A unified approach to Markov decision problems and performance sensitivity analysis

Xi-Ren Cao

Department of Electrical and Electronic Engineering, The Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong

Received 27 April 1999; revised 18 September 1999

Abstract

We propose a simple approach that provides a unified formulation for the performance sensitivity analysis of Markov chains and Markov decision problems with both infinite horizon average-cost and discounted performance criteria. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Discounted Poisson equations; a-potentials; Discounted Lyapunov equations; Perturbation analysis

1. Introduction

In this note, we propose a simple approach that provides a unified formulation for the performance sensitivity analysis of Markov chains and Markov decision problems (MDP) with both infinite horizon average-cost and discounted performance criteria. We first study the effect of the change in the transition probability matrix of a Markov chain on its performance; the main results then follow naturally. This approach is based on the discounted Poisson equation, an extension of the Poisson equation that has been studied widely in the literature (Glynn & Meyn, 1996; Makowski & Shwartz, 1994; Meyn & Tweedie, 1993). The results for MDP are the same as the standard ones (Bertsekas, 1995; Puterman, 1994), and the results for sensitivity analysis can be viewed as extensions of the recent developments in perturbation analysis (Glasserman, 1991; Ho & Cao, 1991).

Perturbation analysis is a technique that provides performance sensitivities based on a single-sample path of a discrete event dynamic system. Most works have been on continuous variables (Glasserman, 1991; Ho & Cao, 1991). Recently, the theory has been extended to Markov systems where the parameters of interest are transition matrices representing different policies; relations between perturbation analysis and Markov decision processes with long-term average cost are found (Cao, 1998). The results in this paper also cover the case with discounted performance criteria.

2. The problem

We study the infinite horizon performance criteria for discrete time Markov chains. For simplicity, we assume that the Markov chain, denoted as X = \{X_t, n \geq 0\}, has a finite state space \( S = \{1, 2, \ldots, M\} \). Let \( P = \{p(i,j)\}_{i,j=1}^{M} \) be the transition matrix, and \( \pi = (\pi(1), \pi(2), \ldots, \pi(M)) \) the row vector of the steady-state probabilities. We also assume that \( P \) is irreducible and aperiodic so that the Markov chain is positive recurrent; thus, \( \pi(i) > 0 \) for all \( i \). We have \( \pi P = \pi \). Let \( f(x), x \in S \), be a performance function and \( \alpha, 0 < \alpha \leq 1 \), be a discount factor. The performance cost is defined as a column vector \( \eta = (\eta_1(1), \eta_1(2), \ldots, \eta_1(M)) \), where \( ^{\top} \) denotes transpose, and for \( 0 < \alpha < 1 \)

\[
\eta_1(i) = (1 - \alpha)^E \left\{ \sum_{s=0}^{\infty} \alpha^s f(X_s) | X_0 = i \right\}.
\]
The factor \((1 - \alpha)\) in (1) is used to obtain the continuity of \(\eta_n\) at \(\alpha = 1\). In fact, we define

\[ \eta_1 = \lim_{\alpha \to 1^-} \eta_n. \]

It is proved later (in (3) and (11)) that the above limit exists. Let \(e = (1, 1, \ldots, 1)\) be an \(M\)-dimensional column vector with all its components being one.

**Lemma 1.** \(\eta_1 = \eta_0\) with \(\eta\) being the average-cost performance.

\[ \eta = \eta_0 = \lim_{N \to \infty} \left\{ E \left[ \frac{1}{N} \sum_{n=0}^{N-1} f(X_n) \right] \right\}. \tag{2} \]

**Proof.** In a matrix form, we have

\[ \eta_n = (1 - \alpha) \sum_{n=0}^{\infty} \alpha^n P f = (1 - \alpha)(I - \alpha P)^{-1} f, \]

\[ 0 < \alpha < 1. \tag{3} \]

The second equation in (3) holds because for \(0 < \alpha < 1\) all the eigenvalues of \(\alpha P\) are within the unit circle (Berman & Plemmons, 1994). The lemma then follows directly from (11) proved in the next section. □

Suppose that the transition matrix \(P\) and the performance function \(f\) change to \(\tilde{P}\) and \(\tilde{f}\), respectively, with \(\tilde{P}\) being another irreducible and aperiodic transition matrix. Our focus is to study the change in performance, \(\tilde{\eta}_n - \eta_n\).

