Semi-Markov Decision Problems and Performance Sensitivity Analysis

Xi-Ren Cao, *Fellow, IEEE*

***Abstract*—Recent research indicates that Markov decision processes (MDPs) can be viewed from a sensitivity point of view; and perturbation analysis (PA), MDPs, and reinforcement learning (RL) are three closely related areas in optimization of discrete-event dynamic systems that can be modeled as Markov processes. The goal of this paper is two-fold. First, we develop PA theory for semi-Markov processes (SMPs); and second, we extend the aforementioned results about the relation among PA, MDP, and RL to SMPs. In particular, we show that performance sensitivity formulas and policy iteration algorithms of semi-Markov decision processes (SMDPs) can be derived based on performance potential and realization matrix. Both the long-run average and discounted-cost problems are considered; this approach provides a unified framework for both problems, and the long-run average problem corresponds to the discounted factor being zero. The results indicate that performance sensitivities and optimization depend only on first-order statistics. Single sample path-based implementations are discussed.**

***Index Terms*—Discounted Poisson equations, discrete-event dynamic systems (DEDS), Lyapunov equations, Markov decision processes (MDPs), perturbation analysis (PA), perturbation realization, Poisson equations, policy iteration, potentials, reinforcement learning (RL).**

I. Introduction

MARKOV decision processes (MDPs), perturbation analysis (PA), and reinforcement learning (RL) are three different approaches to optimization of discrete event dynamic systems (DEDSs). In MDPs, performance depend on policies; a policy with a better performance can be identified by analyzing the behavior of the Markov process under the current policy; the policy with the best performance can then be obtained by *policy iteration* [1], [20]. With PA, the derivative of a performance measure with respect to a parameter can be obtained by analyzing the behavior of a DEDS. (For PA of queueing systems; see [4], [11], [15], and PA of Markov processes [9]). The derivatives obtained by PA can then be used to determine the optimal value of the performance measure [13], [18]. The goal of RL [1], [3], [21], [22] is to learn how to make decisions to improve a system's performance by observing its behavior.

Recent research shows that MDPs can be viewed from a sensitivity point of view [6], and policy iteration, in fact, chooses its next policy along the direction with the steepest performance derivative provided by PA. Both MDP and PA of Markov processes are based on an important concept, called *performance potential*, which is strongly related to *perturbation realization* in PA. These concepts provide an intuitive explanation for both PA and MDPs (in view of discrete sensitivity) and their relations [6], [7]. The performance potential at state i can be approximated by the mean of the sum of the performance functions at the first N transitions on a sample path starting from state i, and hence it can be estimated online. A number of other single sample path-based estimation algorithms for potentials have been derived [6]. With the potentials estimated, performance derivatives (i.e., PA) can be obtained and policy iteration (i.e., MDPs) can be implemented based on a single sample path. Stochastic approximation methods can be used in these two cases to improve the convergence speeds and to reduce stochastic errors. The sample path based implementation of PA and MDP resembles RL, in particular the Q-learning method [21], which estimates Q-factors, a variant of potentials when the system structure is completely unknown. The analysis based on performance potentials provides a unified framework to both MDPs and PA with both average- and discounted-cost performance measures [7]. The results for the average-cost problems (in the discrete time case) correspond to the case with the discounted factor being one [7]. The sensitivity point of view of PA, MDP, and RL brings in some new insight to the area of learning and optimization. For more details, see [8].

In this paper, we develop the PA theory and extend the above results to semi-Markov processes (SMPs) with a continuous-time model. The previous results on MPDs and PA of Markov processes become special cases. Therefore, our approach provides a unified framework to both decision problems and sensitivity analysis (PA) with both average- and discounted-cost performance measures for both semi-Markov and Markov processes. In this approach, decision problems are viewed as performance sensitivities in a discrete policy space, and PA is regarded as performance sensitivities in a continuous parameter space. Both of them depend on the concept of performance potential. The average-cost problem is a special case of the discounted-cost problem with discount factor $\beta = 0$. RL methods can be developed to estimate the potentials, Q-factors, or even the performance derivatives.

In Section II, we review the fundamentals of semi-Markov processes. In particular, we show that the steady-state performance (average cost) depends on an equivalent infinitesimal generator which depends only on the first order statistics of the semi-Markov kernel. We study the average-cost semi-Markov decision problem in Section III. We start with introducing the concept of perturbation realization, which is fundamental in PA

Manuscript received February 18, 2002; revised September 25, 2002 and November 7, 2002. Recommended by Associate Editor Y. Wardi. This work was supported in part by a grant from Hong Kong UGC.

The author is with the Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: eecao@ust.hk).

Digital Object Identifier 10.1109/TAC.2003.811252

[4], [9]. We define realization matrix and prove that it satisfies the Lyapunov equation. From the realization matrix, we define performance potentials and prove that it satisfies the Poisson equation. With realization matrix and performance potential, sensitivity formulas and policy iteration algorithms of semi-Markov decision processes (SMDPs) can be derived easily in the same way as for Markov processes. It is also shown that the potentials can be estimated on a single sample path and, hence, online algorithms can be derived for performance sensitivities and policy iteration of SMDP. Section IV deals with the discounted-cost problem. We derive the equivalent infinitesimal generators with discounted factor β and the corresponding discounted Poisson and Lyapunov equations. The sensitivity formulas and policy iteration can be derived using performance potentials which is the solution to the discounted Poisson equation. By carefully defining the discounted-cost performance measure, we show that the average-cost problem is the limiting case as the discount factor β goes to zero. Thus, the potential based approach applies to both average- and discounted-cost problems. Section V summarizes the results with a few remarks on its significance and future research topics.

II. Fundamentals for SMPs

We study an SMP $\{X_t, t \geq 0\}$ defined on a finite state–space $\mathcal{E} = \{1, 2, \ldots, M\}$. Let $T_0, T_1, \ldots, T_n, \ldots$, with $T_0 = 0$, be the transition epoches. Each interval $[T_n, T_{n+1})$ is called a period. The process is right continuous so the state at each transition epoch is the state after the transition. Let $X_n = X_{T_n}$, $n = 0, 1, 2, \ldots$.

Define the semi-Markov kernel [12] as

$$Q(i,j,t) = P\{X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i\}.$$

Set

$$\begin{aligned} Q(i,t) &= \sum_{j\in\mathcal{E}} Q(i,j,t) = P\{T_{n+1} - T_n \leq t | X_n = i\} \\ H(i,t) &= 1 - Q(i,t) \\ Q(i,j) &= \lim_{t\to\infty} Q(i,j,t) = P\{X_{n+1} = j | X_n = i\} \end{aligned}$$

and

$$\begin{aligned} G(i,j,t) &= \frac{Q(i,j,t)}{Q(i,j)} \\ &= P\{T_{n+1} - T_n \leq t | X_n = i, X_{n+1} = j\}. \end{aligned}$$

Normally, $Q(i,i) = 0$, for all $i \in \mathcal{E}$. However, in general, we may allow the process jumps from a state into itself at the transition epoches; in such a case $Q(i,i)$ may be nonzero and our results still hold. Furthermore, a Markov process with transition rates $\lambda(i)$ and transition probabilities $Q(i,j)$ can be viewed as a SMP whose kernel is $Q(i,j,t) = Q(i,j)[1 - e^{-\lambda(i)t}]$.

We assume that the matrix $Q(i,j)$ is irreducible and nonperiodic [2]. Let

$$m(k) = \int_0^\infty sQ(k,ds) = E[T_{n+1} - T_n | X_n = i]$$

be the mean of the sojourn time at state k. We also assume that $m(k) < \infty$ for all k. Under these assumptions the semi-Markov process is irreducible and nonperiodic and hence ergodic. Define the hazard rates as

$$q(i,t) = \frac{\frac{d}{dt}Q(i,t)}{H(i,t)}$$

and

$$q(i,j,t) = \frac{\frac{d}{dt}Q(i,j,t)}{H(i,t).}$$

The latter is the rate that the process jumps from i to j in $[t, t+dt)$ given that the process does not jump out from state i in $[0,t)$.

