The Relations Among Potentials, Perturbation Analysis, and Markov Decision Processes

XI-REN CAO*  
The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong  
eceso@ee.ust.hk

Received September 9, 1996; Revised October 10, 1997; Accepted November 25, 1997

Abstract. This paper provides an introductory discussion for an important concept, the performance potentials of Markov processes, and its relations with perturbation analysis (PA), average-cost Markov decision processes (MDP), Poisson equations, α-potentials, the fundamental matrix, and the group inverse of the transition matrix (or the infinitesimal generators). Applications to single sample path-based performance sensitivity estimation and performance optimization are also discussed. On-line algorithms for performance sensitivity estimates and on-line schemes for policy iteration methods are presented. The approach is closely related to reinforcement learning algorithms.

Keywords: Policy iterations, Poisson equations, α-potentials, group inverse, fundamental matrices, on-line optimization, reinforcement learning.

1. Introduction

In this paper, we show that two important research areas in discrete event dynamic systems, perturbation analysis (PA) and Markov decision processes (MDP), are closely related. The central piece of these two areas is the fundamental concept of performance potentials. We show that performance potential is very naturally related to Poisson equations, the Markov potential theory, the fundamental matrix, the group inverse of the transition matrix, and other well-known research topics. We also show that performance potentials can be estimated based on a single sample path of a discrete event system; thus single sample path based sensitivity analysis and on-line optimization can be implemented by using these potential estimates.

The research presented in this paper was a result of the author’s effort in developing on-line optimization algorithms for practical systems such as high-speed communication networks. The objective of on-line optimization is to achieve the optimal performance of a discrete event dynamic system by using the information contained in a single sample path of the system. Such a sample path can be obtained by observing the evolution of a real system, in the case of real time control and management, or by simulation, in the case of engineering system design. Markov decision process (MDP) theory is a fundamental tool in optimization; thus, on-line optimization algorithms can be developed by implementing MDP principles on a single sample path of a discrete event system to achieve the optimal performance.

A parallel research area is the single sample path-based sensitivity analysis. Perturbation analysis (PA) is one of the major techniques in this area, which provides efficient algorithms

* Supported in part by Hong Kong UGC under grant HKUST 690/95E.
that yield strongly consistent estimates for performance sensitivities of many discrete event
dynamic systems. There have been many works in this area; the results presented in this
paper represent a new direction that extends PA to a wide range of systems and links PA to
MDP and on-line optimization over a discrete decision space.

The main contribution of this paper is that it presents a simple approach that links all
the above concepts and research topics together and thus provides some new insights to
these problems. Most results about each particular topic presented in this paper are either
documented or will appear in the literature in a more general setting and with a more
rigorous approach (e.g., for performance sensitivity analysis, see (Cao and Chen, 1997; Cao
and Wan; Jaakkola et al., 1995); for MDP, (Bertsekas, 1995; Puterman, 1994; Gallager,
1995); for Markov potential theory, (Çinlar, 1975); for fundamental matrices and group
inverse, (Golub and Meyer, 1986; Kemeny and Snell, 1960; Meyer, 1975).) Our purpose
is to show that these results can be linked together very naturally through the concept of
performance potentials. As such, we shall focus on main ideas, concepts, methodologies,
and the relations among them. For most results, we shall give only intuitive explanations
and omit the lengthy and tedious proofs, if it does not affect the understanding of the key
concepts. We shall study the simplest model, i.e., irreducible and aperiodic finite chains
(or processes) with finite action spaces.

We first give a brief review to the important concept of performance potentials, which
plays a dominant role in both performance sensitivity analysis and performance optimization
problems. The performance potential is simply defined as the solution to the well-known
Poisson equations. With the performance potentials, the sensitivity of the steady state
performance of a Markov chain (or a Markov process) with respect to the changes in its
transition probability matrix (or its infinitesimal generator) can be easily obtained, and the
policy iteration algorithm of MDP can be easily derived without resorting to the standard
optimality equations. Our results reveal a fundamental relation between PA and the policy
iteration method in MDP: the policy iteration algorithm in fact chooses a policy that
represents the "steepest" direction obtained by PA as the new policy in the next step.

We then present several equations for potentials (Cao and Chen, 1997); with these equations,
a number of algorithms for estimating the potentials are developed (Cao and Wan); these algorithms are based on a single sample path of the Markov system. Using these
potential estimates in the sensitivity formulas or policy iteration algorithms, we can obtain
estimates for performance sensitivities or achieve the optimal performance by simply
analyzing a single sample path of the system.

We show that the performance potential defined in this paper can be viewed as an extension
of the $\alpha$-potentials, $0 < \alpha < 1$, defined in the modern Markov potential theory (Çinlar,
1975) to the case with the discount factor $\alpha = 1$. When $\alpha = 1$, the $\alpha$-potential for ergodic
Markov processes as defined in modern Markov theory takes an infinite value. We show,
however, the "difference" of any two component potentials is finite and these differences
have important applications in sensitivity analysis and performance optimization. In this
sense, the potential defined in this paper is the final version of the $\alpha$-potential, with $\alpha = 1$,
in the modern Markov potential theory.

Our results also relate the sensitivity analysis and optimization to the fundamental matrix
and the group inverse of the transition matrices (or infinitesimal generators). It was realized
in Kemeny and Snell (1960) and Meyer (1975) that the fundamental matrix or the group inverse of the transition matrices contains almost all the information about a Markov chain; however, their results are related to performance itself; our study provides their relation with performance sensitivities and their applications to performance optimization.