3. **Discounted Poisson equations**

The discounted Poisson equation is defined as

\[ (I - \alpha P + \alpha e f) \eta_n = f. \tag{4} \]

\(\eta_n\) is called the \(\alpha\)-potential. When \(\alpha = 1\), it is the standard Poisson equation, its solution is simply called the potential, which is the same as the relative cost in Bertsekas (1995) or the bias in Puterman (1994). From (4), we have

\[ \eta_n = (I - \alpha P + \alpha e f)^{-1} f \]

\[ = \left\{ \sum_{n=0}^{\infty} \alpha^n (P - e f) \right\} f \]

\[ = \left\{ I + \sum_{n=1}^{\infty} \alpha^n (P - e f) \right\} f, \quad 0 < \alpha \leq 1. \tag{5} \]

The above expansion holds because the eigenvalues of \(P - e f\) are all in the unit circle. In particular,

\[ g_1 = (I - P + e f)^{-1} f = \left\{ I + \sum_{n=1}^{\infty} (P - e f) \right\} f. \]

This is the same as the performance potentials defined for perturbation analysis (Cao & Chen, 1997) and the average cost MDP. \((I - P + \alpha e)^{-1}\) is called the fundamental matrix (Kemeny & Snell, 1990). It is easy to verify the following equations:

\[ \pi(I - \alpha P + \alpha e f)^{-1} = \eta_n \tag{6} \]

\[ (I - \alpha P + \alpha e f)^{-1} e = e, \tag{7} \]

\[ (I - \alpha P)^{-1} e = \frac{1}{1 - \alpha} e \tag{8} \]

and

\[ (I - \alpha P)^{-1} = (I - \alpha P + \alpha e f)^{-1} + \frac{\alpha}{1 - \alpha} e. \tag{9} \]

Eq. (9) is obtained by using (8), (6), and the following equation:

\[ (I - \alpha P)^{-1}(I - \alpha P + \alpha e f) = I + (I - \alpha P)^{-1} \alpha e f. \]

In addition, we have

\[ \lim_{\alpha \to 1^-} \eta_n = \eta_0, \tag{10} \]

From (9), we obtain

\[ \lim_{\alpha \to 1^-} (I - \alpha(P - e f))^{-1} e = e. \tag{11} \]

**Lemma 2.** We have

\[ \tilde{\eta}_n - \eta_n = (1 - \alpha)(I - \alpha \tilde{P})^{-1} \left\{ \tilde{f} - \alpha \tilde{P} \eta_n \right\} - \left\{ f + \alpha P \eta_n \right\}, \quad 0 < \alpha \leq 1, \tag{12} \]

and

\[ \tilde{\eta}_n - \eta_n = \tilde{\eta} \left\{ \tilde{f} - \alpha \tilde{P} \eta_n \right\} - \left\{ f + P \eta_n \right\}. \tag{13} \]

**Proof.** From (3), we have

\[ \tilde{\eta}_n - \eta_n = (1 - \alpha)(\tilde{f} - f) + \alpha(\tilde{P} \eta_n - P \eta_n), \]

\[ = (1 - \alpha)\tilde{f} - f + \alpha(\tilde{P} - P) \eta_n + \alpha(\tilde{P} \eta_n - P \eta_n). \]

This leads to

\[ \tilde{\eta}_n - \eta_n = (1 - \alpha)(\tilde{P} - P)^{-1} (\tilde{f} - f) \]

\[ + \alpha(\tilde{P} - P)^{-1} (\tilde{f} - f) + \alpha(\tilde{P} - P)^{-1} (\tilde{f} - f) + \alpha(\tilde{P} - P)^{-1} (\tilde{f} - f). \tag{14} \]

From (3) and (9), we obtain

\[ \eta_n = (1 - \alpha) \eta_0 + \alpha e. \]

Substituting this into the right-hand side of (14) and noting that \((\tilde{P} - P)e = 0\), we obtain (12). Finally, letting \(\alpha \to 1\) in (12) and using (11), we get (13). □

This lemma leads to the fundamental results in Markov decision problems (MDP) and performance sensitivity analysis, which are discussed in the following two sections.
4. Markov decision problems

For two $M$-dimensional vectors $a$ and $b$, we define $a + b$ if $a(i) = b(i)$ for all $i = 1, 2, \ldots, M$; $a < b$ if $a(i) < b(i)$ for all $i = 1, 2, \ldots, M$; and $a \leq b$ if $a(i) \leq b(i)$ for all $i = 1, 2, \ldots, M$; and $a \leq b$ if $a(i) < b(i)$ for at least one $i$, and $a \leq b$ for other components. Similar definitions are used for the relations $>$, $\geq$, and $\geq$.

**Theorem 1.** If $\alpha + xPQg_{\alpha} \leq f + xPQg_{\lambda}$, then $\tilde{\eta}_{\alpha} = \eta_{\alpha}, 0 < \alpha \leq 1$; if $\alpha + xPQg_{\alpha} \leq f + xPQg_{\lambda}$, then $\tilde{\eta}_{\alpha} = \eta_{\alpha}, 0 < \alpha \leq 1$.

