Available online at www.sciencedirect.com
SCIENCE DIRECT®
Automatica 40 (2004) 1749–1759
automatica
www.elsevier.com/locate/automatica

Brief paper

A unified approach to Markov decision problems and performance sensitivity analysis with discounted and average criteria: multichain cases☆

Xi-Ren Cao[a,*], Xianping Guo[b]

[a] *Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*
[b] *Zhongshan University, Guangzhou, PR China*

Received 11 April 2003; received in revised form 13 March 2004; accepted 5 May 2004

Abstract

We propose a unified framework to Markov decision problems and performance sensitivity analysis for multichain Markov processes with both discounted and average-cost performance criteria. With the fundamental concept of performance potentials, we derive both performance-gradient and performance-difference formulas, which play the central role in performance optimization. The standard policy iteration algorithms for both discounted- and average-reward MDPs can be established using the performance-difference formulas in a simple and intuitive way; and the performance-gradient formulas together with stochastic approximation may lead to new optimization schemes. This sensitivity-based point of view of performance optimization provides some insights that link perturbation analysis, Markov decision processes, and reinforcement learning together. The research is an extension of the previous work on ergodic Markov chains (Cao, Automatica 36 (2000) 771).

Keywords: Policy iteration; Potentials; Perturbation analysis; Performance sensitivity; Reinforcement learning

1. Introduction

Recently, it was discovered that in the area of learning and optimization of Markov systems, performance sensitivity analysis (PSA), Markov decision problems (MDPs), and reinforcement learning (RL) are closely related (Cao, 1998a, 2000, 2003). At the center of the subjects are the two performance sensitivity formulas, one for performance gradients in continuous parameter spaces, and the other for performance differences in discrete "policy" spaces. Both of them depend on the fundamental concept of performance *potentials* (or equivalently, *perturbation realization* (Cao, 1994; Cao & Chen, 1997), in the terminology of perturbation analysis (PA) (Cao, 1994; Ho & Cao, 1991). These two sensitivity formulas can be explained and derived in a simple and intuitively clear way by applying the PA principles (Cao, accepted). The standard policy iteration algorithms are the natural consequence of the performance difference formulas (Cao, 1998a, 2000), and the performance-gradient formulas together with stochastic approximation may lead to new optimization schemes (Baxter & Bartlett, 2001; Cao, 1999; Cooper, Henderson, & Lewis, 2003; Marbach & Tsitsiklis, 2001).

This sensitivity view of optimization provides a unified framework for PSA and MDPs with both infinite horizon discounted and average-reward performance criteria (Cao, 1998a, 2000, 2003). In particular, both the discounted- and average-reward performance cases are treated in the same way with the average case corresponding to the discount factor α being one.

All the existing results stated above are for ergodic chains. The goal of this paper is to extend the results in Cao (1998a, 2000, 2003) to *multichain* Markov processes. That is, we

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Ioannis Paschalidis under the direction of Editor Ian Petersen. Supported by a grant from Hong Kong UGC. The second author was also supported by the Natural Science Foundations of China and Guangdong Province and by Zhongshan University Advanced Research Center, China.

* Corresponding author. Tel.: +852-2358-7048; fax: +852-2358-1485.
E-mail addresses: eecao@ust.hk (X.-R. Cao), mcsgxp@zsu.edu.cn (X. Guo).

0005-1098/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2004.05.003

propose a unified formulation for PSA and MDPs for both infinite horizon discounted- and average-reward performance criteria for Markov processes with multichain structures, and we show that the above statements on ergodic chains also hold for the multichain case.

In Section 2, we briefly review the results for ergodic chains developed in Cao (2000), which will become special cases of the results presented in this paper. This section serves as a load map for the rest of the paper. In Section 3, we define performance potentials and derive performance-difference formulas for problems with discounted and average performance criteria for multichain Markov processes. In Section 4, we show that the standard policy iteration algorithms for the both problems can be derived using the performance-difference formulas in a clear and intuitive way. In Section 5, we discuss the performance gradients for Markov chains whose transition probability matrices depend on continuous parameters. Performance optimization can be implemented using these gradient formulas together with stochastic approximation algorithms. Both the performance-gradient and performance-difference formulas are based on the performance potentials of one of the Markov chains. Section 6 concludes the paper with some discussions.

It is straightforward to extend the derivation of policy iteration to the discounted performance problems with multichains using the performance difference formula, but this is not true for the average-reward problems. The crucial point in the derivation for the average-reward case is an observation based on the canonical form of the transition probability matrix that the effect of the transient and recurrent states on the performance difference can be "decoupled". With this observation, we can easily establish the optimality equations for multichain MDPs with the average-reward criterion, construct the policy iteration algorithm, and prove its convergence to the optimal policy in a finite number of iterations.

It was brought to the authors' attention in the reviewing process of an earlier version of this paper that the algebraic derivation in our proof of convergence of the policy iteration algorithm for the average-reward case follows the same ideas of the early work of Veinott (1966) about the canonical structure of the transition probability matrix (see also Section 9.2.4 of Puterman, 1994). However, our presentation with the performance difference framework is more intuitive and clearer than Veinott's paper (Veinott, 1966). We also emphasize the uniform applicability of our approach to the discounted- and average-reward problems and performance sensitivity analysis.

2. Results for ergodic chains

We first briefly review the basic concepts and results about the sensitivity-based approach to performance optimization with an ergodic model (Cao, 2000).

Consider an irreducible and aperiodic (hence ergodic) Markov chain on a finite state space $S = \{1,2,\dots,M\}$ with a transition probability matrix $P = [p(j\,|\,i)]$. Let $\pi = (\pi(1),\dots,\pi(M))$ be the (row) vector representing its steady-state probabilities, and $r = (r(1),\dots,r(M))^{\mathrm{T}}$ be a (column) reward vector, where "T" denotes transpose. Then $Pe = e$, where $e = (1,\dots,1)^{\mathrm{T}}$ is an M-dimensional vector whose components are all equal to 1, and $\pi e = 1$. The steady-state probability flow balance equation is $\pi = \pi P$. Let α, $0 < \alpha \leqslant 1$, be a discount factor. Let $\{X_0, X_1, \dots, X_n, \dots\}$ denote a sample path of the Markov chain. The discounted-reward performance criterion is defined as a column vector $\eta_\alpha = (\eta_\alpha(1), \eta_\alpha(2), \dots, \eta_\alpha(M))^{\mathrm{T}}$ with:

$$\eta_\alpha(i) = (1-\alpha)E\left\{\sum_{n=0}^{\infty}\alpha^n r(X_n)\,|\,X_0 = i\right\} \quad \forall i \in S.$$

The average reward performance is $\eta = \pi r$. We have $\lim_{\alpha\to 1-}\eta_\alpha = \eta e$ (Blackwell, 1962; Cao, 2000). The α-*potential* is defined by the *discounted Poisson equation* (Cao, 2000):

$$(I - \alpha P + \alpha e\pi)g_\alpha = r. \tag{1}$$

When $\alpha = 1$, (1) is the *Poisson equation*, and $g := g_1$ is simply called the *potential*.