Let $P_t(i,j) = P\{X_t = j | X_0 = i\}$. By the total probability theorem, we can easily derive

$$\begin{aligned} P_{t+\Delta t}(i,j) = &\sum_{k\in\mathcal{E}} P_t(i,k) \int_0^\infty p_t(s|k) \\ &\times \{I(k,j)[1 - q(k,s)\Delta t] + q(k,j,s)\Delta t\}\, ds \end{aligned} \tag{1}$$

where $I(j,k) = 1$ if $j = k$, $I(j,k) = 0$ if $j \neq k$ ($I(i,j)$ is the (i,j)th entry in the identity matrix I), $p_t(s|k)ds$ is the probability that given the state at time t is k the process has been in state k for a period of s to $s + ds$, which may depend on the initial state. Precisely, let n_t be the integer such that $T_{n_t} \leq t < T_{n_t+1}$. Then

$$p_t(s|k)ds = P(s \leq t - T_{n_t} < s + ds | X_t = k). \tag{2}$$

It is proved in the Appendix that

$$\lim_{t\to\infty} p_t(s|k) = \frac{H(k,s)}{m(k)}. \tag{3}$$

Now, set $\Delta t \to 0$ in (1), and we obtain

$$\begin{aligned} \frac{dP_t(i,j)}{dt} = &-\sum_{k\in\mathcal{E}} P_t(i,k) \int_0^\infty \\ &\times \{p_t(s|k)[I(k,j)q(k,s) - q(k,j,s)]\}\, ds. \end{aligned} \tag{4}$$

Since the semi-Markov process is ergodic, when $t \to \infty$, we have $P_t(i,j) \to p(j)$ [12] and $(dP_t(i,j)/dt) \to 0$, where $p(j)$ is the steady-state probability of j. Letting $t \to \infty$ in both sides of (4), we get

$$\begin{aligned} 0 = &-\sum_{k\in\mathcal{E}} p(k) \int_0^\infty \frac{1}{m(k)} \\ &\times \left\{ I(k,j)\frac{d}{ds}[Q(k,s)] - \frac{d}{ds}[Q(k,j,s)] \right\} ds \\ = &-\sum_{k\in\mathcal{E}} p(k) \left\{ \frac{1}{m(k)} [I(k,j) - Q(k,j)] \right\} \\ = &-\sum_{k\in\mathcal{E}} p(k) \{\lambda(k)[I(k,j) - Q(k,j)]\} \end{aligned}$$

where

$$\lambda(k) = \frac{1}{m(k)}.$$

Finally, we have

$$\sum_{k\in\mathcal{E}} p(k) A(k,j) = 0 \qquad \forall\, j \in \mathcal{E}$$

where

$$A(k,j) = -\lambda(k)\left[I(k,j) - Q(k,j)\right]. \tag{5}$$

In matrix form, we can write

$$pA = 0 \tag{6}$$

where $p = (p(1), \dots, p(M))$ is the steady-state probability vector and A is a matrix with elements $A(k,j)$. (6) is consistent with [12, Th. 10.5.22]. In addition, we have

$$Ae = 0 \tag{7}$$

where $e = (1, 1, \dots, 1)^T$ is an M-dimensional column vector whose components are all ones (the superscript "T" denotes transpose). It is well known that for ergodic processes (6) and (7) have a unique solution.

Equation (6) is exactly the same as the Markov process with A being its infinitesimal generator. This means that the steady-state probability is insensitive to the high order statistics of the sojourn times at the states, and is independent of whether the sojourn time at state i depends on j, the state it jumps into from i. In the next section, we will see that A plays the same role for semi-Markov processes as the infinitesimal generator for Markov processes in policy iteration and PA.

III. Average-Cost Problems

A. Perturbation Realization Matrices

Consider a semi-Markov process starting from a transition epoch $T_0 = 0$ in state $X_0 = j$. At any time $t \in [T_n, T_{n+1})$, denote $Y_t = X_{n+1}$, i.e., the state that the process jumps into at the next transition epoch. We define the performance value at any time t as $f(X_t, Y_t)$, where $f : \mathcal{E} \times \mathcal{E} \to R$. The long-run average performance measure is

$$\eta = \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T f(X_t, Y_t)dt\right]$$

where E denotes the expectation operator. Denote the instant at which the process jumps into state i for the first time as

$$S^j(i) = inf\left\{t \geq 0 | X_t = i, X_0 = j\right\}.$$

Following the same approach as for the PA of Markov processes [9], we define the *perturbation realization factors* as (the only difference is that $T_0 = 0$ must be a transition epoch in the semi-Markov case)

$$D(i,j) = E\left[\int_0^{S^j(i)} \left[f(X_t, Y_t) - \eta\right] dt | X_0 = j\right]. \tag{8}$$

As we will see, $D(i,j)$ measures the effect of a change from state j to i on the long-run integration of performance function f. The matrix $D = [D(i,j)]$ is called a *perturbation realization matrix.*

Let $p(i,j)$ be the steady-state probability of $X_t = i$ and $Y_t = j$ and $p(j|i)$ be the conditional steady-state probability of $Y_t = j$ given that $X_t = i$, e.g., $\lim_{t\to\infty} P(Y_t = j | X_t = i)$ (not to be confused with $\lim_{t\to\infty} P(X_{n+1} = j | X_n = i)$). It is proved in the Appendix that

$$p(j|i) = \frac{\int_0^\infty sQ(i,j,ds)}{\int_0^\infty sQ(i,ds)} = \frac{Q(i,j)m(i,j)}{m(i)} \tag{9}$$

where

$$\begin{aligned} m(i,j) &= \int_0^\infty sG(i,j,ds) \\ &= E[T_{n+1} - T_n | X_n = i, X_{n+1} = j] \end{aligned}$$

and

$$m(i) = \sum_{j\in\mathcal{E}} Q(i,j)m(i,j) = \int_0^\infty sQ(i,ds). \tag{10}$$

Thus

$$p(i,j) = p(j|i)p(i) = p(i)\frac{Q(i,j)m(i,j)}{m(i)}.$$

By ergodicity, we have

$$\eta = \sum_{i,j\in\mathcal{E}} p(i,j)f(i,j) = \sum_{i\in\mathcal{E}} p(i)f(i) = pf$$

where $f = (f(1), f(2), \dots, f(M))^T$, and (Note that we use f for both $f(i)$ and $f(i,j)$)

$$f(i) = \frac{\sum_{j\in\mathcal{E}} Q(i,j)f(i,j)m(i,j)}{m(i)}. \tag{11}$$

From (8), we have

$$\begin{aligned} D(i,j) =& E\left[\int_0^{T_1} \left[f(X_t,Y_t) - \eta\right] dt | X_0 = j\right] \\ &+ E\left[\int_{T_1}^{S^j(i)} \left[f(X_t,Y_t) - \eta\right] dt | X_0 = j\right] \\ =& \sum_{k\in\mathcal{E}} Q(j,k) \\ &\times \left\{ E\left[\int_0^{T_1} \left[f(X_0,Y_0) - \eta\right] dt | X_0 = j, X_1 = k\right]\right. \\ &\left. + E\left[\int_{T_1}^{S^j(i)} \left[f(X_t,Y_t) - \eta\right] dt | X_0 = j, X_1 = k\right]\right\} \\ =& \sum_{k\in\mathcal{E}} Q(j,k) \\ &\times \left\{\left[f(j,k) - \eta\right] E[T_1 | X_0 = j, X_1 = k]\right. \\ &\left. + E\left[\int_{T_1}^{S^k(i)} \left[f(X_t,Y_t) - \eta\right] dt | X_1 = k\right]\right\} \\ =& \sum_{k\in\mathcal{E}} Q(j,k) \\ &\times \left\{\left[f(j,k) - \eta\right] m(j,k)\right. \end{aligned}$$

$$+E\left[\int_{T_1}^{S^k(i)}[f(X_t,Y_t)-\eta]\,dt|X_1=k\right]\Bigg\}.$$

From (10) and (11), the previous equation leads to

$$D(i,j)=m(j)\,[f(j)-\eta]+\sum_{k\in\mathcal{E}}Q(j,k)D(i,k)$$

or, equivalently

$$\begin{aligned}-[f(j)-\eta]&=\sum_{k\in\mathcal{E}}\{-\lambda(j)\,[I(j,k)-Q(j,k)]\,D(i,k)\}\\&=\sum_{k\in\mathcal{E}}\{A(j,k)D(i,k)\}\,.\end{aligned}$$

In matrix form, this is

$$DA^T=-[ef^T-\eta ee^T] \qquad (12)$$

where D is a matrix whose components are $D(i,j)$.

