Available online at www.sciencedirect.com
ScienceDirect
Automatica 44 (2008) 1055–1061
automatica
www.elsevier.com/locate/automatica

Brief paper

Policy iteration based feedback control☆

Kan-Jian Zhang[a,1], Yan-Kai Xu[b,2], Xi Chen[b,2], Xi-Ren Cao[c,*,3]

[a] *Research Institute of Automation, Southeast University, Nanjing 210096, China*
[b] *CFINS, Department of Automation, Tsinghua University, Beijing 100084, China*
[c] *Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*

Received 7 March 2006; received in revised form 22 April 2007; accepted 22 August 2007
Available online 21 December 2007

Abstract

It is well known that stochastic control systems can be viewed as Markov decision processes (MDPs) with continuous state spaces. In this paper, we propose to apply the policy iteration approach in MDPs to the optimal control problem of stochastic systems. We first provide an optimality equation based on performance potentials and develop a policy iteration procedure. Then we apply policy iteration to the jump linear quadratic problem and obtain the coupled Riccati equations for their optimal solutions. The approach is applicable to linear as well as nonlinear systems and can be implemented on-line on real world systems without identifying all the system structure and parameters.

Keywords: Nonlinear control; Performance potentials; Jump linear system; Coupled Riccati equation

1. Introduction

Optimal control of stochastic dynamic systems is a difficult problem. Because exact solutions can only be obtained under some rather strict restrictions on system structures, numerical and approximate approaches have to be developed. Dynamic programming (Bellman, 1957) is one of the commonly used approaches to the problem, which solves, analytically or numerically, the well-known optimality equation called the Bellman equation.

The research in this paper was motivated by the results in two areas: optimal control of stochastic dynamic systems and Markov decision processes (MDPs). It has been realized for a long time that stochastic control systems can be viewed as Markov decision processes. For example, numerical algorithms based on value iterations (Kushner & Paul, 1992; Puterman, 1994; Tsitsiklis & Van Roy, 1996) are developed for solving the Bellman equation (Kushner, 1977). Hernandez-Lerma and Lasserre (1996) deal with general state MDPs and provide conditions for existence of stationary optimal policies. However, with value iteration and dynamic programming, the transition probabilities, or equivalently the system structure and parameters, have to be known. When the system structure and/or parameters are unknown, identification methods have to be used, and this further complicates the problem. Therefore, numerical methods and learning based approaches have to be developed.

In this paper, we consider average cost Markov decision problems. We propose a policy iteration based approach for the optimal control problem. At each iteration, we analyze the system's behavior under one policy and find another policy under which the system performs better. The main concept of this approach is the performance potential (Cao, 2000, 2003) (or the bias Puterman, 1994, or relative cost Bertsekas, 1995). When the system structure and parameters are known, the potentials can be obtained by solving the Poisson equation; otherwise, they can be estimated from a sample path (Cao & Wan, 1998). Compared with value iteration and dynamic programming, the policy iteration based approach can be implemented on-line

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Bart De Schutter under the direction of Editor Ian Petersen.

* Corresponding author. Tel.: +852 2358 7048; fax: +852 2358 1485.
E-mail address: eecao@ee.ust.hk (X.-R. Cao).

[1] Partially supported by the National Natural Science Foundation (60404006) of China.
[2] Partially supported by the National Natural Science Foundation (60574064) of China.
[3] Supported in part by a grant from Hong Kong UGC.

0005-1098/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2007.08.014

without knowing all the system parameters, and learning based implementation algorithms can be developed. The approach treats nonlinear systems in the same way as the linear systems.

In Section 2, we describe how an (nonlinear) optimal control problem can be modelled as a Markov decision problem. In Section 3, we develop the policy iteration theory for MDPs with continuous state spaces. In Section 4, we apply the policy iteration approach to the jump linear quadratic (JLQ) problem. We obtain the closed form of the potentials for the problem and show that the optimal solution can be obtained via the coupled Riccati equations. Section 5 concludes the paper with a discussion.

The contributions of the paper is to define the performance potential for continuous state space, and propose the policy iteration approach to solve the optimal control problems and apply it to the JLQ problem. The policy iteration based approach can be implemented on-line and learning algorithms can be developed.