Let $\tilde{P}$ and $\tilde{\pi}$ be another ergodic transition probability matrix and its steady-state probability defined on the same state space S, respectively. Let $\tilde{r}$, $\tilde{\eta}$, and $\tilde{\eta}_\alpha$ be the reward vector, the average, and discounted performance criteria for the system with $\tilde{P}$, respectively. Then we have

$$\tilde{\eta}_\alpha - \eta_\alpha = (1-\alpha)(I - \alpha\tilde{P})^{-1}\{[\tilde{r} + \alpha\tilde{P}g_\alpha] - [r + \alpha P g_\alpha]\}, \quad 0 < \alpha < 1, \tag{2}$$

$$\tilde{\eta} - \eta = \tilde{\pi}\{[\tilde{r} + \tilde{P}g] - [r + Pg]\}. \tag{3}$$

Now, suppose that P changes to $P(\delta) = P + \delta Q = \delta\tilde{P} + (1-\delta)P$, and r changes to $r(\delta) = r + \delta h$, with $Q = \tilde{P} - P$, $h = \tilde{r} - r$, and $\delta \in [0,1]$. Then the performance measure changes to $\eta(\delta) = \eta + \Delta\eta(\delta)$. From (2) and(3), we can easily get

$$\frac{\mathrm{d}\eta_\alpha}{\mathrm{d}\delta} = (1-\alpha)(I - \alpha P)^{-1}[\alpha Q g_\alpha + h], \quad 0 < \alpha < 1, \tag{4}$$

$$\frac{\mathrm{d}\eta}{\mathrm{d}\delta} = \pi(Qg + h). \tag{5}$$

Formulas (2)–(5) are two sets of performance sensitivity formulas: (2) and (3) for performance differences and (4) and (5) for performance gradients. These two sets of sensitivity formulas are fundamental for performance optimization. Many optimization schemes originate from them. For example, the policy iteration algorithm (for the ergodic case with average reward) of MDPs can be easily derived from (3) (Cao, 1998a). Roughly speaking, since $\tilde{\pi} > 0$ component wisely, if $Qg + h = (\tilde{P} - P)g + (\tilde{r} - r) \geqslant 0$ component wisely, then we have $\tilde{\eta} \geqslant \eta$. Policy iteration essentially uses

this fact to find a policy that has a better performance than the current one.

The sensitivity point of view of performance optimization provides a simple and intuitive way to prove the results for policy iteration algorithms. It links policy iteration naturally with the gradient-based optimization approaches. A direct comparison of (2) and (4), (3) and (5) shows that in policy iteration one simply chooses the direction with the steepest gradients as the improved policy in the next iteration (Cao, 1998a). This approach brings in some new insights and new research topics (see, e.g., Baxter & Bartlett, 2001; Cao, 1999; Cooper et al., 2003). In a recent paper (Cao, accepted), it is shown that both performance-gradient and performance-difference formulas can be constructed by using perturbation realization (Cao, 1994; Cao & Chen, 1997), or equivalently, performance potentials, as building blocks. With this view, we can derive performance gradient and difference formulas for many optimization problems that cannot be formulated as the standard MDPs. With these sensitivity formulas, we can further develop policy iteration algorithms and gradient-based optimization schemes, using the structural properties of the sensitivity formulas. In this paper, we extend the above results with the sensitivity-based approach to the multichain case.

3. Performance differences

3.1. Performance criteria

We study the infinite horizon performance with discounted- and average-reward criteria for discrete-time Markov chains. Let $S=\{1,2,\dots,M\}$ be the state space and $P=[p(j\,|\,i)]$, $i,j\in S$, be the transition probability matrix. Denote a sample path as $\{X_0,X_1,\dots,X_n,\dots\}\in\Omega:=(S)^{\infty}$, with $X_n\in S$ being the system state at time $n\geqslant 0$. For any initial state $i\in S$ and a given transition matrix P, by the Kolmogoroff theorem there exists a unique probability space $\{\Omega,\mathscr{F},P_i\}$ such that, for any sequence $i_j\in S$, $j=1,2,\dots,n$, we have

$$P_i(X_0=i,X_1=i_1,\dots,X_n=i_n)$$
$$=p(i_1\,|\,i)\,p(i_2\,|\,i_1)\dots p(i_n\,|\,i_{n-1}).$$

We denote by E_i the corresponding expectation operator. We assume that all operators, such as limit on matrices or vectors, etc., are component-wise; and we denote by "0" the matrix and the vector with zero as all of their components.

Let $r(i)$ $(i\in S)$ be a reward function, and $\alpha\in(0,1)$ be a discount factor. Then, the discounted- and average-performance criteria are defined as column vectors η_α and η, respectively. Their ith components $\eta_\alpha(i)$ and $\eta(i)$ are given as

$$\eta_\alpha(i):=(1-\alpha)E_i\left\{\sum_{n=0}^{\infty}\alpha^n r(X_n)\right\},\quad 0<\alpha<1,\ \text{and}\quad (6)$$

$$\eta(i):=\lim_{N\to\infty}\frac{E_i[\sum_{n=0}^{N-1}r(X_n)]}{N},\quad (7)$$

respectively. The weighting factor $(1-\alpha)$ in (6) is added for maintaining the continuity of η_α at $\alpha=1$, see (11) below.

For any given transition matrix P, let P^* be the Cesaro-limit defined as

$$P^*:=\lim_{N\to\infty}\frac{\sum_{n=0}^{N-1}P^n}{N}.\quad (8)$$

Lemma 1. *Let I be the identity matrix, $0<\alpha<1$. Then*

(a) *$PP^*=P^*P=P^*P^*=P^*$, $P^*e=e$, and $\eta=P^*r$.*
(b) *The matrices $(I-P+P^*)$, $(I-\alpha P)$, and $(I-\alpha P+\alpha P^*)$ are all nonsingular.*
(c) *$(I-\alpha P)^{-1}=(I-\alpha P+\alpha P^*)^{-1}+(\alpha/(1-\alpha))P^*$.*

Proof. Part (a) follows directly from (7) and (8). Theorems A.7 and C.2 in Puterman (1994) gives part (b). Next, by parts (a) and (b), a straightforward calculation gives

$$I-\alpha P+\alpha P^*=I-\alpha P+\alpha(I-\alpha I+\alpha P^*)P^*.$$

Thus,

$$I=(I-\alpha P+\alpha P^*)^{-1}(I-\alpha P)+\alpha P^*$$
$$=(I-\alpha P+\alpha P^*)^{-1}(I-\alpha P)+\frac{\alpha}{1-\alpha}P^*(I-\alpha P),$$

which, together with part (b), yields part (c). □

From Lemma 1(b) and (c), we have

$$\lim_{\alpha\to 1^-}(1-\alpha)(I-\alpha P)^{-1}=P^*.\quad (9)$$

From (6), we have

$$\eta_\alpha=(1-\alpha)\sum_{n=0}^{\infty}\alpha^n P^n r=(1-\alpha)(1-\alpha P)^{-1}r.\quad (10)$$

From (10), (9), and Lemma 1(a), we have

$$\lim_{\alpha\to 1^-}\eta_\alpha=\eta.\quad (11)$$

Thus, a discount factor $\alpha=1$ corresponds to the case of average performance criteria.