**Proof.** For $0 < \alpha < 1$, we have

\[
(I - \alpha P)^{-1} = I + \alpha P + \alpha^2 P^2 + \cdots.
\]

Since the Markov chain is positive recurrent, every item in $(I - \alpha P)^{-1}$ is positive for the theorem for $0 < \alpha < 1$ then follows directly from (12); for $\alpha = 1$ it follows from (13) and the fact that $\tilde{x} > 0$.

In MDP, there is an action space denoted as $\mathcal{A}$, which we assume to be finite. At any time $n \geq 0$, an action is taken from $\mathcal{A}$ and is applied to the Markov chain. The actions that are available for $i \in \mathcal{S}$ form a nonempty subset $\mathcal{A}(i) \subseteq \mathcal{A}$. A time-invariant policy is a mapping $\mathcal{S} \times \mathcal{A} \rightarrow \mathcal{A}$, i.e., for any state $i$, $\mathcal{A}(i)$ specifies an action $\mathcal{A}(i) \in \mathcal{A}(i)$. Let $\mathcal{A}$ be the policy space. If action $a$ is taken at state $i$, then the state transition probabilities at state $i$ are denoted as $P(i,j), j = 1, 2, \ldots, M$. With a policy $\mathcal{A}$, the Markov process evolves according to the transition matrix $P^\pi = \{P(i,j)\}^M_{i,j=1}$. We use superscript $\pi$ to denote the quantities associated with policy $\mathcal{A}$.

Theorem 1 can be used to compare the performance of two policies. Policy iteration algorithms for MDP with $0 < \alpha \leq 1$ can be derived easily from this theorem: at each iteration, we choose the policy with the largest (component-wise) $f^\pi + xP^\pi g_{\alpha}$ as the policy for the next iteration; the procedure stops at the optimal when no improvement can be achieved.

5. Performance sensitivity

The concept of perturbation realization plays an important role in perturbation analysis (Cao, 1994; Cao & Chen, 1997). We define the perturbation realization matrix as

$$D_{\alpha} = eg_{\alpha} - g_{\alpha}e, \quad 0 < \alpha \leq 1.$$  

From this equation and (10), we have

$$D_{\alpha} = g_{\alpha} - \eta c.$$  

**Theorem 2.** The realization matrix satisfies the (discounted) Lyapunov equation:

$$-D_{\alpha} + xPD_{\alpha}P = -F,$$

where $F = e^T - fc$.  

**Proof.** We have

$$-D_{\alpha} + xPD_{\alpha}P = -(eg_{\alpha} - g_{\alpha}e) + x[P(g_{\alpha} - g_{\alpha}e)P]$$

$$= -(eg_{\alpha} - g_{\alpha}e) + x[ eg_{\alpha}P(e_i - P)g_{\alpha}e]$$

$$= -[e_{\alpha}g_{\alpha} - xP_{\alpha}g + x\eta e_{\alpha}e']$$

$$= -F.$$  

Eq. (16) reduces to the standard Lyapunov equation when $\alpha = 1$ (Cao & Chen, 1997).

Next, we let $Q = \hat{P} - P, h = \hat{f} - f$, and $P(\delta) = P + \delta\hat{Q}$ and $f(\delta) = f + \delta\hat{h}$. Thus, $Qe = 0, P(1) = \hat{P}, P(0) = P, f(1) = f, f(0) = \hat{f}$. All the quantities corresponding to the Markov chains with $P(\delta)$ and $f(\delta)$ are functions of $\delta$.

The performance derivative is defined as

$$\frac{dn_{\pi}}{dQ} = \lim_{\delta \to 0} \frac{n_{\pi}(\delta) - n_{\pi}(0)}{\delta}.$$  

**Theorem 3.** We have

$$\frac{dn_{\pi}}{dQ} = (1 - \alpha)(I - \alpha P)^{-1}[Q(I - \alpha P + \alpha n_{\pi})^{-1}f + h]$$

$$= (1 - \alpha)(I - \alpha P)^{-1}[QD_{\pi}g + h]$$

$$= (1 - \alpha)(I - \alpha P)^{-1}[QD_{\pi} + h], \quad 0 < \alpha < 1$$  

and

$$\frac{dn_{\pi}}{dQ} = \tau(Q(I - P + \alpha n_{\pi})^{-1}f + h)$$

$$= \tau(QD_{\pi}g + h)$$

$$= \tau(QD_{\pi} + h).$$  

**Proof.** The equations follow directly from (4), (12), (15), and $Qe = 0$.

From the theorem, we have

$$\frac{dn_{\pi}}{dQ} = \frac{dn_{\pi}}{d\eta} \cdot \frac{dn_{\pi}}{dQ} = \frac{dn_{\pi}}{dQ} \cdot \frac{dn_{\pi}}{d\eta}, \quad \alpha \rightarrow 0$$