In this paper, we study Markov chains except in section 6. In section 2, we introduce the concept of performance potentials and relate it to Poisson equations, the fundamental matrix, the group inverse of the transition matrix, and the \( \alpha \) potentials. A number of equations that are based on sample paths of Markov chains are derived. In section 3, using the potentials defined in section 2, we derive equations for the performance derivatives with respect to the changes in transition matrices. These equations serve as the basis for PA of Markov chains (in a wide sense, i.e., as a single sample path-based sensitivity analysis technique). In section 4, we show that with performance potentials the policy iteration algorithms in MDP can be easily obtained. The study clearly indicates the relation between PA and MDP. In section 5, single sample path-based algorithms are developed for estimating the performance potentials; these algorithms, together with the results in sections 3 and 4, provide single sample path-based solutions to PA and MDP. Interestingly, these algorithms are in the same spirit as the reinforcement learning algorithms (see, e.g., Jaakkola, Singh and Jordan (1995)). Results for Markov processes are stated in section 6. The paper concludes with some discussions in section 7.

2. Performance Potentials

Consider a Markov chain \( \mathbf{X} = \{X_n; n \geq 0\} \) on a finite state space \( \mathcal{S} = \{1, 2, \ldots, M\} \). Let \( P = [p(i, j)]_{i=1}^{M} |_{j=1}^{M} \) be the transition probability matrix. Assume that \( \mathbf{X} \) is ergodic, i.e., irreducible and aperiodic. Then \( P \) is an irreducible stochastic matrix. Let \( \pi = (\pi(1), \pi(2), \ldots, \pi(M)) \) be a row vector representing the steady-state probability of \( \mathbf{X} \). Let \( e = (1, 1, \ldots, 1)^T \) be an M-dimensional column vector in which each component is equal to 1, then

\[
P e = e,
\]

and \( \pi \) is the unique solution to the following set of equations

\[
\pi = \pi P,
\]

and

\[
\pi e = 1.
\]

Let \( f: \mathcal{S} \rightarrow \mathcal{R} \), where \( \mathcal{R} = (-\infty, \infty) \) is the space of real numbers. \( f \) is called a performance function. We also denote \( f = (f(1), f(2), \ldots, f(M))^T \) as a column vector.
(Therefore, $f$ is used for both a function and a vector). The performance measure is defined as:

$$\eta = E_\pi(f) = \sum_{i=1}^{M} \pi(i) f(i) = \pi f,$$

(4)

where $E_\pi$ denotes the expectation with respect to the steady state measure $\pi$.

**Lemma 1** The eigenvalues of $(P - e\pi)$ for an ergodic Markov chain lie within the unit circle.

**Proof:** From $Pe = e$, 1 is an eigenvalue of $P$ with eigenvector $e$. Let $\{1, \alpha_1, \ldots, \alpha_{M-1}\}$ be the set of the eigenvalues of $P$ (some of the eigenvalues may be repeated in the list). Since the Markov chain is ergodic, it is known that $|\alpha_i| < 1$ for $i = 1, 2, \ldots, M - 1$ (Berman and Plemmons, 1994). Next, we show that $\alpha_i$, $i = 1, 2, \ldots, M - 1$, are also eigenvalues of $P - e\pi$. To see this, let $x_i$ be the eigenvector corresponding to $\alpha_i$, i.e., $Px_i = \alpha_i x_i$. If $\alpha_i \neq 0$, then $(e\pi)x_i = \frac{1}{\alpha_i}(e\pi)Px_i = \frac{1}{\alpha_i}e(\pi P)x_i = \frac{1}{\alpha_i}e\pi x_i$. Thus $e\pi x_i = 0$ since $\alpha_i \neq 1$. Now we have $(P - e\pi)x_i = Px_i = \alpha_i x_i$. That is, $\alpha_i$ is an eigenvalue of $P - e\pi$ with the same eigenvector $x_i$. If $\alpha_i = 0$, then $Px_i = 0$ and $(P - e\pi)x_i = Px_i - e\pi Px_i = 0$. That is, $\alpha_i = 0$ is also an eigenvalue of $P - e\pi$ with the same eigenvector $x_i$. Thus, all $\alpha_i$, $i = 1, 2, \ldots, M - 1$, are eigenvalues of $P - e\pi$. In addition, 0 has another eigenvalue $e$ since $(P - e\pi)e = 0$. Therefore, the eigenvalues of $P - e\pi$ are $\{0, \alpha_1, \ldots, \alpha_{M-1}\}$, with all $|\alpha_i| < 1$, $i = 1, \ldots, M - 1$.

It is easy to verify that the eigenvalues of $I - P + e\pi$ are $\{1, 1 - \alpha_1, \ldots, 1 - \alpha_{M-1}\}$; none of them is zero. Therefore, $I - P + e\pi$ is invertible. The matrix $(I - P + e\pi)^{-1}$ is called the *fundamental matrix* in (Kemeny and Snell, 1960).

Let $g = (g(1), \ldots, g(M))^T$ be a column vector satisfying the Poisson equation

$$(I - P + e\pi)g = f,$$

(5)
i.e.,

$$g = (I - P + e\pi)^{-1} f.$$  

(6)

Multiplying both sides of (5) and using (2) and (3), we obtain

$$\pi g = \pi f = \eta.$$  

(7)

We shall see that in all the applications only the differences between the components of $g$ are important; thus, we give the following definition.

**Definition 1** For any constant $c$, the vector $g + ce$, where $g$ is defined in (5), is called a performance potential vector of the Markov chain; its $k$th component is called the performance potential of state $k$.