Next, on the process X_t, with $T_0=0$ being a transition epoch and $X_0=j$, for any state $i\in\mathcal{E}$ we define two sequences $u_0,u_1,\ldots$, and $v_0,v_1,\ldots$, as follows:

$$u_0=T_0=0 \qquad (13)$$

$$v_n=inf\{t\geq u_n|X_t=i\}. \qquad (14)$$

and

$$u_{n+1}=\inf\{t\geq v_n|X_t=j\} \qquad (15)$$

e.g., v_n is the first time when the process reaches i after u_n and u_{n+1} is the first time when the process reaches j after v_n. Apparently, $u_0,u_1,\ldots$ are stopping times and X_t is a regenerative process with $\{u_n,n=0,1,\ldots\}$ as its associated renewal process. By the theory of regenerative processes, we have

$$\begin{aligned}\eta&=\frac{E\left[\int_{u_0}^{u_1}f(X_t,Y_t)dt\right]}{E(u_1-u_0)}\\&=\frac{E\left[\int_0^{v_0}f(X_t,Y_t)dt\right]+E\left[\int_{v_0}^{u_1}f(X_t,Y_t)dt\right]}{E[v_0]+E[u_1-v_0]}.\end{aligned}$$

Thus

$$E\left[\int_0^{v_0}[f(X_t,Y_t)-\eta]\,dt\right]+E\left[\int_{v_0}^{u_1}[f(X_t,Y_t)-\eta]\,dt\right]=0.$$

By the definition of u_0, v_0 and u_1, we know that the aforementioned equation is

$$D(i,j)+D(j,i)=0.$$

Therefore, the matrix D is skew-symmetric

$$D^T=-D.$$

Taking the transpose of (12), we get

$$-AD=-[fe^T-\eta ee^T].$$

From this equation and (12), D satisfies the following Lyapunov equation:

$$AD+DA^T=-F \qquad (16)$$

where $F=ef^T-fe^T$. This is the same as the Lyapunov equation for Markov processes [9]. When $\lambda(i)\equiv 1$ for all i, we have $A=Q-I$, $Q=[Q(i,j)]$ is the transition probability matrix of the Markov chain embedded in the transition epoches. From (12) and $Ae=0$, we get $ADA^T=0$. Thus, (16) becomes

$$-D+QDQ^T=-F.$$

This is the Lyapunov equation for discrete-time systems [6].

B. Performance Potentials

Similar to (13) to (15), for any three states i, j, k, we define three sequences $u_0,u_1,\ldots$; $v_0,v_1,\ldots$; and $w_0,w_1,\ldots$ as follows: $u_0=T_0=0$, $X_0=j$, $v_n=inf\{t\geq u_n,X_t=i\}$, $w_n=inf\{t\geq v_n,X_t=k\}$, and $u_{n+1}=\inf\{t\geq w_n,X_t=j\}$. By a similar approach, we can prove

$$D(i,j)+D(j,k)+D(k,i)=0.$$

In general, we can prove that for any closed circle $i_1-i_2-\cdots-i_n-i_1$ in the state–space, we have

$$D(i_1,i_2)+D(i_2,i_3)+\cdots+D(i_{n-1},i_n)+D(i_n,i_1)=0.$$

This is similar to the conservative law of the potential energy in physics. Therefore, we can define a potential $g(i)$ at any state and write $D(i,j)=g(j)-g(i)$ and

$$D=eg^T-ge^T \qquad (17)$$

where $g=(g(1),\ldots,g(M))^T$ is a column vector. Note that if g fits (17), so does $g+ce$ for any constant c. g is called a *performance potential* vector, and $g(i)$ the performance potential at state i. Similar to the potential energy, g may have different versions, each differs by a constant. ([9] contains more discussions on $D(i,j)$ and $g(i)$ for Markov processes).

Substituting (17) into (12), we get

$$Ag=-f+\eta e. \qquad (18)$$

This is the *Poisson equation*. Since $Ae=0$, A is not invertible. Thus, the solutions are not unique. Now suppose g is any solution to (18). Set $c=\eta-pg$ and choose $g'=g+ce$. Then, $pg'=\eta$. Thus, there always exists a solution to (18) such that $pg=\eta$. Putting this into (18), we get

$$Ag=-f+(pg)e=-f+e(pg)$$

and

$$(-A+ep)g=f. \qquad (19)$$

This is the same as the Poisson equation for Markov processes. For ergodic semi-Markov processes, $(-A+ep)$ is invertible. Equation (19) only defines a particular version of the performance potentials. Define the *group inverse* of A as

$$A^{\#}=(-A+ep)^{-1}.$$

We have

$$g = A^{\#} f.$$

Multiplying both sides of (19) by p on the left, we get

$$pg = \eta = pf.$$

This can be viewed as the "normalizing" condition to the potential defined in (19).

C. Sensitivity and SMDPs

We have shown that with the properly defined g and A, Poisson equation and Lyapunov equation hold for potentials and realization matrices, respectively, for semi-Markov processes. Thus, performance sensitivity formulas can be derived in a similar manner, and the results are briefly stated here.

First, for two SMPs with A', η' f' and A, η, f, multiplying both sides of (19) on the left by p' and using $p'A' = 0$, we get

$$\begin{aligned}\eta' - \eta &= p'\left[(A' - A)g + (f' - f)\right] \\ &= p'\left[(A'g + f') - (Ag + f)\right]. \qquad (20)\end{aligned}$$

This serves as a foundation for SMDPs. Policy iteration for SMPs can be derived from (20) by noting $p' > 0$ componentwisely. This is the same as MDPs and we shall only briefly state the results.

Specifically, a semi-Markov decision problem is defined as follows. At any transition epoch t with $X_t = i$, an action α is taken from an action space $\mathcal{A}$ and applied to the SMP. This action determines the ith item of the semi-Markov kernel $Q^{\alpha}(i,j,t)$, $j = 1, 2, \ldots, M$, and performance function $f^{\alpha}(i,j)$, $j = 1, 2, \ldots, M$, which in turn determine $\lambda^{\alpha}(i)$, $Q^{\alpha}(i,j)$ and $f^{\alpha}(i)$ (or equivalently $A^{\alpha}(i,j)$ and $f^{\alpha}(i)$). A stationary policy is a mapping $\mathcal{L} : \mathcal{E} \rightarrow \mathcal{A}$. For any state $i \in \mathcal{E}$, $\mathcal{L}$ specifies an action $\mathcal{L}(i) \in \mathcal{A}$. A policy $\mathcal{L}$ specifies an infinitesimal generator $A^{\mathcal{L}}$. The policy space is denoted as $\mathcal{C}$. We use superscript $\mathcal{L}$ to denote the quantities associated with policy $\mathcal{L}$, e.g.,

$$\eta^{\mathcal{L}} = \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T f^{\mathcal{L}(X_t)}(X_t, Y_t)dt\right].$$

The objective is to minimize the cost over the policy space $\mathcal{C}$, i.e., to obtain $\min_{\mathcal{L}\in\mathcal{C}} \eta^{\mathcal{L}}$. We first choose any initial policy $\mathcal{L}_0$, which determines $A^{\mathcal{L}_0}$. Given a policy $A^{\mathcal{L}_n}$, $n = 0, 1, \ldots$, we solve the Poisson equation (19) for the potential $g^{\mathcal{L}_n}$. Then, we choose a policy $A^{\mathcal{L}_{n+1}}$ that minimizes $A^{\mathcal{L}'}g^{\mathcal{L}_n} + f^{\mathcal{L}'}$, i.e., $\mathcal{L}_{n+1} = arg\{\min_{\mathcal{L}'\in\mathcal{C}}[A^{\mathcal{L}'}g^{\mathcal{L}_n} + f^{\mathcal{L}'}]\}$, componentwisely. From (20), we have $\eta^{\mathcal{L}_{n+1}} \leq \eta^{\mathcal{L}_n}$. If $\eta^{\mathcal{L}_{n+1}} < \eta^{\mathcal{L}_n}$, we set $n := n + 1$ and continue the procedure until $\eta^{\mathcal{L}_{n+1}} = \eta^{\mathcal{L}_n}$; at this point the policy iteration reaches the optimal policy.

Since we minimize $A^{\mathcal{L}'}g^{\mathcal{L}_n} + f^{\mathcal{L}'}$ componentwisely, it requires that the actions at different states do not interact with each other. This indeed is the case: by examining (5), we can see that the ith row of A is determined completely by $Q(i,j,t)$ with the same i, which is controlled by the action $\mathcal{L}(i)$.