2. Control systems modelled as MDPs

Consider a stochastic control system of the form

$$X_{l+1} = H(X_l, u_l) + \xi_l, \quad l = 0, 1, \ldots, \tag{1}$$

where $l = 0, 1, \ldots,$ denotes the discrete time, $X_l \in \mathscr{R}^n$, $\mathscr{R} = (-\infty, +\infty)$, is an n-dimensional vector representing the system state at time l, $\xi_l \in \mathscr{R}^n$ is the random noise at time l, and $u_l \in U$ is an m-dimensional vector representing the control applied to the system at time l, with U being a specified control constraint set of $\mathscr{R}^m$. We assume that ξ_l, $l = 0, 1, \ldots,$ is a sequence of independent and identically distributed (i.i.d.) random variables with a distribution density function $p_\xi(y)$, $y \in \mathscr{R}^n$ whose mean is zero and variance is finite.

A feedback control law is an m-dimensional function of the state denoted as $u_l(X_l)$. If it is independent of l, the control law is called *stationary* and is denoted as $u(x)$. In optimal control, we consider only the control laws that make the system stable.

Suppose that at time l, $X_l = x$. Then we have $X_{l+1} = H(x, u(x)) + \xi_l$. Thus, $X_{l+1} - H(x, u(x))$ has a distribution function $p_\xi(y)$. Therefore, the transition at time l from state $X_l = x$ to $X_{l+1} \in [y, y + \mathrm{d}y)$ has the following transition probability function:

$$P^u(\mathrm{d}y|x) = p_\xi[y - H(x, u(x))]\,\mathrm{d}y.$$

The superscript u indicates the dependency on the control variable. Clearly, $u(x)$ plays the same role as the actions do in MDPs, and the control function u is the same as a policy. The performance measure to be minimized is defined as

$$\eta^u(x) = \lim_{L\to\infty} \frac{1}{L} E\left\{\sum_{l=0}^{L-1} f(X_l, u_l)|X_0 = x\right\}, \tag{2}$$

where $f(x, u)$ is a cost function which is assumed to be a measurable function. In fact, the limit in (2) may not exist always. However, in many engineering problems, such limit usually exists. Therefore, a stochastic control system (1) with performance measure (2) can be modelled as an MDP problem with performance measure (2). The state spaces are usually continuous, and the relevant results will be discussed in Section 3.

3. MDPs with continuous state spaces

As shown in Section 2, we need to extend the theory of policy iteration to continuous state spaces. The transition probability with a continuous state space is described by an operator (integration) on the function space. We will present the main ideas and will not go deep into the operator theory; especially, we will not study the general conditions for the infinite dimensional operators to be interchangeable in their orders. There are standard theorems for such interchangeability (e.g., see Durrett, 1996, Section 1.3).

3.1. Transition probability functions and steady-state probability

Consider a discrete-time Markov chain $\mathbf{X} := \{X_0, X_1, \ldots\}$ with a continuous state space $\mathscr{S} = \mathscr{R}^n$. Let $\mathscr{B}$ be the σ-field of $\mathscr{R}^n$ containing all the (Lebesgue) measurable sets. Given the current state $x \in \mathscr{R}^n$, the probability that the next state lies in a set $B \in \mathscr{B}$ can be denoted as a state transition function $P(B|x)$ with $P(\mathscr{R}^n|x) = \int_{\mathscr{R}^n} P(\mathrm{d}y|x) = 1$ for all $x \in \mathscr{R}^n$. Without specifically mentioning, we will assume that all sets and functions discussed in this paper are (Lebesgue) measurable. Define a linear (right) operator, $\mathbf{P}$, corresponding to $P(B|x)$ on the function space as follows:

$$\mathbf{P}h(x) := \int_{R^n} h(y) P(\mathrm{d}y|x), \tag{3}$$

where $h(x)$ is any measurable function.

For any two operators $\mathbf{P}_1$ and $\mathbf{P}_2$, their product is defined as

$$(\mathbf{P}_1\mathbf{P}_2)(B|x) = \int_{\mathscr{R}^n} P_2(B|y) P_1(\mathrm{d}y|x), \quad x \in \mathscr{S},\ B \in \mathscr{B}. \tag{4}$$

Define function $e(x) = 1$ for all $x \in \mathscr{R}^n$ and an identity operator, I: $I(B|x) = 1$ if $x \in B$; $I(B|x) = 0$ otherwise. Then $Pe = e$ for any transition function P, and $(Ih)(x) = h(x)$, $x \in \mathscr{R}^n$, for any function h.