All the applications of the performance potentials $\tau_{\alpha}$ depend only on the differences of the components of $\tau_{\alpha}$. In other words, we can replace $\tau_{\alpha}$ with $\tau_{\alpha} + ce$, where $c$ is any constant. In particular, after replacing $\tau_{\alpha}$ with $\tau_{\alpha} + c$, $D_{\alpha}$ takes the same form and Lemma 2, Theorems 1–3, still hold. This situation is similar to the potential energy in physics and thus the terminology is justified. For any $c, \tau_{\alpha} + ce$ is called a version of the $\alpha$-potential.
6. Single sample path-based estimation

Equation (5) is equivalent to

\[ g_0(t) = E \left\{ \sum_{k=0}^{\infty} \alpha_k \delta(t - X_k) \mid X_0 = i \right\} , \quad 0 < \alpha \leq 1. \]

This can be approximated by a finite sum:

\[ g_0(t) \approx E \left\{ \sum_{k=0}^{n} \alpha_k \delta(t - X_k) \mid X_0 = i \right\} , \quad 0 < \alpha \leq 1. \]

As commented in the last section, the constant term \( E[\sum_{k=0}^{\infty} \alpha_k \delta^k] \) can be removed from the potentials, and we obtain

\[ g_0(t) \approx E \left\{ \sum_{k=0}^{n} \alpha_k \delta(t - X_k) \mid X_0 = i \right\} , \quad 0 < \alpha \leq 1. \]

(19)

From this equation, we can estimate \( g_0(t) \) by taking the average of \( \sum_{k=0}^{n} \alpha_k \delta(t - X_k) \) on the segments of a sample path, where each segment starts with state \( i \) and consists of \( N + 1 \) consecutive states.

Another way to obtain \( g_0 \) is to estimate \( D_\alpha \) by using

\[ D_\alpha(i,j) = g_\alpha(j) - g_\alpha(i) \]

\[ = E \left\{ \sum_{k=0}^{\infty} \alpha_k \delta(t - X_k) - f(X_k) \mid X_0 = i, X_\alpha = j \right\} , \quad (20) \]

where \( \mathcal{X} = \{ X_n, n \geq 0 \} \) and \( \mathcal{X'} = \{ X'_n, n \geq 0 \} \) are two independent Markov chains with the same transition matrix \( P \) and different initial states \( X_0 = i \) and \( X'_0 = j \), respectively, and \( N_{ij} = \min\{ n : X_n = X'_n \} \). The estimate based on (20) requires two sample paths, one starting from \( i \) and the other \( j \), but it requires only a finite number of transitions on each path. This approach for estimating performance sensitivity is in fact an extension of the method described in Dai and Ho (1995) to the discounted case (0 < \( \alpha \) < 1).

\[ \pi \] in (18) can be easily estimated on a sample path. For (17), we denote the \( i \)-th row of \( (1 - \alpha)$P$)^{-1} as \( \pi_{\alpha,i} \).

From \( (1 - \alpha)$P$)^{-1} = I + \alpha P + \alpha^2 P^2 + \cdots \), the \( k \)-th component of \( \pi_{\alpha,i} \)

\[ \pi_{\alpha,i}(k) = (1 - \alpha)E \left\{ \sum_{l=0}^{\infty} \alpha^l I(X_k) \mid X_0 = i \right\} , \]

where \( I(x) = 1 \), if \( x = k \), and \( I(x) = 0 \), otherwise, which can also be estimated on a sample path. With this notation, (17) becomes

\[ \frac{d\pi_{\alpha,i}}{dq} = \pi_{\alpha,i}[2(q(I - \alpha P + \alpha q)^{-1})^r f + h] \]

\[ = \pi_{\alpha,i}[2qD\pi_{\alpha} + h] \]

\[ = \pi_{\alpha,i}[2q\pi_{\alpha} + h] , \quad 0 < \alpha < 1. \]

We have for all \( i \),

\[ \lim_{\alpha \to 1} \pi_{\alpha,i} = \pi. \]

Using (19) or (20), (17) and (18), we can implement sensitivity analysis and MDP by analyzing sample paths of Markov chains.

7. Conclusions

We have proposed a simple and unified approach for both the performance sensitivity analysis of Markov chain and Markov decision problems (MDP) with infinite horizon average-cost and discounted performance criteria. The approach is based on the concept of \( \alpha \)-potential, which can be estimated based on a single sample path.

References


