimensional column vector with all components being 1. Thus $m_{ji^*}(\theta)$ is locally Lipschitz continuous, since $\pi_j(\theta)$ is locally Lipschitz continuous. Similarly, $m_{i^*j}(\theta)$ is also locally Lipschitz continuous. Since $m_{i^*j}(\theta) > 0$, $m_{ji^*}(\theta) > 0$ if $i^* \neq j$, and $\pi_{i^*}(\theta) > 0$ for any θ, then by (12) it follows that $\lambda(\theta)$ is locally Lipschitz continuous. ■

Let $F_k = \sigma\{x_l, 1 = 0, 1, \ldots, k\}$.

Lemma 3.

$$E(\hat{d}_{k+1}(j) \mid F_{t_k}) = u_k g^{\theta_k}(j)(1 - \lambda(\theta_k)) + \lambda(\theta_k)\hat{d}_k(j)1_{\{\sigma_k = \sigma_{k-1}\}}$$
$$+ (u_k \eta^{\theta_k} - \tilde{\eta}_k)C^{\theta_k}(j),$$

where

$$g^\theta(j) = E^\theta\{\sum_{n=0}^{t(i^*)-1}(f(X_n, \theta) - \eta(\theta)) \mid x_0 = j\},$$

$$C^\theta(j) = E^\theta\{t(j)1_{\{t(j) < t(i^*)\}}\}.$$

Proof. From (8)-(11) we have

$$E(\hat{d}_{k+1}(j) \mid F_{t_k}) - (u_{k+1}\eta^{\theta_k} - \tilde{\eta}_k)C^{\theta_k}(j)$$
$$= u_k E\{d_{k+1}(j)1_{\{t_k(j) < t_{k+1}\}} \mid F_{t_k}\} + \lambda(\theta_k)\hat{d}_k(j)1_{\{\sigma_k = \sigma_{k-1}\}}$$
$$= u_k g^{\theta_k}(j)(1 - \lambda(\theta_k)) + \lambda(\theta_k)\hat{d}_k(j)1_{\{\sigma_k = \sigma_{k-1}\}}.$$

The last equality is from that

$$E\{d_{k+1}(j)1_{\{t_k(j) < t_{k+1}\}} \mid F_{t_k}\} = E\{E[d_{k+1}(j) \mid F_{t_{kj}}]1_{\{t_k(j) < t_{k+1}\}} \mid F_{t_k}\}$$
$$= g^{\theta_k}(j)(1 - \lambda(\theta_k)).$$ ■

Proof of Theorem 2. Note that

$$h_k(\theta) \triangleq Eu_{k+1}Eu_k \sum_{i,j} \pi_i(\theta)\nabla p_{ij}(\theta)d^\theta(j)$$
$$= Eu_{k+1}Eu_k \nabla\eta(\theta).$$

By Theorem 1 of [4], we need check that for any convergent subsequence of $\{\theta_k\}$ and any $N > 0$ $t \in [0, T]$

$$\limsup_{k \to \infty} \left\| \sum_{l=n_k}^{m(n_k, t)} a_k(y_{k+1} + h_k(\theta_k))1_{\{\|\theta_k\| < N\}} \right\| = o(T). \quad (13)$$

Let $F_k(j) = -\Sigma_{n=t_k}^{t_{k+1}-1} \nabla p_{x_n j}(\theta_k)$. Note that

$$y_{k+1} + h_k(\theta_k) = \sum_{j=1}^{M}(F_k(j) - E\{F_k(j) \mid F_k\})\hat{d}_k(j)$$
$$+ \sum_{j=1}^{M} E\{F_k(j) \mid F_k\}\hat{d}_k(j) + h_k(\theta_k) \quad (14)$$

By Lemma 1, we have

$$E\hat{d}_k^2(j)1_{\{\|\theta_k\| < N\}} \le \sup_{\|\theta_k\| \le N} Ed_k^2(j) < \infty.$$

This yields

$$\sum_{k=1}^{\infty} a_k^2 \hat{d}_k^2(j) 1_{\{\|\theta_k\|<N\}} < +\infty, \text{ a.e.}$$

since

$$E\sum_k a_k^2 \hat{d}_k^2(j) 1_{\{\|\theta_k\|<N\}} = \sum_k a_k^2 E\hat{d}_k^2(j) 1_{\{\|\theta_k\|<N\}} < +\infty.$$

Then, by the martingale convergence theorem we have

$$\limsup_{k\to\infty} \left\| \sum_{l=n_k}^{m(n_k,t)} a_l \xi_l(j) \hat{d}_l(j) 1_{\{\|\theta_k\|<N\}} \right\| = 0,$$

where $\xi_l(j) = F_l(j) - E\{F_l(j) \mid F_{t_l}\}$.