Suppose that Markov chain $\mathbf{X}$ is time-homogeneous. Then the l-step transition probability functions, denoted as $P^l(B|x)$, are defined as $P^0(B|x) = I(B|x)$ and

$$P^l(B|x) = \int_{\mathscr{R}^n} P(\mathrm{d}y|x) P^{l-1}(B|y), \quad l \geqslant 1, \tag{5}$$

and by (4) $\mathbf{P}^l$ also denotes the operator corresponding to the l-step transition function $P^l(B|x)$, in particular, $\mathbf{P}^0 := I$.

A probability measure $v(B)$ itself can be viewed as a special state transition function $v(B|x)$ which takes the same value $v(B)$ for all $x \in \mathscr{R}^n$. Thus, any probability measure $v(B)$ can be viewed as an operator v.

Let $f(x)$ be a cost function. The long-run average performance is defined as

$$\eta(x) = \lim_{L\to\infty} \frac{1}{L} E\left\{\sum_{l=0}^{L-1} f(X_l)|X_0 = x\right\}. \tag{6}$$

By (3), we have $\mathbf{P}f(x) = \int_{\mathscr{R}^n} f(y)P(\mathrm{d}y|x) = E[f(X_1)|X_0 = x]$, and because $P^l(B|x), l = 1, 2, \ldots,$ is the l-step transition function and $\mathbf{P}^l$ is the operator corresponding to $P^l(B|x)$, we have $\mathbf{P}^l f(x) = \int_{\mathscr{R}^n} f(y)P^l(\mathrm{d}y|x) = E[f(X_l)|X_0 = x]$. Thus, it is clear that

$$\eta(x) = \lim_{L\to\infty} \frac{1}{L} \left\{\sum_{l=0}^{L-1} (\mathbf{P}^l f)(x)\right\}. \tag{7}$$

The steady-state probability distribution of a transition function $P(B|x)$, $B \in \mathscr{B}$ and $x \in \mathscr{R}^n$, is defined as a probability distribution π satisfying $\pi = \pi\mathbf{P}$. Normally, we hope that as $l \to \infty$, P^l will converge to the transition function π. However, with a continuous state space, there are many ways to define the convergence of P^l, $l = 1, 2, \ldots$. The convergence is related to the topic of ergodicity. For the analysis in this paper, we only need to assume that for the performance function $f(x)$, we have

$$\lim_{l\to\infty} (\mathbf{P}^l f)(x) = (\pi f)e(x), \quad \forall x \in \mathscr{R}^n. \tag{8}$$

Therefore, from (7), we have

$$\eta(x) = \lim_{l\to\infty} \mathbf{P}^l f(x) = (\pi f)e(x), \tag{9}$$

with $\pi f = \int_{\mathscr{R}^n} f(x)\pi(\mathrm{d}x)$.

With a slightly abused notation, we also use $\eta := \pi f$ as a constant. Thus, we have $\eta(x) = \eta e(x)$.

3.2. Potentials and policy iteration

3.2.1. Performance potentials

Suppose that a Markov chain $\{X_l, l = 0, 1, \ldots\}$ with continuous state space on $\mathscr{R}^n$ has a steady-state probability π. The performance potential g is a function that satisfies the Poisson equation

$$(I - \mathbf{P})g(x) + \eta(x) = f(x). \tag{10}$$

Notice that if g is a solution to (10), so is $g + ce$ with any constant c. In fact, the performance potential is nothing but the differential cost. We shall use the same notation for these different versions of potentials with only a constant difference.

With a continuous state space, we have to be careful on exchanging the order of mathematical operations such as integration and limit. We define

$$g_L := \left\{I + \sum_{l=1}^{L} (\mathbf{P}^l - \boldsymbol{\pi})\right\} f. \tag{11}$$

Set $g := \lim_{L\to\infty} g_L$, assuming the limit exists. We know that $\mathbf{P}g_L(x) = \int_{\mathscr{R}^n} g_L(y)P(\mathrm{d}y|x)$. Therefore, if this integration uniformly converges for all L, we can exchange the order of the "*lim*" and the integration "$\int$" and obtain $\lim_{L\to\infty} \mathbf{P}g_L = \mathbf{P}g$. All the property are related to the topic of ergodicity and we will not go deep.