In particular, $g$ in (6) is a potential (with $c = 0$) of the Markov chain. (7) can be viewed
as its normalizing equation. Setting \( c = -\eta \) we get another version of the potential

\[
\tilde{g} = [(I - P + e\pi)^{-1} - e\pi] f = B^# f, \tag{8}
\]

where

\[
B^# = (I - P + e\pi)^{-1} - e\pi \tag{9}
\]

is called the group inverse of \( B = I - P \) ((Meyer, 1975)). The normalizing equation for \( \tilde{g} \) is

\[
\pi \tilde{g} = 0.
\]

Let us briefly discuss the meaning of group inverse. From \( B = I - P \), (1), and (2), we have \( \pi B = 0 \) and \( Be = 0 \). For any \( \pi \), the matrix that satisfies these two equations is not unique. Define

\[
B = \{ B : \pi B = 0, \ Be = 0, \ (B + e\pi)^{-1} \ exists \}.
\]

It is not difficult to verify that \( B \) is a group under the operation of matrix multiplication. In particular, we can verify that if \( B_1 \in B \) and \( B_2 \in B \), then \( B_2B_1 \in B \):

\[
(B_2B_1 + e\pi)((B_1 + e\pi)^{-1}(B_2 + e\pi)^{-1})
= ((B_1 + e\pi)^{-1}(B_2 + e\pi)^{-1})(B_2B_1 + e\pi) = I,
\]

i.e., \( (B_2B_1 + e\pi)^{-1} = (B_1 + e\pi)^{-1}(B_2 + e\pi)^{-1} \). We can also verify the following simple facts: the identity element of the group is \( B_0 = I - e\pi \), \( B^# \) defined in (9) belongs to \( B \), i.e.,

\[
B^#e = 0, \ \pi B^# = 0,
\]

and \( B^# \) is the inverse of \( B \) in group \( B \), i.e.,

\[
BB^# = B^#B = B_0 = I - e\pi. \tag{10}
\]

From Lemma 1, we can expand the fundamental matrix into a Taylor series:

\[
(I - P + e\pi)^{-1} = \sum_{k=0}^{\infty} (P - e\pi)^k = I + \sum_{k=1}^{\infty} (P^k - e\pi). \tag{11}
\]

Note that the \((i, j)\)th entry of \( P^k \) is

\[
P^k(i, j) = \text{Prob}(X_k = j|X_0 = i).
\]

Then from (6) and (11), we obtain

\[
g(i) = \lim_{N \to \infty} \left\{ E \left[ \sum_{n=0}^{N-1} f(X_n)|X_0 = i \right] - (N - 1)\eta \right\}. \tag{12}
\]
From (8), we have

\[ \tilde{g}(i) = \lim_{N \to \infty} \left\{ E \left[ \sum_{n=0}^{N-1} f(X_n) | X_0 = i \right] - N \eta \right\} \]

\[ = \lim_{N \to \infty} E \left\{ \sum_{n=0}^{N-1} [f(X_n) - \eta] | X_0 = i \right\}. \]

(13)

Define the difference of the performance potentials of any two states as

\[ d(i, j) = g(j) - g(i) = \tilde{g}(j) - \tilde{g}(j), \quad i, j = 1, 2, \ldots, M. \]

We have

\[ d(j, i) = \lim_{N \to \infty} \left\{ E \left[ \sum_{n=0}^{N} f(X_n) | X_0 = i \right] - E \left[ \sum_{n=0}^{N} f(X_n) | X_0 = j \right] \right\}. \]

(14)

In modern Markov theory, for any \( \alpha \in [0, 1] \), the \( \alpha \)-potential of function \( f \) is defined as a function \( g_\alpha \) on \( S \) (Çinlar, 1975):

\[ g_\alpha(i) = E \left[ \sum_{l=0}^{\infty} \alpha^l f(X_l) | X_0 = i \right]. \]

It is the expected value of the discounted performance return in an infinite horizon when the Markov chain starts from \( X_0 = i \). When \( \alpha = 1 \), the 1-potential (or simply called the potential) becomes

\[ g_1(i) \equiv g(i) = E \left[ \sum_{l=0}^{\infty} f(X_l) | X_0 = i \right]. \]

(15)

The sum in (15) is usually infinity for ergodic chains. (14) shows that the differences between any two components of \( g_1 \) truncated at any finite \( N \) are finite and converge to finite values, and they can be obtained by using the fundamental matrix or the group inverse of \( B = I - P \) (see (11) and (12)). In the analogy of the potential energy, (15) takes an infinite value because it sets the reference point at infinity. (6) and (8) can be viewed as two finite versions of (15).

Equation (5) provides a simple way to compare the performance of two Markov chains. Let \( P' \) be another transition matrix, which is also assumed to be irreducible and aperiodic; let \( f' \) be the performance function and \( \pi' \) and \( \eta' \) be the corresponding steady-state probability and performance measure, respectively. Define

\[ Q = P' - P, \quad h = f' - f. \]

Then \( Qe = 0 \). Multiplying both sides of (5) with \( \pi' \) and rearranging the obtained equation, we get

\[ \pi'[I - P]g = \pi'f - \pi g. \]

(16)
Thus,
\[
\eta' - \eta = \pi'f' - \pi g
= \pi'f' - \pi'f + \pi'(I - P)g \quad \text{using (16)}
= \pi'[(P'g + f') - (Pg + f)] \quad \text{using } \pi'P' = \pi'
= \pi'(Qg + h).
\]
(17) (18)

This is the basic equation that directly leads to simple solutions to PA and MDP, which will be discussed in the next two sections.