Note that

$$E\{F_k(j) \mid F_{t_k}\} = Eu_{k+1} \sum_j \pi^{\theta_k}(j) \nabla p_{ij}(\theta_k).$$

Thus, to prove (13), we only need to show that for any $t \in [0, T]$

$$\lim_{k\to\infty} \left| \sum_{l=n_k}^{m(n_k,t)-1} a_l \left(\hat{d}_l(j) - Eu_l g^{\theta_l}(j) \right) \right| = o(T),$$

Note that

$$\left| \sum_{l=n_k}^{m(n_k,t)-1} a_l \left(\hat{d}_l(j) - Eu_l g^{\theta_l}(j) \right) \right|$$

$$\leq \left| \sum_{l=n_k}^{m(n_k,t)-1} a_l \left(\hat{d}_l(j) - Eu_l g^{\theta_{l-1}}(j) \right) \right|$$

$$+ \left| \sum_{l=n_k}^{m(n_k,t)-1} a_l (g^{\theta_l}(j) - g^{\theta_{l-1}}(j)) Eu_l \right|$$

$$+ \left| \sum_{l=n_k}^{m(n_k,t)-1} a_l g^{\theta_l}(j)(Eu_l - Eu_{l-1}) \right|, \tag{15}$$

where the last two terms are of $o(T)$, since $a_k y_{k+1} \xrightarrow[k\to\infty]{} 0$. Therefore, it suffices to show that the first term on the right hand side of (15) is of $o(T)$.

In what follows, let $\lambda_k = \lambda(\theta_k)$. Then

$$\hat{d}_l(j) - Eu_l g^{\theta_{l-1}}(j) = \frac{1}{1-\lambda_{l-1}} (\hat{d}_l(j) - E(\hat{d}_l(j) \mid F_{t_l})) - Eu_l g^{\theta_{l-1}}(j)$$

$$+ \frac{1}{1-\lambda_{l-1}} E(\hat{d}_l(j) \mid F_{t_l}) + \frac{\lambda_{l-1}}{1-\lambda_{l-1}} \hat{d}_l(j)$$

By Lemma 3, we have

$$\frac{1}{1-\lambda_{l-1}} E(\hat{d}_l(j) \mid F_{t_l}) - u_l g^{\theta_{l-1}}(j)$$

$$= -\frac{\lambda_{l-1}}{1-\lambda_{l-1}} \hat{d}_{l-1}(j) + (u_l \eta^{\theta_l} - \tilde{\eta}_l) C^{\theta_l}(j).$$

So, the first term on the right hand side of (15) is

$$\left| \sum_{l=n_k}^{m(n_k,t)-1} a_l \left(\frac{\lambda_{l-1}}{1-\lambda_{l-1}} - \frac{\lambda_l}{1-\lambda_l} \right) \hat{d}_l(j) \right|$$

$$+ \left| \sum_{l=n_k}^{m(n_k,t)} a_l (t_l \eta^{\theta_l} - \tilde{\eta}_l) C^{\theta_l}(j) \right| + o(T),$$

which is of $o(T)$ by Lemma 2. The assertion of the theorem is proved. ■

IV. CONCLUDING REMARKS

In this paper, two stochastic approximation methods solving the Markov decision problem are presented, and the system parameters are optimized based on the observations of the sample path of the system. These methods are useful when a complex system should be improved online.

There are two directions which one can work with further: one is to extend "potential" to some general state problems so that the gradient of the performance can be estimated directly from the sample path; another one is to optimize the system when only the partial information of the system is known. This situation occurs in the Semi-Markov decision problems and in the "state aggregation'' problems as well.

REFERENCES.