Lemma 1. *For any transition function P and performance function $f(x)$, if*

$$\lim_{l\to\infty} (\mathbf{P}^l f) = (\pi f)e,$$

$$\lim_{L\to\infty} g_L = g \quad \textit{and} \quad \lim_{L\to\infty} \mathbf{P}g_L = \mathbf{P}g \tag{12}$$

hold for every $x \in \mathscr{R}^n$, then

$$g = \left\{I + \sum_{l=1}^{\infty} (\mathbf{P}^l - \boldsymbol{\pi})\right\} f \tag{13}$$

is a solution of (10).

This lemma can be established by directly verifying that g in (13) satisfies (10). In addition, we can easily verify that

$$\pi g = \pi f = \eta. \tag{14}$$

This is a normalizing condition of the potential g in (13). With (14), the Poisson equation (10) becomes $(I - \mathbf{P} + \pi)g(x) = f(x)$.

3.2.2. Performance optimization

First, we modify the definition of the relations $=$, $\leqslant$, $<$, and $\preccurlyeq$ for two functions in $\mathscr{R}^n$. Given a probability measure v on $\mathscr{R}^n$, for two functions $h(x)$ and $h'(x)$, $x \in \mathscr{R}^n$, we define $h' =_v h$, $h' \leqslant_v h$, and $h' <_v h$, respectively, if $h'(x) = h(x)$, $h'(x) \leqslant h(x)$, and $h'(x) < h(x)$, respectively, for all $x \in \mathscr{R}^n$ except on a set E with $v(E) = 0$. We further define $h'(x) \preccurlyeq_v h(x)$ if $h'(x) \leqslant_v h(x)$ and $h'(x) < h(x)$ on a set E with $v(E) > 0$. Similar definitions are used for the relations $>_v$, $\succcurlyeq_v$, and $\geqslant_v$. With these definitions, we have the following results.

Let (P, f) and (P', f') be the transition functions and performance functions of two Markov chains with the same state space $\mathscr{S} = \mathscr{R}^n$. Let η, g, π and η', g', and π' be their corresponding long-run average performances, performance potential functions, and steady-state probability measures, respectively. Then average performance difference formula is

$$\eta' - \eta = \pi'[(f' + \mathbf{P}'g) - (f + \mathbf{P}g)]. \tag{15}$$

And the Comparison Lemma is

$$\text{If } f' + \mathbf{P}'g \preccurlyeq_{\pi'} f + \mathbf{P}g \text{ then } \eta' < \eta. \tag{16}$$

To develop the optimality equation, we need to further restrict the policy space. First, we use u to denote a policy. In general, $u(x), x \in \mathscr{R}^n$, represents an action that determines the transition function $P(B|x)$, $B \in \mathscr{B}$. Thus, policy u determines the steady-state probability measure.

Two policies u and u' are said to have the same *support* if for any set B in $\mathscr{R}^n$, if $\pi^u(B) > 0$ then $\pi^{u'}(B) > 0$ and vice versa. If the Markov chain is irreducible for any policy then all the policies have the same support. Now we assume that all the policies in the policy space have the same support. So we can

drop the subscript π^u in the relationship notations such as $\leqslant$ and $\preccurlyeq$, etc. Then from (15) we have the optimality condition:

A policy $\hat{u}$ is optimal, if and only if

$$f^{\hat{u}} + \mathbf{P}^{\hat{u}} g^{\hat{u}} \leqslant f^u + \mathbf{P}^u g^{\hat{u}} \quad \text{for all policies } u. \tag{17}$$

From condition (17), the optimality equation is

$$\mathbf{P}^{\hat{u}} g^{\hat{u}} + f^{\hat{u}} = \min_u \{\mathbf{P}^u g^{\hat{u}} + f^u\}. \tag{18}$$

This equation holds with probability one with respect to the steady-state probability measure of any policy. We also assume that the policy space is, in a sense, compact and has some sort of continuity and hence the minimum can be reached.