3.2. Performance potentials

Similar to the ergodic case (Cao, 2000), the *α-potential* is defined as

$$g_\alpha:=(I-\alpha P+\alpha P^*)^{-1}r,\quad 0<\alpha\leqslant 1.\quad (12)$$

By Lemma 1(b), the inverse exists and g_α is well defined on $\alpha\in(0,1]$. In addition, $g:=g_1$ is simply called the *potential*. By Lemma 1(b) we can easily prove

$$\lim_{\alpha\to 1^-}g_\alpha=g.$$

Lemma 2.

(a) $P\eta = \eta$, *i.e.*, $\sum_{j\in S} p(j\,|\,i)\eta(j) = \eta(i)$ *for each* $i \in S$.
(b) η *and* g *are a unique solution to the following two equations*

$$\eta + g = r + Pg, \tag{13}$$

$$\eta = P^* g. \tag{14}$$

(c) $\eta_\alpha = (1-\alpha)g_\alpha + \alpha\eta$.

Proof. Lemma 1(a) gives part (a). Next, by Lemma 1(a) and (b), we have

$$(I-P) = (I-P^*)(I-P+P^*),$$

$$P^* = P^*(I-P+P^*), \text{ and}$$

$$(I-P)(I-P+P^*)^{-1} = I - P^*.$$

From these equations and Lemma 1(a) and (12) with $\alpha = 1$, we obtain (13) and (14). On the other hand, suppose that x and y satisfy (13) and(14). Then by Lemma 1(b) and (12), we get

$$y = (I-P+P^*)^{-1} r = g,$$

which gives $P^* y = P^* r$. Moreover, by (14) and Lemma 1(a) we have $x = P^* y = P^* r = \eta$. This shows the uniqueness of the solution to (13) and (14). Finally, by (10) and Lemma 1, part (c) holds. □

Eq. (13) is the *Poisson equation*. Its solution is unique up to an additive constant; i.e., if g satisfies (13), so does $g + ce$ for any constant c. In this sense, (14) is used to normalize the potential g.

3.3. Performance differences

Suppose that the transition matrix P and the performance function r change to $\tilde{P}$ and $\tilde{r}$, defined on the same state space S, respectively. Let $\tilde{\eta}_\alpha$ and $\tilde{\eta}$ be the discounted and average performance criteria associated with $\tilde{P}$ and $\tilde{r}$, respectively.

Theorem 1.

(a) *For the discounted performance criterion* $(0 < \alpha < 1)$, *we have*

$$\tilde{\eta}_\alpha - \eta_\alpha = (I-\alpha\tilde{P})^{-1}[(1-\alpha)(\tilde{r}-r) + \alpha(\tilde{P}-P)\eta_\alpha] \tag{15}$$

$$= (1-\alpha)(I-\alpha\tilde{P})^{-1}[(\tilde{r}+\alpha\tilde{P}g_\alpha) - (r+\alpha P g_\alpha)] + \alpha^2(I-\alpha\tilde{P})^{-1} \times (\tilde{P}-I)\eta. \tag{16}$$

(b) *For the average criterion, we have*

$$\tilde{\eta} - \eta = \tilde{P}^*[(\tilde{r}+\tilde{P}g) - (r+Pg)] + (\tilde{P}^* - I)\eta, \tag{17}$$

$$(I-\tilde{P})(\tilde{g}-g) = (\tilde{r}+\tilde{P}g) - (r+Pg) - (\tilde{\eta}-\eta). \tag{18}$$

Proof. (a) By (10), we have

$$(I-\alpha P)\eta_\alpha = (1-\alpha)r,$$

which gives

$$\tilde{\eta}_\alpha - \eta_\alpha = (1-\alpha)(\tilde{r}-r) + \alpha(\tilde{P}\tilde{\eta}_\alpha - P\eta_\alpha) = (1-\alpha)(\tilde{r}-r) + \alpha(\tilde{P}-P)\eta_\alpha + \alpha\tilde{P}(\tilde{\eta}_\alpha - \eta_\alpha).$$

This is

$$(I-\alpha\tilde{P})(\tilde{\eta}_\alpha - \eta_\alpha) = (1-\alpha)(\tilde{r}-r) + \alpha(\tilde{P}-P)\eta_\alpha \tag{19}$$

and so (15) follows. Multiplying both sides of (19) by $(I-\alpha\tilde{P})^{-1}$ leads to

$$\tilde{\eta}_\alpha - \eta_\alpha = (1-\alpha)(I-\alpha\tilde{P})^{-1}(\tilde{r}-r) + \alpha(I-\alpha\tilde{P})^{-1}(\tilde{P}-P)\eta_\alpha,$$

which together with Lemma 2(c) and (a) yields (16).

(b) By Lemma 1(a) and (13), we have

$$\tilde{\eta} - \eta = \tilde{P}^*\tilde{r} - \eta = \tilde{P}^*\tilde{r} + \tilde{P}^* g - \tilde{P}^* g - \tilde{P}^*\eta + \tilde{P}^*\eta - \eta = \tilde{P}^*[\tilde{r} + \tilde{P}g - g - \eta] + (\tilde{P}^* - I)\eta = \tilde{P}^*[(\tilde{r}+\tilde{P}g) - (r+Pg)] + (\tilde{P}^* - I)\eta.$$

This is (17). Next, by (13), we have $\tilde{g} = \tilde{P}\tilde{g} + \tilde{r} - \tilde{\eta}$ and $g = Pg + r - \eta$. Thus,

$$\tilde{g} - g = \tilde{P}\tilde{g} - \tilde{P}g + \tilde{P}g + \tilde{r} - \tilde{\eta} - Pg - r + \eta = \tilde{P}(\tilde{g}-g) + (\tilde{r}+\tilde{P}g) - (Pg+r) - (\tilde{\eta}-\eta),$$

which gives (18). □

These equations lead to the fundamental results in MDPs and PSA, which will be discussed in the following sections. Results for ergodic chains (Cao, 2000, 2003) become special cases.

4. Markov decision problems

In Markov decision problems, there is an action space denoted by A, which we assume to be finite. At any state $i \in S$ at time $n \geqslant 0$, an action a_i is taken from an available action set $A(i) \subset A$. The transition probability from state i to state $j \in S$ depends on a_i and is denoted by $p(j\,|\,i, a_i)$. Also, the reward function is denoted by $r(i, a_i)$. Let F be the set of all decision rules f with $f(i) \in A(i)$ for all $i \in S$. Then F is called the (stationary) policy space. Thus, for a given policy $f \in F$, the Markov chain $X^f := \{X_n^f, n \geqslant 0\}$ evolves according to the transition matrix $P^f := [p(j\,|\,i, f(i))]$, and the corresponding reward function is defined as r^f with

$r^f(i) := r(i, f(i))$ for all $i \in S$. We will use the index f to indicate the quantities associated with policy f, such as η_α^f, η^f and g_α^f, etc.