3. Perturbation Analysis

Perturbation analysis is a single sample path-based technique that provides estimates for the performance sensitivities of discrete event dynamic systems by analyzing a single sample path of the system (see, e.g., Ho and Cao (1991)). Its fundamental method is called infinitesimal perturbation analysis (IPA). It has been shown that IPA is very efficient and yields unbiased or strongly consistent estimates for performance sensitivities of many discrete event systems. However, it is also known that IPA does not always provide such nice estimates for many other systems. In this section, we shall show that using the concept of performance potentials, we can derive a simple formula for the performance sensitivity of a Markov chain; single sample path-based algorithms can be developed based on the formula. Since the Markov model is widely applicable, the approach developed here can be applied to a wide range of systems, including those for which IPA does not work well.

Suppose that \( P \) changes to \( P' = P + Q \) with \( Qe = 0 \) and \( P' \) is a transition matrix. We set \( P_\delta = P + \delta Q \) where \( \delta > 0 \) is a small real number. Then \( P' = P_1 \) and \( P = P_0 \). Since \( P \) is irreducible and aperiodic, when \( \delta \) is small enough, all the positive items in \( P \) remain positive in \( P(\delta) \), and hence \( P(\delta) \) is also irreducible and aperiodic. Let \( f_\delta = f + \delta h \) be the performance function corresponding to \( P_\delta \), \( \pi_\delta \) be the steady-state probability of the Markov chain with \( P_\delta \), and \( \eta_\delta = \pi_\delta f_\delta \) be the performance measure. The performance derivative of \( \eta \) in the direction of \( Q \) is defined as

\[
\frac{\partial \eta}{\partial Q} = \lim_{\delta \to 0} \frac{\eta_\delta - \eta}{\delta}.
\]
(19)

Similarly, \( \frac{\partial \pi}{\partial Q} = \lim_{\delta \to 0} \frac{\pi_\delta - \pi}{\delta} \), and \( \frac{\partial P}{\partial Q} = \lim_{\delta \to 0} \frac{P_\delta - P}{\delta} = Q \). Note that \( \pi_\delta \) is a solution to \( \pi_\delta = \pi_\delta[P + \delta Q] \) and \( \pi_\delta e = 1 \); the existence of \( \frac{\partial \pi}{\partial Q} \) can be easily seen by matrix inversion.

**THEOREM 1** For ergodic chains, we have

\[
\frac{\partial \eta}{\partial Q} = \pi(Qg + h)
= \pi(QB^\# f + h).
\]
(20)
Proof: Replacing $\eta'$ with $\eta_\delta$, $Q$ with $Q\delta$, and $h$ with $h\delta$ in (18), we obtain

$$\eta_\delta - \eta = \pi_\delta (Qg + h)\delta,$$

which leads to the first equation in (20). The second equation follows by replacing $g$ by $\bar{g}$ defined in (8). \hfill \blacksquare

Higher order derivatives can also be obtained. Taking derivative of both sides of

$$\pi_\delta [I - P_\delta] = 0,$$

with respect to $\delta$ at $\delta = 0$, we have

$$\frac{d\pi}{dQ} (I - P) = \pi Q.$$

Continuing taking derivatives of both sides of the resulting equations, we obtain for any $n \geq 1$,

$$\frac{d^n\pi}{dQ^n} (I - P) = n \frac{d^{n-1}\pi}{dQ^{n-1}} Q.$$

Multiplying both sides of the above equation on the right with $B^#$ and using (10) and $\pi e = 1$, we get

$$\frac{d^n\pi}{dQ^n} = n \frac{d^{n-1}\pi}{dQ^{n-1}} Q B^#.$$

Thus,

$$\frac{d^n\pi}{dQ^n} = n!\pi (QB^#)^n.$$

(For $n = 1$, this equation was derived in (Golub and Meyer, 1986)). Finally, for any performance function $f_\delta = f + \delta h$, we have

$$\frac{d^n\eta}{dQ^n} = n!\pi (QB^#)^{n-1}[QB^#f + h].$$

Note that in (20), $Qg = Q(g + ce)$ for any constant $c$. This confirms the claim that only the difference between the potentials are important. In fact, the potential difference has a particular meaning in PA (Ho and Cao, 1991). Recall $d(i, j) = g(j) - g(i)$. Suppose that at some time $t$, the Markov chain was originally at state $i$, but for some reason, the system was perturbed to state $j$ (in PA, it is said that a perturbation from $i$ to $j$ is generated at time $t$). Then $d(i, j)$ equals the average of the difference between $\sum_{t=1}^{\infty} f(X_t)$ starting from $X_0 = i$ and the same sum starting from $X_t = j$. In other words, it represents the long term average effect of a change from state $i$ to state $j$ on the performance measure. Following
the terminology of IPA, we call $d(i, j)$ a **perturbation realization factor** (Cao, 1994). We define

$$D = eg^T - ge^T,$$

whose $(i, j)$th component is $d(i, j)$. $D$ is called a **realization matrix**. Let

$$F = ef^T - fe^T.$$

Both $D$ and $F$ are skew-symmetric matrices, i.e., $D^T = -D$ and $F^T = -F$.