With the Comparison Lemma (16), policy iteration algorithms can be designed. Roughly speaking, we may start with any policy u_0. At the kth step with policy u_k, $k = 0, 1, \ldots$, we set

$$u_{k+1}(x) = \arg\left\{\min_u [f^u(x) + \mathbf{P}^u g^{u_k}(x)]\right\}, \tag{19}$$

with g^{u_k} being any solution to the Poisson equation (10) for (P^{u_k}, f^{u_k}). If at some x, $u_k(x)$ attains the minimum, we set $u_{k+1}(x) = u_k(x)$. The iteration stops if u_{k+1} and u_k differ only on a set with zero measure. In this case, $\eta^{u_{k+1}} = \eta^{u_k}$. Comparison Lemma (16) implies that performance improves at each step. Optimality condition (17) shows that the minimum is reached when performance can no longer be improved. If the policy space is finite, the policy iteration will stop in a finite number of steps. Otherwise, the iteration scheme may not stop at a finite number of steps, although the sequence of the performance η^{u_k} is decreasing. In practical applications, PI always performs well with a fast convergence speed. In on-line algorithm, to evaluate a policy, we estimate potentials by averaging samples along the sample path. Thus computation of policy evaluation is $O(n)$, where n is the length of a sample path.

3.3. Comparison: dynamic programming and policy iteration

As we know, the standard approach to the control problem (1) is based on dynamic programming. Thus its procedure goes *backward* in time, and an optimal policy is obtained at each iteration for a finite-step problem. The long-run average problem is treated as the limit of the finite-step problem when the number of steps goes to infinity.

In contrast, with the policy iteration approach, at each iteration we deal with a (stationary) policy (not necessary an optimal one) with an infinite horizon. First, we work *forward* in time to obtain the performance potentials $g(x)$, $x \in \mathscr{R}^n$, of the policy. Then we find a better policy. In this way we iterate in the policy space to reach an optimal policy.

There is another difference between the two approaches. The dynamic programming approach requires to know the transition probabilities for all actions. Because working backward is somehow unrealistic in practice, it cannot be implemented on a real system. On the other hand, policy iteration can be implemented on-line, the potentials can be learned on a sample path, and in many cases this approach does not require to know all the transition probabilities.

4. JLQ problem

In this section, we derive the performance potentials for the JLQ problem. We show that with the approach developed in Section 3, we can directly obtain the optimal feedback and coupled Riccati equation, which are usually obtained by dynamic programming.

In a discrete time JLQ problem, we consider a two-level stochastic control system. The system state at time l, $l = 0, 1, \ldots$, is denoted as (M_l, X_l), where $M_l \in \mathscr{M} := \{1, 2, \ldots, M\}$ represents the *mode* (high level) that the system is in, and $X_l \in \mathscr{R}^n$ denote the continuous part of the *state* (low level). The system changes its mode as an ergodic Markov chain with transition probabilities $p_{\text{jump}}(j|i)$, $j, i \in \mathscr{M}$. When the system is in mode $M_l = i$, $l = 0, 1, \ldots$, the continuous part X_l, evolves as

$$X_{l+1} = A_{M_l} X_l + B_{M_l} u_l + \sigma_{M_l} \xi_l, \quad M_l = i \in \mathscr{M}. \tag{20}$$

The dimensions of A_i and B_i are $n \times n$ and $n \times m$, respectively. σ_i is $n \times n$ matrix, ξ_l is an i.i.d. random variable with zero mean and covariance I (identity matrix). ξ_l is not necessary to be Gaussian distributed. Denote the density function of $\sigma_i \xi$ as $p_{\xi_i}(x)$. The transitions among the modes and among the continuous states are assumed to be independent. The performance criterion is

$$\eta^u(i, x) = \lim_{L\to\infty} \frac{1}{L} E\left\{\sum_{l=0}^{L-1} [X_l^{\mathrm{T}} Q_{M_l} X_l + u_l^{\mathrm{T}} V_{M_l} u_l]| \right.$$
$$\left. X_0 = x, M_0 = i\right\}, \tag{21}$$

where Q_i is an $n \times n$ positive semi-definite matrix and V_i is an $m \times m$ positive definite matrix for all $i \in \{1, \ldots, M\}$, respectively. Our goal is to find a control law $u = u(i, x)$ that minimizes $\eta^u(i, x)$.