For two vectors u and v, we define $u = v$ if $u(i) = v(i)$ for all $i \in S$; $u \geqslant v$ if $u(i) \geqslant v(i)$ for all $i \in S$; $u \geqslant v$ if $u \geqslant v$ and $u(i) > v(i)$ for at least one $i \in S$.

A policy $f^* \in F$ is called *α-discounted optimal* if $\eta_\alpha^{f^*}(i) \geqslant \eta_\alpha^f(i)$ for all $f \in F$ and all $i \in S$; f^* is called *average-optimal* if $\eta^{f^*}(i) \geqslant \eta^f(i)$ for all $f \in F$ and all $i \in S$.

Most of the results in this section have appeared in literature (see, e.g., Blackwell, 1962; Federgruen & Shweitzer, 1984; Guo & Shi, 2001; Guo, Yu, & Li, 2002; Hastings, 1969; Howard, 1960; Kallenberg, 2002; Ng, 1999; Puterman, 1994; Shweitzer & Brower, 1987; Shweitzer & Federgruen, 1978; Spreen, 1985; Veinott, 1966). We show that these results can be obtained from the performance difference equations in a simple and direct way.

4.1. MDPs with discounted performance criterion

We show that the standard policy iteration algorithm for MDPs with discounted performance criterion follows easily from the performance difference formula (15) (or (16)). To this end, for each $f \in F, i \in S$, $a \in A(i)$, and $0 < \alpha < 1$, we let

$$G_\alpha^f(i,a) := (1-\alpha) r(i,a) + \alpha \sum_{j \in S} p(j \mid i,a) \eta_\alpha^f(j) \tag{20}$$

and define an action set $B_\alpha^f(i)$ as

$$B_\alpha^f(i) := \{a \in A(i) \mid G_\alpha^f(i,a) > G_\alpha^f(i, f(i))\}. \tag{21}$$

Then we define a policy $h \in F$ (depending on f) as follows:

$$h(i) \in B_\alpha^f(i) \quad \text{when } B_\alpha^f(i) \neq \emptyset,$$
$$\text{and } h(i) = f(i) \text{ when } B_\alpha^f(i) = \emptyset. \tag{22}$$

Theorem 2. *For any given $f \in F$, let h be defined as in (22). If $h \neq f$, then $\eta_\alpha^h \geqslant \eta_\alpha^f$.*

Proof. Let the ith components of vectors $G_\alpha^f(h)$ and $G_\alpha^f(f)$ be $G_\alpha^f(i, h(i))$ and $G_\alpha^f(i, f(i))$ for all $i \in S$, respectively. Let $S_\alpha^1 := \{i \in S \mid B_\alpha^f(i) \neq \emptyset\}$, $S_\alpha^2 := \{i \in S \mid B_\alpha^f(i) = \emptyset\} = \{i \in S - S_\alpha^1\}$. Then, by (22), we have $h(i) = f(i)$ for all $i \in S_\alpha^2$. Therefore, by (21) and (22) we get

$$G_\alpha^f(i, h(i)) > G_\alpha^f(i, f(i)) \quad \forall i \in S_\alpha^1, \tag{23}$$

$$G_\alpha^f(i, h(i)) = G_\alpha^f(i, f(i)) \quad \forall i \in S_\alpha^2, \tag{24}$$

which imply (S_α^1 is not empty since $h \neq f$)

$$G_\alpha^f(h) \geqslant G_\alpha^f(f) \quad \text{when } h \neq f. \tag{25}$$

Next, by (15) and (20), we have

$$\eta_\alpha^h - \eta_\alpha^f = (I - \alpha P^h)^{-1} [G_\alpha^f(h) - G_\alpha^f(f)]. \tag{26}$$

Since $(I - \alpha P^h)^{-1} = I + \sum_{n=1}^\infty \alpha^n (P^h)^n \geqslant I$, by (25) and (26), we have $\eta_\alpha^h \geqslant \eta_\alpha^f$. □

Following the same argument as for the ergodic case (Cao, 2000), from Theorem 2, we can obtain the standard *policy iteration algorithm*:

1. Set $n = 0$ and select an arbitrary decision rule $f_0 \in F$.
2. (Policy evaluation) Obtain $\eta_\alpha^{f_n}$ by using (10).
3. (Policy improvement) Set $f = f_n$ in (21) and obtain f_{n+1} as the policy h in (22).
4. If $B_\alpha^{f_n}(i) = \emptyset$ for all $i \in S$, then stop and f_n is optimal. Otherwise, increment n by 1 and return to step 2.

The above policy iteration algorithm is based on the performance difference formula (15). It is easy to see that the performance difference formula (16) also leads naturally to a policy iteration algorithm that uses g_α and η. This algorithm is essentially the same as the standard one; it, however, may provide some insights that may link to the policy iteration procedure of the average performance MDPs discussed in Section 4.3. We will not get into more details.

4.2. The canonical forms

We first review some results related to the canonical forms of P and P^*. Let $C_k \subset S$, $k = 1, 2, \ldots, m$, denote all the disjoint closed irreducible sets of the recurrent states of a Markov chain $\mathbf{X}$ with the transition probability matrix P, with m being the number of such sets; and let C_{m+1} be the set of transient states. First, it is well known (see, e.g., Puterman, 1994) that by reordering the states, P takes the *canonical form*

$$P = \begin{bmatrix} P_1 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & P_2 & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & P_m & 0 \\ Q_1 & Q_2 & \cdot & \cdot & Q_m & Q_{m+1} \end{bmatrix}, \tag{27}$$

in which P_k corresponds to the transitions among states in $C_k, k=1,2,\ldots,m$; $Q_k, k=1,2,\ldots,m$, to the transitions from the transient states in C_{m+1} to the recurrent states in C_k, $k=1,\ldots,m$; and Q_{m+1} to the transitions among the transient states in C_{m+1}. Next, P^* takes the following form Puterman (1994):

$$P^* = \begin{bmatrix} P_1^* & 0 & 0 & \cdot & \cdot & 0 \\ 0 & P_2^* & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & P_m^* & 0 \\ Q_1^* & Q_2^* & \cdot & \cdot & Q_m^* & 0 \end{bmatrix}, \tag{28}$$

in which $P_k^* = e_k\pi_k$, where π_k is the steady-state probability (row) vector of P_k obtained by $\pi_k P_k = \pi_k$ subject to $\pi_k e_k = 1$, e_k is a column vector of ones (with the same dimension as π_k), and $Q_k^* := (I - Q_{m+1})^{-1} Q_k P_k^*$.

By (27) and (28), we have the following simple observations which will be used to prove the existence of optimal policies for MDPs with the average criterion.

Lemma 3. *Let P be a transition probability matrix of a Markov chain* **X** *on S, and u be a vector on S.*

(a) *If $(P^*u)(i) > 0$ (or $(P^*u)(i) < 0$) for some state $i \in S$, then there exists a recurrent state of* **X**, *denoted as $j \in S$, such that $u(j) > 0$ (or $u(j) < 0$).*
(b) *Suppose $P^*u = 0$ and $u \leqslant 0$ (or $u \geqslant 0$). If $u(i) < 0$ (or $u(i) > 0$) for some $i \in S$, then i is a transient state of* **X**.
(c) *Suppose $P^*u = 0$ and $u \leqslant 0$ (or $u \geqslant 0$), then $u(i) = 0$ for all recurrent states i of* **X**.