1. Bertsekas, D.P., *Dynamic Programming and Optimal Control*, Athena Scientific, Belmont, MA, Vol. 1 and 2. (1995).
2. Cao, X.R. and H.F. Chen, "Perturbation Realization, Potentials, and Sensitivity Analysis of Markov Processes," *IEEE Trans. Automat. Contr.*, Vol. 42, pp. 1382-1393 (1997).
3. Cao, X.R. and Y.W. Wan, "Algorithms for Sensitivity Analysis of Markov Systems Through Potentials and Perturbation Realization," *IEEE Trans. Contr. Syst. Technol.*, to appear.
4. Chen, H.F, "Stochastic Approximation with State-

dependent Noise," Technical Report (1999).
5. Chen, H.F., T.E. Duncan and B.Pasik-Duncan, "A Kiefer-Wolfowitz Algorithm with Randomized Differences," *IEEE Trans. Automat. Contr.*, Vol. 44, pp. 442-453 (1999).
6. Karlin, S. and H.M. Taylor, *A Second Course in Stochastic Processes*, Academic Press, New York (1981).
7. Marbach, P. and J.N. Tsitsiklis, "Simulation-based Optimization of Markov Reward Processes," Technical Report, Laboratory for Information and Decision Systems, MIT (1997).
8. Meyn, S.P., "The Policy Iteration Algorithm for Average Reward Markov Decision Processes with General State Space," *IEEE Trans. Automat. Contr.*, Vol. 42, pp. 1663-1679 (1997).
9. Plambeck, E.L., B.R. Fu, S.M. Robinson and R. Suri., "Sample-path Optimization of Convex Stochastic Performance Functions," *Math. Program.*, to appear.
10. Puterman, M.L., *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York (1994).
11. Qian, M.P. and M. Qian, "Circulation for Recurrent Markov Chains," *Z. Wahrscheinlichkeits Verw. Geb.*, Vol, 59, pp. 203210 (1982).
12. Spall, J.C., "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation," *IEEE Trans. Automat. Contr.*, Vol. 37, pp. 332-341 (1992).

Hai-Tao Fang received his B.S. degree in probability and statistics in 1990, M.S. degree in applied mathematics in 1993, and Ph.D. degree in 1996, respectively, from the Peking University, Tsinghua University, and Peking University. He now is with the Laboratory of Systems and Control, Institute of Systems Science, the Chinese Academy of Sciences as an Associate Professor.

From 1996-1998, he was a Postdoc at the Institute of Systems Science and joined the Institute as an Assistant Professor in 1998. During 1998-1999, he was with the Hong Kong University of Science and Technology as a Research Associate. His research interests include stochastic optimization and systems control, Markov desicion processes and its applications.

Han-Fu Chen obtained the Diploma from the Department of the Mathematics and Mechanics, Leningrad (St. Petersburg) University, Russia in 1961. After graduation, he joined the Institute of Mathematics, and then the Institute of Systems Science, Chinese Academy of Sciences where he is a Professor of the Laboratory of System and Control. His research interests include stochastic systems, system identification, adaptive control, recursive estimation, and stochastic approximation and it applications. He has authored and coauthored more than 140 papers.

He was elected an Academician of the Chinese Academy of Sciences in 1993 and an IEEE Fellow in 1997. He now serves as the President of the Chinese Association of Automation and the Technical Board of IFAC.

Xi-Ren Cao received the M.S. and Ph.D. degrees from Harvard University, in 1981 and 1984, respectively, where he was a research fellow from 1984 to 1986. He then worked as a principal and consultant engineer/engineering manager at Digital Equipment Corporation, U.S.A. until October 1993. Since then, he is a Professor of the Hong Kong University of Science and Technology (HKUST). He is the director of the Center for Networking at HKUST. He held visiting positions at Harvard University, University of Massachusetts at Amberst, AT&T Labs, University of Maryland at College Park, Shanghai Jiaotong University, Nankei University, and University of Science and Technology of China.

Xi-Ren owns two patents in data communications and published two books: "Realization Probabilities, the Dynamics of Queuing Systems," Springer Verlag, 1994, and "Perturbation Analysis of Discrete-Event Dynamic Systems," Kluwer Academic Publishers, 1991 (co-authored with Y.C. Ho). He received the Outstanding Transactions Paper Award from the IEEE Control Systems Society in 1987 and the Outstanding Publication Award from the Institution of Management Science in 1990. He is a Fellow of IEEE, Associate Editor at large of IEEE Transactions of Automatic Control, and he is/was Board of Governors of IEEE Control Systems Society, associate editor of a number of international journals and chairman of a few technical committees of international professional societies. His current research areas include discrete event systems theory, communication systems, signal processing, stochastic processes, and optimization techniques.