**Theorem 2.** The realization matrix $D$ satisfies the Lyapunov equation

$$D - PDP^T = F. \quad (21)$$

**Proof:** From (1), (8), and (10), we have

$$D - PDP^T = (eg^T - Peg^T P^T) - (ge^T - Pge^T P^T)$$

$$= (eg^T - eg^T P^T) - (ge^T - Pge^T)$$

$$= eg^T (I - P^T) - (I - P)ge^T$$

$$= e(B^h f)^T (I - P^T) - (I - P)B^h f e^T$$

$$= ef^T (I - e\pi)^T - (I - e\pi)fe^T$$

$$= ef^T - fe^T$$

$$= F$$

Repeatedly applying $D = F + PDP^T$ $N$ times, we get

$$D = \sum_{n=0}^{N-1} P^n F(P^T)^n + P^N D(P^T)^N.$$ 

Since $\lim_{N \to \infty} P^N = e\pi$, we have $\lim_{N \to \infty} P^N D(P^T)^N = e\pi (eg^T - ge^T)\pi^T e^T = 0$. Thus, we have

$$D = \sum_{n=0}^{\infty} P^n F(P^T)^n.$$ 

Furthermore, we have

$$D^\pi^T = -D\pi^T = (B^h f e^T - ef^T B^h)^\pi^T$$

$$= B^h f (\pi e)^T - ef^T (\pi B^h)^T$$

$$= B^h f.$$ 

That is

$$\tilde{g} = D^\pi^T. \quad (22)$$
Therefore,
\[
\frac{\partial \eta}{\partial Q} = \pi [Q D^T \pi^T + h].
\] (23)

In PA, we would like to estimate the performance sensitivity by analyzing a single sample path. Thus, the remaining problem is how to estimate \( g \) or \( D \) based on a single sample path. This will be discussed in section 5. The approach developed here can be applied to many problems including those for which IPA does not work well.

4. Average Cost Markov Decision Problems

In this section, we study the infinite horizon average cost per stage Markov decision problem (MDP).

In an MDP, at any transition instant \( n \geq 0 \) of a Markov chain \( X = \{X_n, n \geq 0\} \), an action is chosen from an action space \( \mathcal{A} \) and is applied to the Markov chain. We assume that the number of actions are finite, and we only consider stationary policies. The actions that are available for \( i \in \mathcal{S} \) form a nonempty subset \( A(i) \subseteq \mathcal{A} \). A stationary policy is a mapping \( \mathcal{L} : \mathcal{S} \rightarrow \mathcal{A} \), i.e., for any state \( i \), \( \mathcal{L} \) specifies an action \( \mathcal{L}(i) \in A(i) \). Let \( \mathcal{E} \) be the policy space. If action \( \alpha \) is taken at state \( i \), then the state transition probabilities at state \( i \) are denoted as \( p^\alpha(i, j) \), \( j = 1, 2, \ldots, M \). With a policy \( \mathcal{L} \), the Markov process evolves according to the transition matrix \( P^\mathcal{L} = [p^\mathcal{L}(i)(j, j)]_{i=1}^M \).

For most stable systems in real world applications, the Markov chains under any policy are irreducible. For example, in any stable communication systems, from any state it is always possible to reach the null state where there are no packets in the system, and any state should be reachable from the null state. Therefore, in this paper we assume that the Markov chain is irreducible under any policy. In addition, for simplicity we assume that it is aperiodic; although all the results in this paper hold for periodic chains if we replace the steady state value by the corresponding time average value (Puterman, 1994). Therefore, the Markov chains considered are ergodic (Çinlar, 1975). This corresponds to the problems classified in (Puterman, 1994) as the recurrent case.

The steady state probabilities corresponding to policy \( \mathcal{L} \) is denoted as a vector \( \pi^\mathcal{L} = (\pi^\mathcal{L}(1), \ldots, \pi^\mathcal{L}(M)) \). Suppose that at each stage with state \( i \) and control action \( \alpha \in A(i) \), a cost \( f(i, \alpha) \) is incurred. The long term expected value of the average cost per stage corresponding to policy \( \mathcal{L} \) is then
\[
\eta^\mathcal{L} = \lim_{N \to \infty} \frac{1}{N} E \left\{ \sum_{n=0}^{N-1} f[X_n, \mathcal{L}(X_n)] \right\},
\] (24)

where "E" denotes the expectation on the underlying probability space of the Markov chain. For ergodic chains, the above limit exists and does not depend on the initial state. Our objective is to minimize this average cost per stage over the policy space \( \mathcal{E} \), i.e., to obtain
\[
\min_{\mathcal{L} \in \mathcal{E}} \eta^\mathcal{L}.
\]
As stated above, we assume that all the policies have irreducible transition matrices. We show that the basic results can be easily derived by using the concept of potentials and there is a close relationship between MDP and PA.

Again, we consider two ergodic Markov chains with transition matrices \( P \) and \( P' \) and cost functions \( f \) and \( f' \). Let the steady state probabilities be \( \pi \) and \( \pi' \) and the performance measures be \( \eta \) and \( \eta' \), respectively.

For two \( M \)-dimensional vectors \( a \) and \( b \), we define \( a = b \) if \( a(i) = b(i) \) for all \( i = 1, 2, \ldots, M \); \( a \preceq b \) if \( a(i) < b(i) \) or \( a(j) = b(j) \) for all \( i = 1, 2, \ldots, M \); \( a < b \) if \( a(i) < b(i) \) for all \( i = 1, 2, \ldots, M \); and \( a \preceq b \) if \( a(i) < b(i) \) for at least one \( i \), and \( a(j) = b(j) \) for other components. The relation \( \preceq \) includes =, \( \preceq \), and \( < \). Similar definitions are used for the relations \( >, \succeq \), and \( \succeq \). The next lemma follows directly from (17) and the fact that \( \pi'(i) > 0 \) for all \( i = 1, 2, \ldots, M \).