Proof. From the canonical form (28), the columns in P^* corresponding to transient states in C_{m+1} are all zeros; thus, all $u(j)$s with $j \in C_{m+1}$ contribute nothing to P^*u. Moreover, since all the entries in P_k^*, $k=1,2,\ldots,m$, are positive, parts (a) and (b) follow. (c) is simply another statement of (b). □

Next, for given $r, \tilde{r}, P$ and $\tilde{P}$, define

$$w := (\tilde{r} + \tilde{P}g) - (r + Pg), \quad \Delta g := \tilde{g} - g. \tag{29}$$

We further write $\tilde{P}$ in the canonical form (27) and partition w and Δg according to it, i.e., denote

$$\tilde{P} = \begin{bmatrix} \tilde{P}_1 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & \tilde{P}_2 & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \tilde{P}_{\tilde{m}} & 0 \\ \tilde{Q}_1 & \tilde{Q}_2 & \cdot & \cdot & \tilde{Q}_{\tilde{m}} & \tilde{Q}_{\tilde{m}+1} \end{bmatrix}, \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_{\tilde{m}+1} \end{bmatrix},$$

$$\Delta g = \begin{bmatrix} \Delta g_1 \\ \Delta g_2 \\ \cdot \\ \cdot \\ \cdot \\ \Delta g_{\tilde{m}+1} \end{bmatrix}, \tag{30}$$

where $\tilde{m}$ is the number of ergodic classes under $\tilde{P}$. Note that P and $\tilde{P}$ may have different closed-subset structures (e.g., it is possible that $m \neq \tilde{m}$). We have the following lemma.

Lemma 4. *If $\tilde{\eta} = \eta$, then*

$$\Delta g_{\tilde{m}+1} = (I - \tilde{Q}_{\tilde{m}+1})^{-1} \left\{ w_{\tilde{m}+1} + \sum_{k=1}^{\tilde{m}} \tilde{Q}_k \Delta g_k \right\}. \tag{31}$$

Proof. By (18) and (29) we have

$$\Delta g = w + \tilde{P}\Delta g. \tag{32}$$

Noting that $(I - \tilde{Q}_{\tilde{m}+1})^{-1}$ exists (by Proposition A.3 in Puterman, 1994). By (30) we can solve (32) for $\Delta g_{\tilde{m}+1}$ and get (31). □

This lemma will be used later to prove the anti-cycling rule for the policy iteration procedure in MDPs.

4.3. MDPs with the average performance criterion

In the ergodic case, with the formula for the performance difference (3), because $\tilde{\pi} > 0$, we know that if $Qg + h \geqslant 0$ or equivalently $\tilde{P}g + \tilde{r} \geqslant Pg + r$, then $\tilde{\eta} > \eta$. That is, although we don't know $\tilde{\pi}$, but by comparing $\tilde{P}g + \tilde{r}$ with $Pg + r$, we may know $\tilde{\eta}$ is indeed larger than η. This is the basis for policy iteration.

To extend the above simple observation to the multichain case, we use the performance difference formula in Theorem 1(b). However, there are two terms on the right-hand side. This causes a major problem in extending the above results. Fortunately, as the following example indicates, these two

terms can be "de-coupled". First, we set $u^* = (\tilde{P}^* - I)\eta$ and rewrite (17) as

$$\tilde{\eta} - \eta = \tilde{P}^* w + u^*. \tag{33}$$

Example 1. Let $S := \{1,2,3,4,5\}$, $r =: (5,2,1,3,1)^{\mathrm{T}}$, $\tilde{r} := (4,1,1,2,0)^{\mathrm{T}}$, and

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 & 0 \\ 0 & 0 & 0.7 & 0.3 & 0 \\ 0.1 & 0.2 & 0.2 & 0.3 & 0.2 \end{bmatrix},$$

$$\tilde{P} = \begin{bmatrix} 0.9 & 0.1 & 0 & 0 & 0 \\ 0.8 & 0.2 & 0 & 0 & 0 \\ 0.2 & 0.4 & 0.1 & 0.2 & 0.1 \\ 0.2 & 0.1 & 0.2 & 0.3 & 0.2 \\ 0.3 & 0.1 & 0.2 & 0.1 & 0.3 \end{bmatrix}.$$

After some calculation, we can write the performance difference (33) as

$$\tilde{\eta} - \eta = \tilde{P}^* w + u^*$$

$$= \begin{bmatrix} 0.8889 & 0.1111 & 0.0000 & 0.0000 & 0.0000 \\ 0.8889 & 0.1111 & 0.0000 & 0.0000 & 0.0000 \\ 0.8889 & 0.1111 & 0.0000 & 0.0000 & 0.0000 \\ 0.8889 & 0.1111 & 0.0000 & 0.0000 & 0.0000 \\ 0.8889 & 0.1111 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix}$$

$$\times \begin{bmatrix} 0.3333 \\ 0.3333 \\ -0.2650 \\ -0.5534 \\ -7.7077 \end{bmatrix} + \begin{bmatrix} 0.0000 \\ 0.0000 \\ 2.0832 \\ 2.0832 \\ 1.3019 \end{bmatrix}.$$

Observe the following structure: the components in the second term, $u^*(i)$, for recurrent states of $\tilde{P}$ ($i = 1$ and 2) are zeros; the components in $w(i)$ in the first term for recurrent states of $\tilde{P}$ are all positive, and the columns of $\tilde{P}^*$ for transient states are all zeros. This "decouples" the effect of the two terms: the components of $\tilde{\eta} - \eta$ for the recurrent states are determined by the first term in the performance difference formula; and the components of $\tilde{\eta} - \eta$ for the transient states take additional contribution from the second terms in the performance difference formula. The negative values in the components of w for the transient states do not play a role. □

The above example gives us an idea that we may compare the performance difference between two policies by using their structures; this simple observation leads to the optimality equation and the policy iteration algorithm. We formally state the results below. (The rest of this section is similar to Veinott, 1966.)

Lemma 5. *Suppose that $\tilde{P}$, $\tilde{r}$ and P, r correspond to two Markov chains with the average performance measure $\tilde{\eta}$ and η, and satisfy the following two conditions*:

(a) $\tilde{P}\eta \geqslant \eta$, *and*
(b) $\tilde{r}(i) + (\tilde{P}g)(i) \geqslant r(i) + (Pg)(i)$ *when* $(\tilde{P}\eta)(i) = \eta(i)$ *for some* $i \in S$.

Then $\tilde{\eta} \geqslant \eta$. This lemma also holds if we change all of the symbols "$\geqslant$" to "$\leqslant$".