4.1. Transition operator

Denoting the transition function of JLQ problem as $P(j, B|i, x)$, we have

$$P(j, B|i, x) = p_{\text{jump}}(j|i) P_i(B|x),$$
$$i, j \in \mathscr{M}, \ x \in \mathscr{R}^n, \ B \in \mathscr{B}, \tag{22}$$

where $P_i(\mathrm{d}y|x) = p_{\xi_i}(y - [A_i x + B_i u(i, x)])\,\mathrm{d}y$, which depends on i. For any function $h(i, x)$, the transition operator $\mathbf{P}$ corresponding to $P(j, B|i, x)$ is defined as

$$(\mathbf{P}h)(i, x) = \sum_{j\in\mathscr{M}} \left\{ p_{\text{jump}}(j|i) \int_{\mathscr{R}^n} h(j, y) P_i(\mathrm{d}y|x)\right\}.$$

$\mathbf{P}^l$ denoted the operator corresponding to transition function $P^l(j, B|i, x)$, and $\mathbf{P}^0 = I$, with $I(j, B|i, x) = 1$ if $i = j$ and $x \in B$; 0 otherwise. A probability measure on $\mathscr{M} \times \mathscr{R}^n$ is denoted as $v(i, B)$ which can be viewed as a special state transition function $v(i, B|j, x)$ with the same value $v(i, B)$ for all $j \in \mathscr{M}$ and $x \in \mathscr{R}^n$.

π, as a steady-state probability measure of P, satisfies $\pi = \pi\mathbf{P}$. Similar to (8), we assume that for the performance function $f(i,x), i \in \mathscr{M}, x \in \mathscr{R}^n$

$$\lim_{l\to\infty} \mathbf{P}^l f = \pi f \tag{23}$$

holds. For a stationary feedback control law $u = u(i,x)$, from (21) the cost function for the JLQ problem is

$$f(i,x) = x^{\mathrm{T}} Q_i x + u(i,x)^{\mathrm{T}} V_i u(i,x). \tag{24}$$

4.2. Performance potentials

The long-run average performance of the Markov chain with transition function $P(j, B|i, x)$ is

$$\eta(i,x) = \lim_{L\to\infty} \frac{1}{L} E\left\{\sum_{l=0}^{L-1} f(M_l, X_l)|M_0 = i, X_0 = x\right\}.$$

We have $\eta(i,x) = \lim_{L\to\infty} (1/L)\sum_{l=0}^{L-1} (\mathbf{P}^l f)(i,x)$. From (23), we have $\eta(i,x) = \lim_{l\to\infty} (\mathbf{P}^l f)(i,x) = (\pi f)e(i,x)$. The performance potential g satisfies the Poisson equation $(I - \mathbf{P})g + \eta = f$.

For zero-mean distributions, we have $\int_{\mathscr{R}^n} p_{\xi_i}(z)\,\mathrm{d}z = 1$ and $\int_{\mathscr{R}^n} z p_{\xi_i}(z)\,\mathrm{d}z = 0$. For any quadratic function $h(i,x) = x^{\mathrm{T}} W_i x$, where W_i, $i = 1, 2, \ldots, M$, are positive semi-definite matrices, and a control law $u(i,x)$, we have

$$\begin{aligned}
(\mathbf{P}^u h)(i,x) &= \sum_{j\in\mathscr{M}} \left\{ p_{\mathrm{jump}}(j|i) \int_{\mathscr{R}^n} h(j,y) P_i^u(\mathrm{d}y|x) \right\} \\
&= \sum_{j\in\mathscr{M}} \left\{ p_{\mathrm{jump}}(j|i) \int_{\mathscr{R}^n} \{z + [A_i x + B_i u(i,x)]\}^{\mathrm{T}} W_j \right. \\
&\qquad \left. \times \{z + [A_i x + B_i u(i,x)]\} p_{\xi_i}(z)\,\mathrm{d}z \right\} \\
&= \tilde{c}(i)e(x) + \sum_{j\in\mathscr{M}} p_{\mathrm{jump}}(j|i) \\
&\qquad \times [A_i x + B_i u(i,x)]^{\mathrm{T}} W_j [A_i x + B_i u(i,x)], \qquad (25)
\end{aligned}$$

where $\tilde{c}(i) := \sum_{j\in\mathscr{M}} p_{\mathrm{jump}}(j|i) \int_{\mathscr{R}^n} [z^{\mathrm{T}} W_j z] p_{\xi_i}(z)\,\mathrm{d}z$.