Equation (20) can be viewed as a sensitivity equation in a discrete space, in which each point represents a policy. Next, we consider the performance sensitivity in continuous space. This continuous space can be obtained by randomizing policies in a discrete policy space. Let $A(\delta)$ be a randomized policy which takes policy A with probability $1 - \delta$ and policy $A' = A + B$ with probability δ, where $0 \leq \delta \leq 1$ and B is an $M \times M$ matrix satisfying $Be = 0$. Then, $A(\delta) = A + B\delta$. Let f and $f' = f + h$ be the performance functions associated with policies A and A', respectively. Suppose A changes to $A(\delta) = A + B\delta$, f changes to $f(\delta) = f + h\delta$. (For example, if $\lambda(i)$ changes to $\lambda(i) + (\Delta\lambda)\delta$, $i = 1, 2 \ldots, M$, then according to (5), A changes to $A + \delta(\Delta\lambda)(I - Q)$, i.e., $B = \Delta\lambda(I - Q)$; if Q changes to $Q + \Delta Q$, then $B = \lambda(\Delta Q)$). Then, η will change to $\eta(\delta) = \eta + \Delta\eta$ and p changes to $p(\delta)$. From (20), we have

$$\eta(\delta) - \eta(0) = p(\delta)\left[(A(\delta) - A)\, g + (f(\delta) - f)\right].$$

Letting $\delta \rightarrow 0$, we get

$$\frac{d\eta}{dB} = p(Bg + h) = p(BA^{\#}f + h) \qquad (21)$$

where $d\eta/dB$ denotes the derivative along the direction of B. We can also obtain performance sensitivity using D. From (17), we have

$$Dp^T = (eg^T - ge^T)p^T = \eta e - g.$$

Replacing g with D in the sensitivity equation, we get the performance difference

$$\eta' - \eta = p'\left[(A' - A)D^T p^T + (f' - f)\right]$$

and the performance derivative

$$\frac{d\eta}{dB} = p^T(BD^T p^T + h).$$

Equation (8) provides a way to estimate realization matrices and potentials on a sample path. From (8), we obtain

$$\begin{aligned}D(i,j) = \lim_{T\to\infty} \Bigg\{ &E\left[\int_0^T [f(X_t, Y_t) - \eta]\, dt | X_0 = j\right] \\ &- E\left[\int_0^T \left[f(\tilde{X}_t, \tilde{Y}_t) - \eta\right] dt | \tilde{X}_0 = i\right]\Bigg\}\end{aligned}$$

where $\tilde{X}_t$ has the same kernel as X_t. Therefore

$$g(j) = \lim_{T\to\infty} E\left[\int_0^T [f(X_t, Y_t) - \eta]\, dt | X_0 = j\right] + c \qquad (22)$$

is a performance potential at j, where c is any constant. This is the same as for the Markov process case, except that the integration starts with a transition epoch. The convergence of the right-hand side of (22) can be easily verified by, e.g., using the embedded Markov chain model [6]. Single sample path based algorithms (e.g., Monte Carlo estimates) can be easily developed for potentials and realization matrices, and therefore the performance derivative (21) can be obtained and policy iteration can be implemented with a single sample path. For example

$$D(j,i) = \lim_{N\to\infty} \frac{1}{N+1} \times \left\{ \sum_{n=0}^{N} \int_{u_n}^{v_n} f(X_t, Y_t)dt - \eta \sum_{n=0}^{N} (v_n - u_n) \right\} \qquad (23)$$

where u_n and v_n are defined in (13) to (15). Each segment from u_n to u_{n+1}, $n = 0, 1, \ldots$, can be viewed as an independent sample path starting with an initial state j, $\int_{u_n}^{v_n} f(X_t, Y_t)dt$ is the value of $\int_0^{S^j(i)} f(X_t, Y_t)dt$ in (8) on one sample path, and $\eta(v_n - u_n)$ is that of $\int_0^{S^j(i)} \eta dt$. η can be estimated on a sample path by using

$$\eta = \lim_{N\to\infty} \frac{1}{u_N} \int_0^{u_n} f(X_t, Y_t)dt.$$

More complicated algorithms involving simultaneous estimation of η and $D(i,j)$ can also be developed. $g(i)$, $i \in \mathcal{E}$ can then be obtained by setting $g(i^*) = 0$ for any arbitrarily chosen i^* and using (17). Algorithms based on (23) usually have smaller variannce than those based on (22). This is similar to the case with Markov process [6].

The high-order derivatives are the same as those for Markov processes [5]

$$\frac{d^n\eta}{dB^n} = n!p(BA^{\#})^{n-1}(BA^{\#}f + h).$$

In addition, we have the following expansion:

$$\eta(\delta) = \eta + p\delta \sum_{k=1}^{n} (\delta BA^{\#})^{k-1}(BA^{\#}f + h) + p(\delta)(\delta BA^{\#})^{n}(BA^{\#}f + h).$$

When $h = 0$, this becomes

$$\eta(\delta) = p \sum_{k=0}^{n} (\delta BA^{\#})^{k} f + p(\delta)(\delta BA^{\#})^{n+1} f.$$

Thus, we can use $p\sum_{k=0}^{n} (\delta BA^{\#})^{k} f$ to estimate $\eta(\delta)$ with $p(\delta)(\delta BA^{\#})^{n+1} f$ being the error in the estimation. All the items in p and $A^{\#}$ can be estimated on a sample path of the Markov process with A; see [5].

D. Example

We use an simple example to illustrate the application of the theory previously developed. Consider a communication line (or a switch, a router, etc.) to which packets arrive in a Poisson process with rate λ. The packet length is assumed to have a general distribution function $F(x)$, with the unit of x being bit. For each packet, the system manager can choose the transmission rate θ, whose unit is bit per second. Thus, the transmission time for each packet has a distribution function $\tilde{F}(\tau) = P(t \le \tau) = P(x \le \theta\tau) = F(\theta\tau)$. In a real system, θ takes discrete values, e.g., the number of channels; each channel has a fixed amount of bandwidth. Thus, we can view θ as an action and denote the actions space as $\{\theta_1, \theta_2, \ldots, \theta_K\}$, with $\theta_k = k * \mu$, where μ denotes the transmission rate of one channel. The system can be modeled as an M/G/1 queue; the state at time t is $N(t) = i$ with i being the number of customers in the queue at time t. For stability, we assume $\mu > \lambda\bar{x}$, where $\bar{x}$ is the mean length of a packet. The decision for actions is made at the beginning of the transmission of each packet. Thus, the decision epoches, which consist of all the service completion times and the arrival times to all the idle periods, are denoted as $T_0, T_1, \ldots$. Define $X_t = N(T_n)$ for $T_n \le t < T_{n+1}$, $n = 0, 1, 2, \ldots$, then X_t is a semi-Markov process. It is clear that the following equations hold for X_t:

$$\begin{aligned} Q(0,1,t) &= 1 - e^{-\lambda t} \\ Q(n,t) &= F(\theta t), \qquad n > 0 \end{aligned}$$

and

$$\begin{aligned} Q(i,j,dt) &= P[X_{n+1} = j, t \le T_{n+1} - T_n < t + dt | X_n = i] \\ &= \left\{ \frac{(\lambda t)^{j-i+1}}{(j-i+1)!} e^{-\lambda t} \right\} F(\theta dt), \qquad i - 1 \le j \end{aligned}$$

where the term in braces is the probability that there are $j-i+1$ arrivals in the period of $[0, t)$.

The cost consists of two parts: the holding cost $f_1(i,j)$ and the bandwidth cost $f_2(\theta)$. That is

$$f^{\theta}(i,j) = f_1(i,j) + f_2(\theta).$$

It is well known that if in an interval $[0, t]$ there are k arrivals from a Poisson process, then these k arrivals uniformly distribute over the period (see, e.g., [17]). Thus, the average number of customers in $[0, t]$ is $(i + j)/2$ and we can set

$$f^{\theta}(i,j) = \kappa_1 \frac{i+j}{2} + \kappa_2 \theta$$

where the first term represents the cost for average waiting time. The problem is now formulated as a SMDP problem and the results developed in this paper can be applied.

IV. Discounted-Cost Problems

A. Performance Formula

Instead of the average-cost performance, we consider the problem with discounted performance criteria. For any $\beta > 0$, we define the performance measure as

$$\eta_\beta(i) = \lim_{N\to\infty} E\left[\int_0^{T_N} \beta e^{-\beta t} f(X_t, Y_t)dt | X_0 = i \right], \qquad T_0 = 0 \qquad (24)$$

the performance potentials as

$$g_\beta(i) = \lim_{N\to\infty} E\left[\int_0^{T_N} e^{-\beta t} \left[f(X_t, Y_t) - \eta\right] dt | X_0 = i \right] \qquad i = 1, 2, \ldots, M \qquad (25)$$

where $\eta = pf$ is the average performance, and the performance and potential vectors as $\eta_\beta = (\eta_\beta(1), \ldots, \eta_\beta(M))^T$ and $g_\beta = (g_\beta(1), \ldots, g_\beta(M))^T$.