**Lemma 2.** If \( Pg + f \preceq P'g + f' \), then \( \eta \preceq \eta' \).

It is interesting to note that in the lemma, we use only the potentials with one Markov chain, i.e., \( g \). Thus, to compare the performance measures under two policies, only the potentials with one policy is needed.

Now we add superscripts to distinguish policies. The cost function at state \( i \) for any policy \( \mathcal{L} \) is then \( f[i, \mathcal{L}(i)] \). Define \( f^\mathcal{L} = (f[1, \mathcal{L}(1)], \ldots, f[M, \mathcal{L}(M)])^T \). (5) and (7) become

\[
(I - P^\mathcal{L} + e\pi^\mathcal{L})g^\mathcal{L} = f^\mathcal{L},
\]

and

\[
\pi^\mathcal{L}g^\mathcal{L} = \pi^\mathcal{L}f^\mathcal{L}.
\]

The following optimality theorem follows almost immediately from Lemma 2 ((Gallager, 1995) essentially uses the same approach for this result).

**Theorem 3.** A policy \( \mathcal{L} \) is optimal if and only if

\[
P^\mathcal{L}g^\mathcal{L} + f^\mathcal{L} \preceq P'^\mathcal{L}g^\mathcal{L} + f'^\mathcal{L}
\]

for all \( \mathcal{L}' \in \mathcal{E} \).

The optimality condition (27) is, of course, equivalent to the other conditions in the literature. To see this, we rewrite (25) in the following form:

\[
\eta^\mathcal{L}e + g^\mathcal{L} = f^\mathcal{L} + P^\mathcal{L}g^\mathcal{L}.
\]

Then Theorem 3 becomes: A policy \( \mathcal{L} \) is optimal if and only if

\[
\eta^\mathcal{L}e + g^\mathcal{L} = \min_{\mathcal{L}' \in \mathcal{E}} \{P^\mathcal{L}'g^\mathcal{L} + f'^\mathcal{L}\}.
\]

The minimum is taken for every component of the vector. (29) is the optimality equation, or the Bellman equation. From (29), \( g^\mathcal{L} \) is equivalent to the "differential" or "relative cost vector" in (Bertsekas, 1995), or the "bias" in (Puterman, 1994). In our approach, \( g \) is directly
related to the long term expected performance, and many results, such as the existence and the uniqueness of Equation (28), the optimality condition (29), and the convergence of the optimal algorithms, become almost obvious. In addition, as shown in the next section, \( g^L \) can be easily estimated based on a single sample path. This is an important feature and certainly can be used to implement the optimization algorithms on real world systems.

Policy iteration algorithms for determining the optimal policy can be easily developed by combining Lemma 2 and Theorem 3. We shall not state these algorithms here because they are standard.

Finally, we relate MDP to PA. Consider two Markov chains with \( P', \eta' \) and \( P \) and \( \eta \). Let \( P' = P + Q \) and \( P_\delta = P + \delta Q \). Then \( P_\delta \) corresponds to the randomized policy that chooses \( P \) with probability \( 1 - \delta \) and \( P' \) with probability \( \delta \) (see, e.g., Ross (1983)). We have derived

\[
\frac{\partial \eta}{\partial Q} = \pi'(Qg + h),
\]

and

\[
\eta' - \eta = \pi'(Qg + h).
\]

We note that the difference between the derivative (30) and the finite difference (31) is only that \( \pi \) in (30) is replaced by \( \pi' \) (31). Since the components of both \( \pi \) and \( \pi' \) are positive, we conclude that the policy iteration algorithm in fact chooses the "steepest" direction (with the largest \( \frac{\partial \eta}{\partial Q} \)) to go in the next step.

5. On-Line Algorithms

If there is no confusion, we shall use the same notation \( g \) for different versions of the potentials. From (13), we can choose a large integer \( N \) and use \( E \{ \sum_{n=0}^{N-1} [f(X_n) - \eta]|X_0 = i \} \) to approximate \( g(i) \). Furthermore, we may discard the constant term \( N \eta \) in the expression and simply use

\[
g(i) \approx E \left\{ \sum_{n=0}^{N-1} f(X_n)|X_0 = i \right\}.
\]

Alternatively, we can first estimate \( d(i, j) \) and then apply (22) to get an estimate for \( g(i) \).

Consider two independent Markov chains \( X \) and \( X' \), both have the same transition probability matrix \( P \). The two Markov chains, however, start from two different initial states, \( X_0 = i \) and \( X'_0 = j \), respectively. Define \( N_{ij} = \min\{n : n \geq 0, X_n = X_n' \} \), i.e., at \( n = N_{ij} \) the two chains merge together for the first time. Then we have

\[
d(i, j) = E \left\{ \sum_{n=0}^{N_{ij}} [f(X'_n) - f(X_n)]|X_0 = i, X'_0 = j \right\}.
\]

Intuitively, (33) can be obtained from (14) by noting that starting from \( N_{ij}, X \) and \( X' \) will behave statistically similar.
Another formula for \( d(i, j) \) can be derived as follows. Consider a Markov chain \( X = \{X_n, n \geq 0\} \) starting with \( X_0 = i \). Let \( L_i(j) = \min\{n : n \geq 0, X_n = j\} \), i.e., at \( n = L_i(j) \), the Markov chain reaches state \( j \) for the first time. From (Çinlar, 1975), \( E[L_i(j)|X_0 = i] < \infty \). We have

\[
d(i, j) = E \left\{ \sum_{n=0}^{L_i(j)-1} [f(X_n) - \eta]|X_0 = i \right\}.
\]