For a linear control $u(i,x) = -D_i x$, the system equation (20) becomes $X_{l+1} = C_{M_l} X_l + \sigma_{M_l} \xi_l$, where $C_i = A_i - B_i D_i$, $i \in \mathscr{M}$. The cost function (24) becomes $f(i,x) = x^{\mathrm{T}} W_i x$ with

$$W_i = Q_i + D_i^{\mathrm{T}} V_i D_i, \quad i \in \mathscr{M}. \tag{26}$$

Set $W_{i,0} := W_i$. Then for this jump linear system, we have

$$(\mathbf{P} f)(i,x) = c_1(i)e(i,x) + x^{\mathrm{T}} W_{i,1} x, \tag{27}$$

where

$$\begin{cases}
W_{i,1} = \sum_{j\in\mathscr{M}} p_{\mathrm{jump}}(j|i)(C_i^{\mathrm{T}} W_{j,0} C_i), \\
c_1(i) := c_{1,0}(i), \\
c_{1,0}(i) := \sum_{j\in\mathscr{M}} p_{\mathrm{jump}}(j|i) \int_{\mathscr{R}^n} z^{\mathrm{T}} W_{j,0} z p_{\xi_i}(z)\,\mathrm{d}z.
\end{cases}$$

Next, we have

$$(\mathbf{P}^2 f)(i,x) = c_2(i)e(i,x) + x^{\mathrm{T}} W_{i,2} x, \tag{28}$$

where

$$\begin{cases}
W_{i,2} = \sum_{j\in\mathscr{M}} p_{\mathrm{jump}}(j|i)(C_i^{\mathrm{T}} W_{j,1} C_i), \\
c_2(i) := c_{2,0}(i) + c_{1,1}(i), \\
c_{0,1}(i,j) := \int_{\mathscr{R}^n} z^{\mathrm{T}} W_{j,1} z p_{\xi_i}(z)\,\mathrm{d}z, \\
c_{2,0}(i) := \sum_{j\in\mathscr{M}} c_{1,0}(j) p_{\mathrm{jump}}(j|i), \\
c_{1,1}(i) := \sum_{j\in\mathscr{M}} c_{0,1}(i,j) p_{\mathrm{jump}}(j|i).
\end{cases}$$

Continuing this process, we have

$$(\mathbf{P}^l f)(i,x) = c_l(i)e(i,x) + x^{\mathrm{T}} W_{i,l} x, \tag{29}$$

where

$$\begin{cases}
W_{i,l} = \sum_{j\in\mathscr{M}} p_{\mathrm{jump}}(j|i)(C_i^{\mathrm{T}} W_{j,l-1} C_i), \\
c_l(i) := c_{l,0}(i) + c_{l-1,1}(i) + \cdots + c_{1,l-1}(i), \\
c_{0,n}(i,j) := \int_{\mathscr{R}^n} [z^{\mathrm{T}} W_{j,n} z] p_{\xi_i}(z)\,\mathrm{d}z, \\
c_{1,n}(i) := \sum_{j\in\mathscr{M}} c_{0,n}(i,j) p_{\mathrm{jump}}(j|i), \\
c_{m,n}(i) := \sum_{j\in\mathscr{M}} c_{m-1,n}(j) p_{\mathrm{jump}}(j|i), \quad \forall n < l, 1 < m \leqslant l.
\end{cases}$$

Under the stochastic stabilizability condition (Costa, Fragoso, & Marques, 1995, 2005), we have $W_{i,l} \to 0$ as $l \to \infty$. Therefore, from (29) and (23), we have $c_l(i) \to \eta$, for all $i \in \mathscr{M}$. Moreover, the sum $S_i := \sum_{l=0}^{\infty} W_{i,l}$ exists and, moreover, with $g_L(i,x) = [\sum_{l=1}^{L} (c_l(i) - \eta)]e(i,x)$, conditions in Lemma 1 hold. Therefore, we have

$$g(i,x) = \left[\sum_{l=1}^{\infty} (c_l(i) - \eta)\right] e(i,x) + x^{\mathrm{T}} S_i x. \tag{30}$$

The first term depends on the first state component i and therefore is not a constant. In fact, $x^{\mathrm{T}} S_i x$ corresponds to the potentials of the states in the same mode, while $\sum_{l=1}^{\infty} (c_l(i) - \eta)$ reflects the difference of the potentials at the states in different modes.