Note that the definition (24) differs from the standard one (e.g., in [20], $\eta_\beta(i)$ is defined as $\lim_{N\to\infty} E\{\int_0^{T_N} e^{-\beta t} f(X_t, Y_t)dt|X_0 = i\}$) with a factor "$\beta$." We adopt such a definition for the following reasons. First, the continuity of η_β holds at $\beta = 0$. In fact, we define

$$\eta_0 = \lim_{\beta\to 0} \eta_\beta.$$

We shall prove that the limit exists and $\eta_0 = \eta e$, where η is the performance for the average-cost problem. Second, (24) has its own physical meaning: since $\int_0^\infty \beta e^{-\beta t}dt = 1$, in (24) the performance value is distributed on the sample path according to the weighting factor $\beta e^{-\beta t}$. The average cost performance corresponds to an "even" distribution on the sample path. (With the standard definition, η_β goes to infinity as β approaches 0). Third, with this approach, we can develop a unified theory of PA and MDP that applies to both the discounted-cost and the average-cost problems. Finally, since β is a fixed number for a particular problem, it should be straightforward to translate all the results in this paper to the "standard" definition. A similar definition is used in [7] for discrete time Markov chains.

Similarly, we define

$$g_0 = \lim_{\beta\to 0} g_\beta$$

and will prove $g_0 = g$, the performance potential for the average-cost problem. From (24) and (25), we have

$$\eta_\beta = \beta g_\beta + \eta e. \tag{26}$$

Now, we have

$$\begin{aligned}
\eta_\beta(i) =& E\left[\int_0^{T_1} \beta e^{-\beta t} f(i, Y_0)dt|X_0 = i\right] \\
&+ \lim_{N\to\infty} E\left[\int_{T_1}^{T_N} \beta e^{-\beta t} f(X_t, Y_t)dt|X_0 = i\right] \\
=& \sum_{j\in\mathcal{E}} \int_0^\infty \int_0^{T_1=\tau} \beta e^{-\beta t} f(i,j)dt\, Q(i,j,d\tau) + \lim_{N\to\infty} \sum_{j\in\mathcal{E}} \\
&\times \left\{ \int_0^\infty e^{-\beta\tau} \int_{T_1=\tau}^{T_N} \beta e^{-\beta(t-\tau)} f(X_t, Y_t)dt\, Q(i,j,d\tau) \right\} \\
=& \sum_{j\in\mathcal{E}} \left\{ f(i,j) \int_0^\infty \left(\int_0^\tau \beta e^{-\beta t}dt \right) Q(i,j,d\tau) \right\} \\
&+ \sum_{j\in\mathcal{E}} \left\{ \int_0^\infty e^{-\beta\tau} Q(i,j,d\tau)\eta_\beta(j) \right\} \\
=& \sum_{j\in\mathcal{E}} \left\{ f(i,j) \int_0^\infty (1 - e^{-\beta\tau}) Q(i,j,d\tau) \right\} \\
&+ \sum_{j\in\mathcal{E}} \left\{ \int_0^\infty e^{-\beta\tau} Q(i,j,d\tau)\eta_\beta(j) \right\}.
\end{aligned} \tag{27}$$

To continue our analysis, we set

$$\begin{aligned}
\alpha_\beta(i) &= \int_0^\infty e^{-\beta\tau} Q(i, d\tau) \\
Q_\beta(i,j) &= \frac{\int_0^\infty e^{-\beta\tau} Q(i,j,d\tau)}{\int_0^\infty e^{-\beta\tau} Q(i,d\tau)} \\
m_\beta(i,j) &= \frac{\int_0^\infty (1 - e^{-\beta\tau}) Q(i,j,d\tau)}{\beta \int_0^\infty e^{-\beta\tau} Q(i,j,d\tau)}
\end{aligned} \tag{28}$$

and

$$m_\beta(i) = \frac{\int_0^\infty (1 - e^{-\beta\tau}) Q(i,d\tau)}{\beta \int_0^\infty e^{-\beta\tau} Q(i,d\tau)} = \sum_{j\in\mathcal{E}} Q_\beta(i,j) m_\beta(i,j)$$

or

$$\lambda_\beta(i) = \frac{1}{m_\beta(i)} = \frac{\beta \int_0^\infty e^{-\beta\tau} Q(i,d\tau)}{\int_0^\infty (1 - e^{-\beta\tau}) Q(i,d\tau)}. \tag{29}$$

Then, we can verify that

$$\sum_{j\in\mathcal{E}} Q_\beta(i,j) = 1 \quad or \quad Q_\beta e = e$$

where Q_β is the matrix of $Q_\beta(i,j)$, and (27) becomes

$$\begin{aligned}
\eta_\beta(i) =& f_\beta(i) \int_0^\infty (1 - e^{-\beta\tau}) Q(i,d\tau) \\
&+ \sum_{j\in\mathcal{E}} \left\{ \int_0^\infty e^{-\beta\tau} Q(i,j,d\tau)\eta_\beta(j) \right\}.
\end{aligned} \tag{30}$$

where

$$\begin{aligned}
f_\beta(i) &= \frac{\sum_{j\in\mathcal{E}} \left\{ f(i,j) \int_0^\infty (1 - e^{-\beta\tau}) Q(i,j,d\tau) \right\}}{\int_0^\infty (1 - e^{-\beta\tau}) Q(i,d\tau)} \\
&= \frac{\sum_{j\in\mathcal{E}} f(i,j) Q_\beta(i,j) m_\beta(i,j)}{m_\beta(i)} \\
&= \sum_{j\in\mathcal{E}} f(i,j) p_\beta(j|i) = E_\beta \left[f(X_t, Y_t)|X_t = i \right]
\end{aligned} \tag{31}$$

where $p_\beta(j|i) = ((Q_\beta(i,j)m_\beta(i,j))/m_\beta(i))$ is the "equivalent" conditional probability of $Y_t = j$ given that $X_t = i$, and "E_β" denote the expectation under p_β. We can verify that $\sum_{j\in\mathcal{E}} p_\beta(j|i) = 1$. Dividing both sides of (30) by $\int_0^\infty (1 - e^{-\beta\tau}) Q(i,d\tau)$ yields

$$\eta_\beta(i) = f_\beta(i) - \frac{1}{\beta}\lambda_\beta(i)\eta_\beta + \frac{1}{\beta} \sum_{j\in\mathcal{E}} \lambda_\beta(i) Q_\beta(i,j)\eta_\beta(j). \tag{32}$$

Define $f_\beta = (f_\beta(1), \ldots, f_\beta(M))^T$, then (32) becomes

$$(\beta I - A_\beta)\eta_\beta = \beta f_\beta \tag{33}$$

where A_β is an infinitesimal generator with

$$A_\beta(i,j) = \begin{cases} \lambda_\beta(i) Q_\beta(i,j), & \text{if } i \neq j, \\ -\lambda_\beta(i) \left[1 - Q_\beta(i,i)\right], & \text{if } i = j. \end{cases} \tag{34}$$

We have $A_\beta e = 0$. For any infinitesimal generator A, $\beta I - A$ is invertible. (This can be shown as follows: Let $Q = I + A$. Then, Q is an ergodic Markov matrix; its eigenvalues are in the unit circle except one of them being one [2]. In addition,

$\beta I - A = (1+\beta)I - Q = (1+\beta)(I - (1/(1+\beta))Q)$. For any $\beta > 0$, all the eigenvalues of $(1/(1+\beta))Q$ are in the unit circle and those of $(I - (1/(1+\beta))Q)$ are not zero and, thus, $(I - (1/(1+\beta))Q)$ is invertible). Therefore, we have

$$\eta_\beta = \beta(\beta I - A_\beta)^{-1} f_\beta. \tag{35}$$

If the performance function $f(i,j)$ is independent of j, i.e., $f(i,j) \equiv f(i)$ for all $j \in \mathcal{E}$, then from (31), we have $f_\beta(i) = f(i)$. Thus

$$\eta_\beta = \beta(\beta I - A_\beta)^{-1} f. \tag{36}$$

B. Equivalent Markov Processes

The Markov process with infinitesimal generator A_β defined in (34) and $f_\beta(i)$ in (31) is called an *equivalent Markov process* for the SMP with discount factor β. First, if an SMP is Markov, then the equivalent Markov process with any discount factor β is the Markov process itself. Indeed, a Markov chain with transition rates $\lambda(i)$ and transition probabilities $Q(i,j)$ can be viewed as a semi-Markov process whose kernel is $Q(i,j,t) = Q(i,j)[1 - e^{-\lambda(i)t}]$. Therefore

$$\int_0^\infty e^{-\beta\tau} Q(i,j,d\tau) = \frac{Q(i,j)\lambda(i)}{\beta + \lambda(i)}, \qquad i,j \in \mathcal{E}.$$

Substituting this into (28), (29), and (31), we get

$$Q_\beta(i,j) = Q(i,j) \quad \lambda_\beta(i) = \lambda(i) \quad \text{and } f_\beta(i) = f(i)$$

for all i, j and β. Therefore, for Markov processes

$$\eta_\beta = \beta(\beta I - A)^{-1} f$$

where $f(i) = \sum_{j \in \mathcal{E}} Q(i,j) f(i,j)$ is defined in (11).