Proof. Let $u = \tilde{P}\eta - \eta \geqslant 0$ and $w = (\tilde{r} + \tilde{P}g) - (r + Pg)$. By Lemma 1(a), we have $\tilde{P}^* u = 0$. Then, By Lemma 3(b), we have $u(i) = 0$ for all recurrent states i under $\tilde{P}$, and so it follows from condition (b) that $w(i) \geqslant 0$ for all recurrent states under $\tilde{P}$. Thus, from the canonical form (30) for $\tilde{P}$, we have $\tilde{P}^* w \geqslant 0$. On the other hand, since $\tilde{P}\eta \geqslant \eta$, and so $\tilde{P}^k \eta \geqslant \eta$ for all $k \geqslant 1$. Therefore, by (8) we get $\tilde{P}^* \eta \geqslant \eta$. Finally, by (33), we have $\tilde{\eta} - \eta = \tilde{P}^* w + (\tilde{P}^* - I)\eta \geqslant 0$. This proves the lemma. □

From Lemma 5, we can easily derive the optimality conditions:

Theorem 3. *Let $\hat{\eta}$ and $\hat{g}$ be the average performance measure and the potential with respect to policy $\hat{f} \in F$ (i.e. $\hat{\eta} := \eta^{\hat{f}}, \hat{g} := g^{\hat{f}}$). Suppose the following "optimality conditions" hold*:

$$\hat{\eta}(i) = \max_{a \in A(i)} \left\{ \sum_{j \in S} p(j \mid i,a)\hat{\eta}(j) \right\} \quad \forall i \in S, \tag{34}$$

$$\hat{\eta}(i) + \hat{g}(i) = \max_{a \in B(i)} \left\{ r(i,a) + \sum_{j \in S} p(j \mid i,a)\hat{g}(j) \right\} \quad \forall i \in S, \tag{35}$$

where $B(i) := \{a \in A(i) \mid \sum_{j \in S} p(j \mid i,a)\hat{\eta}(j) = \hat{\eta}(i)\}$. Then $\hat{\eta} \geqslant \eta^f$ for all $f \in F$; that is, policy $\hat{f}$ is average optimal.

Proof. This is a direct consequence of Lemma 5. Let $\hat{P}$, $\hat{r}$, $\hat{\eta}$ and $\hat{g}$ be the P, r, η and g in Lemma 5, respectively; and for any $f \in F$ we let P^f, r^f, η^f and g^f be the $\tilde{P}$, $\tilde{r}$, $\tilde{\eta}$ and $\tilde{g}$ in Lemma 5, respectively. Then (34) means that $P^f \hat{\eta} \leqslant \hat{\eta}$; and (35) together with (13) means that $r^f(i) +$

$(P^f \hat{g})(i) \leqslant \hat{r}(i) + (\hat{P}\hat{g})(i)$, whenever $(P^f \hat{\eta})(i) = \hat{\eta}(i)$. Then Lemma 5 (with relation $\leqslant$) implies $\eta^f \leqslant \hat{\eta}$. □

The goal of MDPs is to find a policy that satisfies the above optimality conditions. This can be achieved in policy iteration by improving performance at each iteration. We will show that for any non-optimal policy we can always construct a "better" policy according to Lemma 5. For a given $f \in F$, $i \in S$ and $a \in A(i)$, let

$$H^f(i,a) := r(i,a) + \sum_{j \in S} p(j \mid i,a) g^f(j) \tag{36}$$

and

$$A^f(i) := \left\{ a \in A(i) : \begin{array}{l} \sum_{j \in S} p(j \mid i,a)\eta^f(j) > \eta^f(i); \text{ or} \\ H^f(i,a) > H^f(i, f(i)) \\ \text{when } \sum_{j \in S} p(j \mid i,a)\eta^f(j) = \eta^f(i) \end{array} \right\}. \tag{37}$$

We then define an improvement policy $h \in F$ (depending on f) as follows:

$h(i) \in A^f(i)$ when $A^f(i) \neq \emptyset$, and

$h(i) = f(i)$ if $A^f(i) = \emptyset$. (38)

Note that such a policy may not be unique, since there may be more than one action in $A^f(i)$ for some state $i \in S$. Let

$$u_f^h := P^h \eta^f - \eta^f, \quad v_f^h := r^h + P^h g^f - r^f - P^f g^f. \tag{39}$$

Theorem 4. *For any given $f \in F$, let h be defined as in (38). Then*

(a) *$\eta^h \geqslant \eta^f$, and $v_f^h(i) \geqslant 0$ for all recurrent states i under P^h.*
(b) *If $v_f^h(i) > 0$ for some recurrent state i under P^h, then $\eta^h \geqslant \eta^f$.*
(c) *If $P^h \eta^f \neq \eta^f$, then $\eta^h \geqslant \eta^f$.*
(d) *If $\eta^h = \eta^f$ and $h \neq f$, then $g^h \geqslant g^f$.*

With Theorem 4 (see Appendix A for its proof), we can state the (standard) *Policy Iteration Algorithm* as follows:

1. Set $n = 0$ and select an arbitrary decision rule $f_0 \in F$.
2. (Policy evaluation) Obtain (by Lemma 2) g^{f_n} and η^{f_n}.
3. (Policy improvement) Obtain policy f_{n+1} as the policy h in (37) and (38).
4. If $f_{n+1} = f_n$, then stop and f_{n+1} is optimal (by Theorem 5 below). Otherwise, increment n by 1 and return to step 2.

Theorem 4 can be used to compare the performance of two policies and to prove the anti-cycling property in the policy iteration procedure. The existence of the solution to the optimality equations can be proved by construction as shown in Theorem 5.

Theorem 5. *The Policy Iteration Algorithm stops at an average optimal policy in a finite number of iterations.*

Proof. By Theorem 4(a), we have $\eta^{f_{n+1}} \geqslant \eta^{f_n}$. That is, as n increases, η^{f_n} either increases or stays the same. Furthermore, by Theorem 4(d), when η^{f_n} stays the same, g^{f_n} increases. Thus, any two policies in the sequence of f_n, $n = 0, 1, \ldots$, either have different performance measures or have different potentials. Thus, every policy in the iteration sequence is different. Since the number of policies is finite, the iteration must stop after a finite number of iterations. Suppose it stops at a policy denoted as $\hat{f}$. Then $\hat{f}$ must satisfy the optimality conditions (34) and (35), because otherwise for some i the set $A^{\hat{f}}(i)$ in (37) is non-empty and we can find the next improved policy in the policy iteration. Thus, by Theorem 3, policy $\hat{f}$ is average optimal. □

5. Performance sensitivity analysis

Now we turn to performance derivatives. Consider a (multichain) transition probability matrix $P(\delta)$ and a reward function $r(\delta)$ that depend on a parameter $\delta \in [0, 1]$. We assume that all the components of both $P(\delta)$ and $r(\delta)$ are (right) differentiable at $\delta = 0$, and denote these derivatives as $P'(0)$ and $r'(0)$, respectively. All quantities associated with $P(\delta)$ and $r(\delta)$ are obviously functions of the parameter $\delta \in [0, 1]$. Therefore, for example, the discounted and average criteria and the potentials are denoted as $\eta_\alpha(\delta)$, $\eta(\delta)$, and $g(\delta)$, respectively. The derivatives of these functions at $\delta = 0$ are viewed as right derivatives (i.e., $\delta \to 0+$).

Theorem 6.