(34) relates \( d(i, j) \) to a finite portion of the sample paths of \( X \). An intuitive explanation can be obtained as follows. In (14), we choose \( N \) large enough so that \( N > L_i(j) \) almost holds. Then

\[
E \left[ \sum_{n=0}^{N} f(X_n)|X_0 = i \right] - E \left[ \sum_{n=0}^{N} f(X_n)|X_0 = j \right]
\]

\[
\approx E \left[ \sum_{n=0}^{L_i(j)-1} f(X_n)|X_0 = i \right] + E \left[ \sum_{n=L_i(j)}^{N} f(X_n)|X_{L_i(j)} = j \right]
\]

\[
- E \left[ \sum_{n=0}^{N-L_i(j)} f(X_n)|X_0 = j \right] - E \left[ \sum_{n=N-L_i(j)+1}^{N} f(X_n)|X_0 = j \right]
\]

The second and third terms cancel each other, because both periods start with the same state \( j \) and have the same length. The first term is the same as the first term in (34). As \( N \) goes to infinity, the Markov chain tends to its steady state; hence the last term converges to \( E[f(X_n)] \times E[L_i(j)] = \eta E[L_i(j)] \).

On-line estimation algorithms can be developed based on either (32), (33), or (34). As an example, we describe an algorithm based on (34). We observe a sample path starting from \( X_0 = i \). Let \( u_0 = 0 \), and \( u_{k+1} = \min\{n : n > u_k, X_n = i\}, k \geq 0 \). Then \( u_k, k \geq 0 \) are regenerative points. For any \( j \neq i \), define \( v_k(j) = \min\{n : u_{k+1} > n > u_k, X_n = j\} \) and \( \chi_k(j) = 1 \), if \( \{u_{k+1} > n > u_k, X_n = j\} \neq \emptyset \); and \( \chi_k(j) = 0 \), otherwise.

Equation (34) can be rewritten as

\[
d(i, j) = E \left\{ \sum_{n=0}^{L_i(j)-1} f(X_n)|X_0 = j \right\} - E[L_j(i)]\eta.
\]

From (35), we have

\[
d(i, j) = \lim_{K \to \infty} \frac{1}{\sum_{k=0}^{K-1} \chi_k(j)} \left\{ \sum_{k=0}^{K-1} \chi_k(j) \sum_{n=v_k(j)}^{u_{k+1}-1} f(X_n) \right\} - \left[ \sum_{k=0}^{K-1} \chi_k(j)[u_{k+1} - u_k(j)] \right] \eta \right\}, \quad w.p.1,
\]

where \( \eta \) can be simply estimated by

\[
\eta = \lim_{L \to \infty} \frac{1}{L} \sum_{l=0}^{L-1} f(X_l), \quad w.p.1.
\]
The algorithm based on (33) requires the comparison of two sample paths. Applying (33) to PA essentially explains the approaches developed in Dai (1994) and Dai and Ho (1995); a similar result was also presented in Fu and Hu (1994).

Using the potential estimates obtained by the algorithms described above in the formulas and algorithms developed in the previous two sections, we can obtain the performance derivatives or implement policy iteration in MDP to obtain an optimal policy. This approach is based on analyzing a single sample path, which can be obtained by simulation or observing the evolution of a real system. The algorithm based on (36) produces unbiased estimates. There is, however, still one issue remaining regarding the practical applications of the approach: the “curse of dimensionality.” That is, when the state space is too large, which is often the case for real systems, to estimate all \( d(i, j) \) or \( g(i) \) accurately may not be feasible. In particular, when the state space becomes larger, each state is visited less frequently and therefore it requires a longer sample path; in addition, \( E(L_j(i)) \) becomes larger and the variance of the estimate becomes larger. Further research is needed to determine the effect of the estimation error on the implementation of MDP (e.g., stopping criteria) and the performance that can be achieved by this approach.

For performance derivatives, we can develop an algorithm that yields \( \pi Q g \) directly without estimating individual components of \( g \). This is based on (32) and the importance sampling technique in simulation. Let us briefly introduce this method. First, from (32) we observe that \( \pi_i p_{ij} g_j \) can be estimated according to the following equation:

\[
\pi_i p_{ij} g_j \approx \lim_{L \to \infty} \frac{1}{L - N + 1} \left\{ \sum_{i=0}^{L-N} \epsilon^i(X_i) \epsilon^j(X_{i+1}) \left[ \sum_{k=0}^{N-1} f(X_{i+k+1}) \right] \right\}, \quad w.p.1,
\]

(37)

where \( \epsilon^i(x) = 1 \) if \( x = i \), and \( \epsilon^i(x) = 0 \) otherwise. Applying importance sampling, we obtain

\[
\pi_i q_{ij} g_j \approx \lim_{L \to \infty} \frac{1}{L - N + 1} \left\{ \sum_{i=0}^{L-N} \epsilon^i(X_i) \epsilon^j(X_{i+1}) \frac{q_{ij}}{p_{ij}} \left[ \sum_{k=0}^{N-1} f(X_{i+k+1}) \right] \right\}, \quad w.p.1.
\]