Second, since (33) for both the SMP and the equivalent Markov process are the same, the equivalent Markov process has the same η_β as the original SMP. Indeed, for the equivalent Markov process, we define the transition function $P_t(i,j) = P(X_t = j | X_0 = i)$, and the transition function matrix $P_t = [P_t(i,j)]$. By a standard result [12], we have

$$P_t = e^{A_\beta t}.$$

Therefore, for the equivalent Markov process, we have

$$\begin{aligned}
\eta_\beta(i) &= \lim_{T\to\infty} \int_0^T \beta e^{-\beta t} E_\beta \left\{ f(X_t, Y_t) | X_0 = i \right\} dt \\
&= \lim_{T\to\infty} \int_0^T \beta e^{-\beta t} \sum_{j \in \mathcal{E}} \\
&\quad \times \left\{ E_\beta \left[f(X_t, Y_t) | X_t = j \right] P(X_t = j | X_0 = i) \right\} dt \\
&= \lim_{T\to\infty} \int_0^T \beta e^{-\beta t} \sum_{j \in \mathcal{E}} \left\{ f_\beta(j) P(X_t = j | X_0 = i) \right\} dt \\
&= \lim_{T\to\infty} \int_0^T \{ \beta e^{-\beta t} e_i e^{A_\beta t} f_\beta \} dt
\end{aligned}$$

where e_i is a row vector whose components are zeros except for its ith component being 1. In matrix form, we have

$$\begin{aligned}
\eta_\beta &= \lim_{T\to\infty} \int_0^T \{ \beta e^{-\beta t} e^{A_\beta t} f_\beta \} dt \\
&= \lim_{T\to\infty} \int_0^T \left\{ \beta e^{-(\beta I - A_\beta)t} f_\beta \right\} dt = \beta(\beta I - A_\beta)^{-1} f_\beta
\end{aligned}$$

which is the same as (36).

C. Limiting Case

We will prove that when $\beta \to 0+$, all of the aforementioned results converge to those for the average-cost problem. Therefore, the average-cost problem can be viewed as a special case of the discounted-cost problem with $\beta = 0$. First, we can easily verify that the following limits exist:

$$\lim_{\beta\to 0+} \alpha_\beta(i) = 1 \tag{37}$$

$$Q_0(i,j) \equiv \lim_{\beta\to 0+} Q_\beta(i,j) = Q(i,j) \tag{38}$$

$$m_0(i,j) \equiv \lim_{\beta\to 0+} m_\beta(i,j) = m(i,j) \tag{39}$$

$$\begin{aligned}
m_0(i) &\equiv \lim_{\beta\to 0+} m_\beta(i) = m(i) \quad \text{or} \\
\lambda_0(i) &\equiv \lim_{\beta\to 0+} \lambda_\beta(i) = \lambda(i)
\end{aligned} \tag{40}$$

and

$$f_0(i) \equiv \lim_{\beta\to 0+} f_\beta(i) = f(i). \tag{41}$$

From (37)–(40), we have

$$A_0(i,j) \equiv \lim_{\beta\to 0+} A_\beta(i,j) = \begin{cases} \lambda(i) Q(i,j), & \text{if } i \neq j, \\ -\lambda(i)\left[1 - Q(i,i)\right], & \text{if } i = j. \end{cases}$$

Denoting $A \equiv A_0$ be the matrix with components $A_0(i,j)$, which is the same as (5) in the average-cost problem, we have

$$\lim_{\beta\to 0+} A_\beta = A.$$

Lemma 1: $\lim_{\beta\to 0+} \beta(\beta I - A_\beta)^{-1} = ep$, where p satisfies $pA = 0$.

Proof: It was proved that for any ergodic Markov matrix P, we have

$$\lim_{\alpha\to 1-} (1-\alpha)(I - \alpha P)^{-1} = ep$$

where p is the steady-state probability vector of P: $pP = p$, and $pe = 1$. Let $P = I + A$, then $pA = 0$. Setting $\beta = 1 - \alpha$, we have

$$\lim_{\beta\to 0+} \beta\left[\beta I - (1-\beta)A\right]^{-1} = \lim_{\alpha\to 1-} (1-\alpha)(I - \alpha P)^{-1} = ep. \tag{42}$$

Next, it is easy to verify that

$$\begin{aligned}
&\left\{ I - \beta\left[\beta I - (1-\beta)A\right]^{-1} A \right\} \beta\left(\beta I - A\right)^{-1} \\
&\qquad = \beta\left[\beta I - (1-\beta)A\right]^{-1}.
\end{aligned}$$

Letting $\beta \to 0+$ on the both sides of the aforementioned equation and using (42) and $pA = 0$, we obtain

$$\lim_{\beta \to 0+} \beta(\beta I - A)^{-1} = ep. \tag{43}$$

Next, from (28) and (29), A_β is an analytical function at $\beta = 0$. Thus

$$A_\beta = A + \frac{dA_\beta}{d\beta}\beta + o(\beta)$$

where $o(\beta)$ represents a matrix with $\lim_{\beta \to 0+}(o(\beta)/\beta) = 0$. From $A_\beta e = 0$, we have

$$\frac{dA_\beta}{d\beta} e = 0.$$

Now, from $\beta(\beta I - A_\beta) = \beta(\beta I - A) - \beta^2(dA_\beta/d\beta) - \beta o(\beta)$, we get

$$\begin{aligned}\beta(\beta I - A)^{-1} = &\, \beta(\beta I - A_\beta)^{-1} - \left[\beta(\beta I - A_\beta)^{-1}\right] \\ &\times \left\{\frac{dA_\beta}{d\beta}\right\}\left[\beta(\beta I - A)^{-1}\right] \\ &+ \left[\beta(\beta I - A_\beta)^{-1}\left\{\frac{o(\beta)}{\beta}\right\}\left[\beta(\beta I - A)^{-1}\right]\right. .\end{aligned}$$

From (43), the lemma follows directly by letting $\beta \to 0+$ on both sides of the previous equation. □

Lemma 2: The discounted performance measure and potentials defined in (24) and (25) converge to their counterparts for the long-run average problem as $\beta \to 0+$, i.e.,

$$g_0 = \lim_{\beta \to 0+} g_\beta = g \quad \textit{and} \quad \eta_0 = \lim_{\beta \to 0+} \eta_\beta = \eta e.$$

Proof: The second equation is a direct consequence of Lemma 1, (35), and (41). The first one follows from (26). □

Next, from (26), we get

$$g_\beta = (\beta I - A_\beta)^{-1} f_\beta - \frac{1}{\beta}\eta e$$

or (noting $A_\beta e = 0$)

$$(\beta I - A_\beta) g_\beta = f_\beta - \eta e \tag{44}$$

which is called the *discounted Poisson equation*. Setting $\beta = 0$ leads to the special case of (18). Let p_β be the steady-state probability of the equivalent Markov process, i.e., $p_\beta A_\beta = 0$. Then, from (33), we have

$$p_\beta \eta_\beta = p_\beta f_\beta = \eta^*_\beta$$

with η^*_β denoting the steady-state performance of the equivalent Markov process. In addition, we have

$$p_\beta g_\beta = \frac{1}{\beta}(\eta^*_\beta - \eta).$$

This applies to the particular potentials defined in (25). Of course, for any constant c, $g_\beta + ce$ is also a potential.

D. Sensitivity and Semi-Markov Decision Problems With Discounted Costs

Now, we consider the sensitivity problem. Let $Q(i, j, t)$, i, $j \in \mathcal{E}$, and $Q'(i, j, t)$, i, $j \in \mathcal{E}$, be two ergodic SMPs defined on the same state–space $\mathcal{E}$. Let the corresponding infinitesimal generators be A_β and A'_β, and their corresponding discounted performance measures be η_β and η'_β.