(a) *For $0 < \alpha < 1$, we have*

$$\begin{aligned} \frac{d\eta_\alpha(\delta)}{d\delta} &= (I - \alpha P(0))^{-1}[(1-\alpha)r'(0) + \alpha P'(0)\eta_\alpha(0)] \\ &= (1-\alpha)(I - \alpha P(0))^{-1}[\alpha P'(0) g_\alpha(0) + r'(0)] \\ &\quad + \alpha^2 (I - \alpha P(0))^{-1} P'(0)\eta(0). \end{aligned}$$

(b) *If $P(0)$ is unichain, then*

$$\frac{d\eta(\delta)}{d\delta} = P^*(0)[P'(0)g(0) + r'(0)].$$

Proof. (a) Set $P(\delta) = \tilde{P}, r(\delta) = \tilde{r}, P(0) = P$ and $r(0) = r$. By Theorem 1(a) we get

$$\begin{aligned} \eta_\alpha(\delta) - \eta_\alpha(0) &= (I - \alpha P(\delta))^{-1}[(1-\alpha)(r(\delta) - r(0)) \\ &\quad + \alpha(P(\delta) - P(0))\eta_\alpha(0)] \end{aligned}$$

$$= (1-\alpha)(I-\alpha P(\delta)^{-1})[(r(\delta)+\alpha P(\delta)g_\alpha(0)) - (r(0)+\alpha P(0)g_\alpha(0))] + \alpha^2(I-\alpha P(\delta))^{-1} \times (P(\delta)-P(0))\eta(0),$$

Dividing both sides with δ and letting $\delta \to 0+$ yield part (a).

(b) If $P(0)$ is unichain, then all components of $\eta(0)$ is a constant number and so $(P(\delta)^* - I)\eta(0) = 0$. Thus, by Theorem 1(b) we get

$$\eta(\delta) - \eta(0) = P^*(\delta)[(P(\delta)-P(0))g(0) + r(\delta) - r(0)]. \quad (40)$$

On the other hand, we have $\lim_{\delta\to 0^+} P^*(\delta) = P^*(0)$. To prove this fact, we choose any arbitrary sequence $\{P^*(\delta_k)\}$ with a limit point G. This is, $P^*(\delta_k) \to G$ as $\delta_k \to 0$. Since $P^*(\delta_k)P(\delta_k) = P^*(\delta_k)$, letting $\delta_k \to 0$, we have $GP(0) = G$. This implies that all rows of G are the same and are equal to the unique stationary distribution of $P(0)$. Since the sequence $\{P^*(\delta_k)\}$ is chosen arbitrarily, we conclude that $\lim_{\delta\to 0^+} P^*(\delta) = P^*(0)$. Thus, by (40), part (b) holds. □

Note that $\lim_{\delta\to 0^+} P^*(\delta) = P^*(0)$ may not hold when $P(0)$ is a multichain (see Cao, 1998b). In that case, the performance derivative for average criterion does not exist.

6. Discussions

We have proposed a unified framework to Markov decision problems and performance sensitivity analysis for multichain Markov processes with both discounted- and average-reward performance criteria. With the fundamental concept of performance potentials, we derived both performance-gradient and performance-difference formulas, which play the central role in performance optimization. In particular, using the performance difference formula, we established policy iteration algorithms for both discounted- and average-reward MDPs. The performance-gradient formulas can be used with stochastic approximation to carry out performance optimization. This leads to the subject of "policy gradients" in the reinforcement learning literature. Previous work on ergodic Markov chains become special cases.

A distinguished feature of our approach is its *simplicity*, *clarity*, *and uniformity*. All the results are based on the two sets of sensitivity formulas. This approach to performance optimization is intuitively clear; it treats the average-reward case in the same way as the discounted performance case.

Sample-path-based estimates for performance potentials g can be derived and can be used in policy iteration and gradient based-optimization; for the ergodic case, see (Baxter & Bartlett, 2001; Cao, 1999; Cao & Wan, 1998; Cooper et al., 2003), among others. The sample-path-based approach is also called learning in literature, because the decision making is based on the information learned from the system behavior. Further research is needed for sample-path-based estimates of potentials for the multichain case.

Appendix A

The Proof of Theorem 4. (a) We take P^f, r^f, P^h and r^h as P, r, $\tilde{P}$ and $\tilde{r}$ in Lemma 5, respectively. Then by the construction in (37)–(39), conditions (a) and (b) in Lemma 5 hold. Thus, it follows from Lemma 5 that $\eta^h \geqslant \eta^f$. Moreover, as in the proof of Lemma 5, we have $v_f^h(i) \geqslant 0$ for all recurrent states i under P^h; thus, part (a) follows.

(b) Since we have $P^h\eta^f \geqslant \eta^f$, so $P^{h*}\eta^f \geqslant \eta^f$. Thus, under the condition in (b), by part (a) we have $P^{h*}v_f^h \geqslant 0$, and so $\eta^h - \eta^f = P^{h*}v_f^h + (P^{h*} - I)\eta^f \geqslant 0$. Then part (b) follows.

(c) By (a), it suffices to prove that $\eta^h \neq \eta^f$. Suppose that $\eta^h = \eta^f$. Then by Lemma 1(a) we have

$$P^h\eta^f = P^h\eta^h = \eta^h = \eta^f, \quad (A.1)$$

which contradicts to the given conditions. Therefore, part (c) is proved.

(d) We first prove that $g^h \geqslant g^f$. In fact, since $\eta^h = \eta^f$, (A.1) holds. Then, by Theorem 1(b) and (37)–(39), we have $(I - P^h)(g^h - g^f) = v_f^h \geqslant 0$. By Lemma 1(a), $P^{h*}(I - P^h)(g^h - g^f) = 0$. Therefore, by Lemma 3(c) we have $(I - P^h)(g^h - g^f)(i) = 0$ for all recurrent states i under P^h. Let

$$\tilde{P} := P^h \quad \text{and} \quad P := P^f$$

and partition $\tilde{P}$ and any vector such as η^h, η^f, g^h, g^f and $\Delta g := g^h - g^f$ as (30). Then, we have $(I - \tilde{P}_k)\Delta g_k = 0$, and so $\Delta g_k = \tilde{P}_k^* \Delta g_k$, for all $k = 1, 2, \ldots, \tilde{m}$. Since $\tilde{P}_k$ are all closed and irreducible, Δg_k is a constant vector denoted by $\rho_k e_k$, for all $k = 1, 2, \ldots, \tilde{m}$. By parts (a) and (b) we have $v_f^h(i) = 0$ for all recurrent states i under P^h. By (38) and (A.1) we have $h(i) = f(i)$ for all such recurrent states i, and so

$$p(j \mid i, h(i)) = p(j \mid i, f(i)) \text{ for all } j \in S$$

and recurrent states i under P^h,

which implies that $\tilde{P}_k$ is also closed and irreducible under P^f. Therefore, $P = P^f$ has the following form:

$$P = \begin{bmatrix} \tilde{P}_1 & 0 & 0 & \cdot & \cdot & 0 & 0 \\ 0 & \tilde{P}_2 & 0 & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \tilde{P}_{\tilde{m}} & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & P_m & 0 \\ Q_1 & Q_2 & \cdot & \cdot & \cdot & Q_m & Q_{m+1} \end{bmatrix},$$