Finally, we have

\[
\frac{\partial \eta}{\partial Q} = \pi Q g
\]

\[
\approx \lim_{L \to \infty} \frac{1}{L - N + 1} \sum_i \sum_j \left\{ \sum_{i=0}^{L-N} \epsilon^i(X_i) \epsilon^j(X_{i+1}) \frac{q_{ij}}{p_{ij}} \left[ \sum_{k=0}^{N-1} f(X_{i+k+1}) \right] \right\}
\]

\[
= \lim_{L \to \infty} \frac{1}{L - N + 1} \left\{ \sum_{i=0}^{L-N} \sum_j \left\{ \epsilon^i(X_i) \epsilon^j(X_{i+1}) \frac{q_{ij}}{p_{ij}} \left[ \sum_{k=0}^{N-1} f(X_{i+k+1}) \right] \right\} \right\}
\]

\[
= \lim_{L \to \infty} \frac{1}{L - N + 1} \left\{ \sum_{i=0}^{L-N} \left( \frac{q_{x_i x_{i+1}}}{p_{x_i x_{i+1}}} \right) \left[ \sum_{k=0}^{N-1} f(X_{i+k+1}) \right] \right\}, \quad w.p.1. \quad (38)
\]

Because the potentials are estimated by a finite length \( N \), the derivative estimate in (38) is biased. Further discussions and numerical examples can be found in (Cao and Wan).
"Curse of dimensionality" is a common issue in MDP. There are some recent works in addressing this issue. One promising approach is the "Neuro-Dynamic Programming" (see Bertsekas and Tsitsiklis (1996) and Tsitsiklis and Roy (1996)). The results in this paper indicates another feasible approach: it is possible to develop an on-line algorithm which compares the performance of any two policies using a single sample path without estimating all the potentials. This is along the same direction as (38) for performance derivatives. The details of such algorithms and its extension and application are being investigated. It is also interesting to note that our approach is closely realated to the reinforcement learning algorithm presented in Jaakkola, Singh and Jordan (1995); when the Markov process is completely observable, the algorithm in Jaakkola, Singh and Jordan (1995) is in the same spirit as our on-line algorithm (32).

6. Results for Markov Processes

In this section, we extend the results to continuous-time Markov processes. Consider a positive recurrent irreducible Markov process $X = (X_t, t \geq 0)$, with a finite state space $S = \{1, 2, \ldots, M\}$ and infinitesimal generator $A = [a(i, j)]$, $a(i, i) < 0$ and $a(i, j) \geq 0$, $i \neq j$, $i, j = 1, 2, \ldots, M$. Let $\pi = (\pi(1), \ldots, \pi(M))$ be the vector of the steady state probabilities; we have

$$Ae = 0, \quad \pi A = 0, \quad \pi e = 1. \quad (39)$$

The fundamental matrix is now $(-A + e\pi)^{-1}$, and the potential vector $g$ is determined by

$$(-A + e\pi)g = f. \quad (40)$$

If we define $P = I + A$, then (39) and (40) are exactly the same as (1–3) and (5). Thus, all the results derived so far for Markov chains can be directly translated to Markov processes. This is the "uniformization" approach to the continuous-time problem. However, uniformization usually requires adding additional transitions from a state to itself on the sample paths; these transitions are not observable for real world systems. Thus, it is impossible to estimate $g$ based on sample paths observed from a real system by using the formulas developed from the uniformization approach. (The approach can be used in simulation, though.) In the following, we state results obtained by using (39) and (40) directly (Cao and Chen, 1997).

Let $d(i, j) = g(j) - g(i)$. Similar to (14), we have

$$d(i, j) = \lim_{T \to \infty} \left\{E \left[ \int_0^T f(X_t)dt | X_0 = j \right] - E \left[ \int_0^T f(X_t)dt | X_0 = i \right] \right\}, \quad i, j \in S. \quad (41)$$

We define the first passage time from state $i$ to state $j$, $S_i(j)$, as follows:

$$S_i(j) = \inf \{ t : t \geq 0, X_t = j | X_0 = i \}.$$
$S_i(j)$ is a stopping time of $X$ and therefore is observable. We have

$$d(i, j) = E \left\{ \int_0^{S_i(j)} [f(X_t) - \eta] dt | X_0 = i \right\}, \quad i, j \in S. \tag{42}$$

In addition, the realization matrix $D$ satisfies the Lyapunov equation (continuous version)

$$AD + DA^T = -F,$$

where $F = ef^T - fe^T$.

Finally, for two Markov processes $A', \eta', f'$, and $A, \eta, f$, with $A' = A + Q, f' = f + h$, set $A' = A + \delta Q$ and $f' = f + \delta h$. From (40), we have

$$\frac{\partial \eta}{\partial Q} = \pi(Qg + h)$$

$$= \pi(-QA^#f + h). \tag{43}$$

In MDP, an action is taken at each transition instant from a finite action space $A$. We have the basic equation

$$\eta' - \eta = \pi'( (A'g + f') - (Ag + f)) \tag{44}.$$ 

Results similar to Lemma 2 and Theorem 3 can be derived by using (44). On-line policy iteration optimization algorithms can be derived by applying (41) or (42) to obtain estimates of the potentials. These algorithms are based on infinitesimal generators $A$ and can be implemented on a single sample path of a Markov process.

7. Conclusion

We have showed that the performance potential plays a crucial role in sensitivity analysis and performance optimization. This important concept naturally links PA, the average cost MDP, the fundamental matrix, the group inverse of the transition matrix, and the $\alpha$ potentials, together. The potentials can be estimated from a single sample path of a Markov chain (or process), and this property leads to many on-line sensitivity estimation and on-line optimization schemes; these schemes are closely related to reinforcement learning algorithms.

Acknowledgment

The author thanks one of the reviewers for pointing out the relations among the results in this paper and those in Jaakkola, Singh and Jordan (1995) and Gallager (1995).
References