Theorem 1: We have

$$\eta'_\beta - \eta_\beta = \beta(\beta I - A'_\beta)^{-1}\left[(A'_\beta g_\beta + f'_\beta) - (A_\beta g_\beta + f_\beta)\right]. \tag{45}$$

Proof: From (33), we have $\beta I \eta_\beta = A_\beta \eta_\beta + \beta f_\beta$. Thus

$$\begin{aligned}\beta I(\eta'_\beta - \eta_\beta) &= \beta(f'_\beta - f_\beta) + A'_\beta \eta'_\beta - A_\beta \eta_\beta \\ &= \beta(f'_\beta - f_\beta) + (A'_\beta - A_\beta)\eta_\beta + A'_\beta(\eta'_\beta - \eta_\beta).\end{aligned}$$

Thus

$$\eta'_\beta - \eta_\beta = (\beta I - A'_\beta)^{-1}\left[(A'_\beta - A_\beta)\eta_\beta + \beta(f'_\beta - f_\beta)\right].$$

From (26) and $A'_\beta e = A_\beta e = 0$, we get

$$\eta'_\beta - \eta_\beta = \beta(\beta I - A'_\beta)^{-1}\left[(A'_\beta - A_\beta)g_\beta + (f'_\beta - f_\beta)\right]$$

which leads directly to (45). □

Theorem 1 forms the basis for the semi-Markov decision problem with discounted performance measures. It is important to note that the transition rates at any state i (i.e., the ith row of A_β) and $f_\beta(i)$ depend on only $Q(i, j, t)$ with this particular i. In other words, each action on state i controls the ith row of A_β. Thus, in policy iteration the new policy can be determined state-by-state. Specifically, at each step we first solve the discounted Poisson equation (44) for the potentials for the current policy, then choose the actions that minimize $(A^{\mathcal{L}}_\beta g_\beta + f^{\mathcal{L}}_\beta)$ componentwisely as the next policy. Of course, the performance potentials can also be estimated on sample paths.

Letting $\beta \to 0$ in the theorem and using Lemmas 1 and 2, we get

$$\eta'_0 - \eta_0 = ep'\left[(A'_0 g_0 + f') - (A_0 g_0 + f)\right].$$

Since $A'_0 = A'$ and $A_0 = A$ with A and A' defined in (5), by Lemma 2, this is equivalent to (20).

Suppose the semi-Markov kernel depends on a continuous parameter θ and is denoted as $Q_\theta(i, j, t)$, and the performance measure is a function of θ, f_θ. With a discount factor β, the equivalent infinitesimal generator becomes [see (28), (29), and (34)]

$$\begin{aligned}A_{\theta;\beta}(i, j) &= \frac{\beta \int_0^\infty e^{-\beta\tau} Q_{\theta;\beta}(i, j, d\tau)}{\int_0^\infty (1 - e^{-\beta\tau}) Q_{\theta;\beta}(i, d\tau)}, \qquad \text{for } i \neq j \\ A_{\theta;\beta}(i, i) &= -\frac{\beta \int_0^\infty e^{-\beta\tau} Q_{\theta;\beta}(i, d\tau)}{\int_0^\infty (1 - e^{-\beta\tau}) Q_{\theta;\beta}(i, d\tau)}, \qquad i \in \mathcal{E}\end{aligned}$$

in which we assume that $Q_\theta(i, i) = 0$ for convenience. Setting $Q_{\theta+\Delta\theta}(i, j, t)$, and $Q_\theta(i, j, t)$ be the two semi-Markov kernels in Theorem 1, we get

$$\begin{aligned}\eta_{\theta+\Delta\theta;\beta} - \eta_{\theta;\beta} = &\, \beta(\beta I - A_{\theta+\Delta\theta;\beta})^{-1} \\ &\times \left[(A_{\theta+\Delta\theta;\beta} g_{\theta;\beta} + f_{\theta+\Delta\theta}) - (A_{\theta;\beta} g_{\theta;\beta} + f_\theta)\right].\end{aligned}$$

Letting $\Delta\theta \to 0$, we get the derivative of the discounted performance measure

$$\frac{d\eta_{\theta;\beta}}{d\theta} = \beta(\beta I - A_{\theta;\beta})^{-1}\left\{\frac{dA_{\theta;\beta}}{d\theta} g_{\theta;\beta} + \frac{df_\theta}{d\theta}\right\}. \tag{46}$$

As a special case, we let $\beta \to 0$ and set $A_\theta = A + \theta B$, $B = A' - A$. It is easy to verify that the performance derivative at $\theta = 0$ has the same form as (21).

Define the *perturbation realization matrix* as

$$D_\beta = e g_\beta^T - g_\beta e^T.$$

D_β is skew-symmetric, i.e., $D'_\beta = -D_\beta$. The performance sensitivities (45) and (46) can be obtained by using the perturbation realization matrix. In particular, we have

$$\frac{d\eta_{\theta;\beta}}{d\theta} = \beta(\beta I - A_{\theta;\beta})^{-1}\left\{\frac{dA_{\theta;\beta}}{d\theta} D_{\theta;\beta}^T p_{\theta;\beta}^T + \frac{df_\theta}{d\theta}\right\}.$$

Equations (45) and (46) are the sensitivity formulas in a discrete-policy space and a continuous-parameter space, respectively.

From the definition, we have $D_\beta(i,j) = g_\beta(j) - g_\beta(i)$. To develop a formula for estimation, we consider two SMPs with the same kernel $Q(i,j,t)$, one starting with $X_0 = i$ and the other with $X'_0 = j$. We have

$$D_\beta(i,j) = \lim_{N\to\infty} \times E\left[\int_0^{T_N} e^{-\beta t}\left[f(X'_t, Y'_t) - f(X_t, Y_t)\right] dt | X'_0 = j, X_0 = i\right].$$

This formula is particularly useful if there is one state, denoted as i^*, at which the system's sojourn time is exponential. In this case, let $T^*(i,j)$ be the random instant such that $X'_{T^*(i,j)} = X_{T^*(i,j)} = i^*$ for the first time. By the memoryless property of the exponential distribution, the behavior of the two SMPs after $T^*(i,j)$ are the same statistically, i.e., $E[f(X'_t, Y'_t)] = E[f(X_t, Y_t)]$, for $t > T^*(i,j)$. Thus

$$D_\beta(i,j) = E\left[\int_0^{T^*(i,j)} e^{-\beta t}\left[f(X'_t, Y'_t) - f(X_t, Y_t)\right] dt | X'_0 = j, X_0 = i\right].$$

If X_t is a Markov process, then $T^*(i,j)$ can be chosen to be the first time that the two processes X_t and X'_t merge together.

From the definition of D_β, $A_\beta e = 0$, and the discounted Poisson equation, it is easy to verify that D_β satisfies

$$A_\beta D_\beta + D_\beta A_\beta^T - \beta D_\beta = -F_\beta$$

where $F_\beta = e f_\beta^T - f_\beta e^T$ or, equivalently

$$\left(A_\beta - \frac{1}{2}\beta I\right) D_\beta + D_\beta \left(A_\beta - \frac{1}{2}\beta I\right)^T = -F_\beta.$$

When $\beta = 0$, this is the same as the Lyapunov equation (16) for the average-cost problem.

V. Conclusion

We have shown that with properly defined A, g and D, the results for potentials, perturbation realization, PA, and MDP, etc., can be extended naturally to SMP with both average and discounted costs. Especially, sensitivity analysis and policy iteration for SMDP can be implemented on a sample path. Performance potentials, which play a crucial role in both sensitivity analysis and policy iteration, can be estimated by the long-run performance integration, which has the same physical meaning as for Markov processes. This approach provides a unified tool in optimization of DEDSs, including PA, MDP, and SMDP for both average- and discounted-cost problems. In addition, RL methods can be developed to estimate potentials, realization matrices, Q-factors, and performance derivatives by analyzing a sample path of a stochastic system that can be modeled as a Markov process or an SMP.

The sensitivity point of view of MDP and SMDP brings out some new thoughts; for example, can we use the performance derivative and/or higher order derivatives, which can be obtained by analyzing a single sample path of the current system, to implement optimization or policy iteration? Other research topics include extensions to more general processes such as generalized semi-Markov processes and applications to queueing networks with general service time distributions. SMDP theory also has applications in the temporal abstraction approach [22]) and the time aggregation approach [10].

Finally, many results about SMP can be obtained by using the embedded Markov chain method (see, e.g., [23]). It is natural to expect that the sensitivity analysis can also be implemented using this approach. However, compared with the embedded-chain-based approach, our approach is more direct and concise and, hence, the results have a clear interpretation. In addition, with the embedded approach, the expected values (time and cost) on a period $T_{n+1} - T_n$ are used; and our approach is easier to be implemented on a sample path [e.g., see (8)].