where, $m \geqslant \tilde{m}$, m is the number of disjoint closed irreducible sets under $P = P^f$. By (14) and the fact that $\eta^h = \eta^f$ (so, $\eta^h_k = \eta^f_k$), we have

$$\tilde{P}^*_k g^h_k = \tilde{P}^*_k(g^f_k + \Delta g_k) = \tilde{P}^*_k(g^f_k + \rho_k e_k)$$
$$= \eta^h_k = \eta^f_k = \tilde{P}^*_k g^f_k \ \forall k = 1, \ldots, \tilde{m},$$

which gives $\rho_k = 0$, and so $\Delta g_k = \rho_k e_k = 0$ for all $k = 1, 2, \ldots, \tilde{m}$. By Lemma 4, we have

$$\Delta g_{\tilde{m}+1} = (I - \tilde{Q}_{\tilde{m}+1})^{-1} v^h_{f,\tilde{m}+1}.$$

Noting that $(I - \tilde{Q}_{\tilde{m}+1})^{-1} = \sum_{k=0}^{\infty} \tilde{Q}^k_{\tilde{m}+1} \geqslant I$ (see Proposition A.3 in Puterman, 1994) and $v^h_f \geqslant 0$, we have $\Delta g_{\tilde{m}+1} \geqslant v^h_{f,\tilde{m}+1} \geqslant 0$. In summary, we have $\Delta g \geqslant 0$, and so $g^h \geqslant g^f$.

The rest is to prove $g^h \neq g^f$. Suppose that $g^h = g^f$. By Lemma 2(b), we have

$$r^h + P^h g^f = r^h + P^h g^h = \eta^h + g^h = \eta^f + g^f = r^f + P^f g^f. \tag{A.2}$$

On the other hand, since $h \neq f$, from (A.1), (37) and (38), we have

$$r^h + P^h g^f \geqslant r^f + P^f g^f,$$

which leads to a contradiction with (A.2), and so we must have $g^h \geqslant g^f$. □

References

Baxter, J., & Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, *15*, 319–350.

Blackwell, D. (1962). Discrete dynamic programming. *Annals of Mathematics and Statistics*, *33*, 719–726.

Cao, X. R. (1994). *Realization probabilities: The dynamics of queueing systems*. New York: Springer.

Cao, X. R. (1998a). The relations among potentials, perturbation analysis, and Markov decision processes. *Discrete Event Dynamic Systems: Theory and Applications*, *8*, 71–87.

Cao, X. R. (1998b). The Maclaurin series for performance functions of Markov chains. *Advances in Applied Probability*, *30*, 676–692.

Cao, X. R. (1999). Single sample path based optimization of Markov chains. *Journal of Optimization: Theory and Application*, *100*, 527–548.

Cao, X. R. (2000). A unified approach to Markov decision problems and performance sensitivity analysis. *Automatica*, *36*, 771–774.

Cao, X. R. (2003). From perturbation analysis to Markov decision processes and reinforcement learning. *Discrete Event Dynamic Systems: Theory and Applications*, *13*, 9–39.

Cao, X. R. (accepted). Constructing performance sensitivities for Markov systems with potentials as building blocks. *IEEE Transactions on Automatic Control*, submitted for publication.

Cao, X. R., & Chen, H. F. (1997). Potentials, perturbation realization, and sensitivity analysis of Markov processes. *IEEE Transactions on Automatic Control*, *42*, 1382–1397.

Cao, X. R., & Wan, Y. W. (1998). Algorithms for sensitivity analysis of Markov systems through potentials and perturbation realization. *IEEE Transactions on Control Systems Technology*, *6*, 482–494.

Cooper, W. L., Henderson, S. G., & Lewis, M. E. (2003). Convergence of simulation-based policy iteration. *Probability in the Engineering and Information Sciences*, *17*, 213–234.

Federgruen, A., & Shweitzer, P. J. (1984). A fixed-point approach to undiscounted Markov renewal programs. *SIAM Journal of Algebraic Discrete Methods*, *5*, 539–550.

Guo, X. P., & Shi, P. (2001). Limiting average criteria for Nonstationary Markov decision processes. *SIAM Journal of Optimization*, *11*, 1037–1053.

Guo, X. P., Yu, W., & Li, X. O. (2002). Minimax control for discrete-time time-varying stochastic systems. *Automatica*, *38*, 1991–1998.

Hastings, N. A. J. (1969). Optimization of discounted Markov decision problems. *Operation Research Quarterly*, *20*, 499–500.

Ho, Y. C., & Cao, X. R. (1991). *Perturbation analysis of discrete-event dynamic systems*. Boston: Kluwer Academic Publisher.

Howard, R. A. (1960). *Dynamic programming and Markov processes*. Cambridge, MA: MIT Press.

Kallenberg, L. C. M. (2002). Finite state and action MDPs. In E. A. Feinberg & A. Shwartz (Ed.), *The handbook of Markov decision processes*. Boston: Kluwer Academic Publishers.

Marbach, P., & Tsitsiklis, T. N. (2001). Simulation-based optimization of Markov reward processes. *IEEE Transaction on Automatic Control*, *46*, 191–209.

Ng, M. K. (1999). A note on policy iteration algorithm for discounted Markov decision problems. *OR Lettere*, *25*, 195–197.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. New York: Wiley.

Shweitzer, P. J., & Brower, A. (1987). Fixed-point mapping approach to communication Markov decision processes. *Journal of Mathematical Analysis and Applications*, *123*, 117–130.

Shweitzer, P. J., & Federgruen, A. (1978). Foolproof convergence in multichain policy iteration. *Journal of Mathematical Analysis and Applications*, *64*, 360–380.

Spreen, D. (1985). A further anti-cycling rule in multi-chain policy iteration for undiscounted Markov renewal programs. *Zeitschrift für Operations Research*, *25*, 153–160.

Veinott, A. F. (1966). On finding optimal policies in discrete dynamic programming with no discounting. *Annals of Mathematics and Statistics*, *37*, 1284–1294.

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), Hong Kong, China. He is the director of the Center for Networking at HKUST.

Dr. Cao owns three patents in data- and tele-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 System 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 on 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 dynamic systems, optimisation theory, performance analysis of communication systems, and signal processing.

Xianping Guo received the B.S. and M.S. degrees in Mathematics from Hunan Normal University, China, in 1987 and 1990, respectively, and the Ph.D. degree in Probability and Statistics from Changsha Railway University, China, in 1996. From September 1996 to August 1998, he worked as a postdoctor at Zhongshan University, China. Then he worked as a Research fellow in Queensland University and in South Australia University, Australia, for about one year. From August 2000 to July 2002, he was a visiting professor at CINVESTAV-IPN, Mexico, and then worked as a visiting scholar at Hong Kong University of Science and Technology, Hong Kong, from August 2002 to July 2003. Now he is a full professor with Zhongshan University, China. His research interests include Markov decision processes, dynamic stochastic games as well as Markov processes.