Appendix

VI. Proof of (3)

Consider an interval $[0, T_N]$, with $N \gg 1$. Let $I_k(x) = 1$ if $x = k$ and $I_k(x) = 0$ if $x \neq k$; and $I(*)$ be an indicator function, i.e., $I(*) = 1$ if the expression in the brackets holds, $I(*) = 0$ otherwise. From (2), by ergodicity we have

$$p_t(s|k)ds = \lim_{T_N\to\infty} \frac{\int_0^{T_N} I\left(s \le t - T_{n_t} < s + ds\right) I_k(X_t) dt}{\int_0^{T_N} I_k(X_t) dt}. \tag{47}$$

Let N_k be the number of periods in $[0, T_N]$ in which $X_t = k$. We have

$$\lim_{T_N\to\infty} \frac{1}{N_k} \int_0^{T_N} I_k(X_t) dt = \int_0^\infty s Q(k, ds). \tag{48}$$

Next, we observe that $\int_0^{T_N} I(s \le t - T_{n_t}) I_k(X_t) dt$ is the total length of the time period in $[0, T_N]$ in which $s \le t - T_{n_t}$. Furthermore, among the N_k periods, roughly $N_k Q(k, d\tau)$ periods terminate with a length in τ to $\tau + d\tau$. For any $s < \tau$, in each of such periods the length of time in which $s \le t - T_{n_t}$ is $\tau - s$. Thus

$$\int_0^{T_N} I(s \le t - T_{n_t}) I_k(X_t) dt \approx N_k \int_s^{\infty} (\tau - s) Q(k, d\tau)$$

or

$$\lim_{T_N \to \infty} \frac{1}{N_k} \int_0^{T_N} I(s \le t - T_{n_t})\, I_k(X_t) dt = \int_s^{\infty} (\tau - s) Q(k, d\tau).$$

Therefore

$$\begin{aligned} &\lim_{T_N \to \infty} \frac{1}{N_k} \int_0^{T_N} I\left(s \le t - T_{n_t} < s + ds\right) I_k(X_t) dt \\ &= - \lim_{T_N \to \infty} \frac{1}{N_k} \left\{ \int_0^{T_N} I\left(s + ds \le t - T_{n_t}\right) I_k(X_t) dt \right. \\ &\qquad \left. - \int_0^{T_N} I\left(s \le t - T_{n_t}\right) I_k(X_t) dt \right\} \\ &= - \frac{d}{ds} \left\{ \int_s^{\infty} (\tau - s) Q(k, d\tau) \right\} ds \\ &= \left[1 - Q(k, s)\right] ds = H(k, s) ds. \end{aligned} \tag{49}$$

From (47), (48), and (49), we get

$$\lim_{t \to \infty} p_t(s|k) ds = \frac{ds H(k, s)}{\int_0^{\infty} sQ(k, ds)}.$$

Therefore,

$$\lim_{t \to \infty} p_t(s|k) = \frac{H(k, s)}{m(k)}.$$

VII. Proof of (9)

Consider a time interval $[0, T_N]$, with $N \gg 1$. Let N_i be the number of periods in which the process X_t is in state i. Then

$$\lim_{T_N \to \infty} \frac{1}{N_i} \int_0^{T_N} I_i(X_t) dt = \int_0^{\infty} sQ(i, ds).$$

Let $I_{i,j}(x, y) = 1$ if $x = i$ and $y = j$, and $I_{i,j}(x, y) = 0$, otherwise. We have

$$\lim_{T_N \to \infty} \frac{1}{N_i} \int_0^{T_N} I_{i,j}(X_t, Y_t) dt = \int_0^{\infty} sQ(i, j, ds).$$

Thus, we have

$$\begin{aligned} p(j|i) &= \lim_{T_N \to \infty} \frac{\int_0^{T_N} I_{i,j}(X_t, Y_t) dt}{\int_0^{T_N} I_i(X_t) dt} = \frac{\int_0^{\infty} sQ(i, j, ds)}{\int_0^{\infty} sQ(i, ds)} \\ &= \frac{Q(i, j) m(i, j)}{m(i)} \end{aligned}$$

where

$$m(i, j) = \int_0^{\infty} sG(i, j, ds) = E[T_{n+1} - T_n | X_n = i, X_{n+1} = j]$$

and

$$m(i) = \sum_{j \in \mathcal{E}} Q(i, j) m(i, j) = \int_0^{\infty} sQ(i, ds).$$

References

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific, 1995, vol. II.

[2] A. Berman and R. J. Plemmons, "Nonnegative matrices in the mathematical sciences," *SIAM J. Numer. Anal.*, vol. 11, pp. 145–154, 1974.

[3] D. P. Bertsekas and T. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996.

[4] X.-R. Cao, *Realization Probabilities: The Dynamics of Queueing Systems*. New York: Springer-Verlag, 1994.

[5] ——, "The Maclaurin series for performance functions of Markov chains," *Adv. Appl. Probab.*, vol. 30, pp. 676–692, 1998.

[6] ——, "The relation among potentials, perturbation analysis, Markov decision processes, and other topics," *J.Discrete Event Dyna. Syst.*, vol. 8, pp. 71–87, 1998.

[7] ——, "A unified approach to Markov decision problems and performance sensitivity analysis," *Automatica*, vol. 36, pp. 771–774, 2000.

[8] ——, "From perturbation analysis to Markov decision processes and reinforcement learning," *J. Discrete Event Dyna. Syst.*, vol. 13, pp. 9–39, 2003.

[9] X.-R. Cao and H. F. Chen, "Potentials, perturbation realization, and sensitivity analysis of Markov processes," *IEEE Trans. Automat. Contr.*, vol. 42, pp. 1382–1393, Oct. 1997.

[10] X.-R. Cao, Z. Y. Ren, S. Bhatnagar, M. Fu, and S. Marcus, "A time aggregation approach to Markov decision processes," *Automatica*, vol. 38, pp. 929–943, 2002.

[11] C. Cassandras and S. Lafortune, *Introduction to Discrete Event Dynamic Systems*. Norwell, MA: Kluwer, 1999.

[12] E. Çinlar, *Introduction to Stochastic Processes*. Upper Saddle River, NJ: Prentice Hall, 1975.

[13] H.-T. Fang and X.-R. Cao, "Single sample path based recursive algorithms for Markov decision processes," *IEEE Trans. Automat. Contr.*, 2003, to be published.

[14] P. W. Glynn and S. P. Meyn, "A Lyapunov bound for solutions of Poisson's equation," *Ann. Probab.*, vol. 24, pp. 916–931, 1996.

[15] Y. C. Ho and X.-R. Cao, *Perturbation Analysis of Discrete-Event Dynamic Systems*. Norwell, MA: Kluwer, 1991.

[16] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*. New York: Van Nostrand, 1960.

[17] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. New York: Wiley, 1975.

[18] P. Marbach and T. N. Tsitsiklis, "Simulation-based optimization of Markov reward processes," *IEEE Trans. Automat. Contr.*, vol. 46, pp. 191–209, Feb. 2001.

[19] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. London, U.K.: Springer-Verlag, 1993.

[20] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: Wiley, 1994.

[21] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[22] R. S. Sutton, D. Precup, and S. Singh, "Between MDPs and Semi-MDPs: a framework for temporal abstraction in reinforcement learning," *Artif. Intell.*, vol. 112, pp. 181–211, 1999.

[23] H. C. Tijms, *Stochastic Models—An Algorithmic Approach*: Wiley, 1994.

Xi-Ren Cao (S'82–M'84–SM'89–F'96) received the M.S. and Ph.D. degrees from Harvard University, Cambridge, MA, in 1981 and 1984, respectively.

He was a Research Fellow at Harvard University from 1984 to 1986. He then worked as a Principal and Consultant Engineer/Engineering Manager at Digital Equipment Corporation, MA, until October 1993. Since then, he has been a Professor with the Hong Kong University of Science and Technology (HKUST). He is the Director of the Center for Networking at HKUST. He has held visiting positions at Harvard University, the University of Massachusetts, Amherst, AT&T Labs, NJ, the University of Maryland, College Park, the University of Notre Dame, Notre Dame, IN, Shanghai Jiaotong University, Shangai, China, Nankei University, Tianjin, China, Tsinghua University, Beijing, China, the University of Science and Technology of China, Hefei, and Tongji University, Shangai, China. He owns three patents in data and telecommunications, and published two books: *Realization Probabilities—the Dynamics of Queuing Systems* (New York: Springer Verlag, 1994) and *Perturbation Analysis of Discrete-Event Dynamic Systems* (Norwell, MA: Kluwer, 1991) (coauthored with Y. C. Ho). His current research areas include discrete-event dynamic systems, communication systems, signal processing, stochastic processes, and system optimization.

Dr. Cao 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/was an Associate Editor at Large of the IEEE TRANSACTIONS OF AUTOMATIC CONTROL, and on the Board of Governors of IEEE Control Systems Society. He is also Associate Editor of a number of international journals and chairman of a few technical committees of international professional societies.