Let $F_i := \sum_{j\in\mathscr{M}} S_j p_{\mathrm{jump}}(j|i)$. We have

$$S_i = C_i^{\mathrm{T}} F_i C_i + W_i, \quad i \in \mathscr{M}. \tag{31}$$

4.3. Optimal policy

Before we apply optimality condition (17), we need to verify that all policies have the same support. To see this, we first assume that the noise distribution $p_{\xi_i}(B)$ has a support of the whole space $\mathscr{R}^n$; i.e., $p_{\xi_i}(B) > 0$ for all $i \in \mathscr{M}$ and B with a nonzero Lebesgue measure. This is common, e.g., the Gaussian distribution has this property. Now, for any B with a nonzero Lebesgue measure, we have $\pi(i,B) = (\pi\mathbf{P})(i,B)$.

In the JLQ problem, since $E(\|X_l\|)$ is bounded under the stabilized control, by Chebyshev's inequality, for

any $1>\varepsilon>0$ there exists a constant $M>0$ such that $\lim\inf_{l\to\infty} P^l(.,\bar{B}|i,x)\geqslant 1-\varepsilon$ for $\bar{B}=\{\|x\|\leqslant M\}$. For any $i,j\in\mathscr{M}$, $x\in\mathscr{R}^n$ we have $P(j,B|i,x)=p_{\text{jump}}(j|i)\int_{y\in B} p_{\xi_i}(y-C_ix)\,\mathrm{d}y$, where C_i depends on the control u. Because $p_{\xi_i}(y-C_ix)>0$, it can be obtained that $P(j,B|i,x)>0$. This lead to $\min_{(i,x)\in\mathscr{M}\times\bar{B}} P(j,B|i,x):=\delta>0$ since $\bar{B}$ is tight and $\mathscr{M}$ is finite. Thus, we have $P^{l+1}(j,B|i,x)\geqslant P^l(.,\bar{B}|i,x)\min_{k\in\mathscr{M}} P(j,B|k,\bar{B})\geqslant(1-\varepsilon)\min_{k\in\mathscr{M}}\int_{\bar{B}} P(j,B|k,x)\,\mathrm{d}x\geqslant(1-\varepsilon)\delta M>0$. This imply that $\pi(i,B)>0$. That is, all policies have the same support of $\mathscr{R}^n$. Of course, the condition that $p_{\xi_i}(x)>0$ for any $x\in\mathscr{R}^n$ is not a necessary condition.

We may apply the optimality equation (18) to determine an optimal policy. We start with a linear policy $u(i,x)=-D_ix$. The performance potential is (30). From (25), for any other policy $u'(i,x)$, which may not be linear, we have

$$\begin{aligned}&(\mathbf{P}^{u'}g)(i,x)+f^{u'}(i,x)\\&\quad=c(i)e(x)+x^{\mathrm{T}}(A_i^{\mathrm{T}}F_iA_i+Q_i)x\\&\qquad+u'(i,x)^{\mathrm{T}}(B_i^{\mathrm{T}}F_iB_i+V_i)u'(i,x)\\&\qquad+x^{\mathrm{T}}A_i^{\mathrm{T}}F_iB_iu'(i,x)+[x^{\mathrm{T}}A_i^{\mathrm{T}}F_iB_iu'(i,x)]^{\mathrm{T}}\end{aligned}$$

in which $c(i)$ is some constant and $c(i)e(i,x)+x^{\mathrm{T}}(A_i^{\mathrm{T}}F_iA_i+Q_i)x$ does not depend on u'. The improved policy is therefore

$$\tilde{u}(i,x)=\arg\min_{u'}\{(\mathbf{P}^{u'}g)(i,x)+f^{u'}(i,x)\}=-\tilde{D}_ix,$$

where $\tilde{D}_i=(B_iF_iB_i+V_i)^{-1}B_i^{\mathrm{T}}F_iA_i$.

Next, if $\tilde{D}_i=D_i$ then the policy $u(i,x)=-D_ix$ satisfies the optimality equation (18) and from (17) it is an optimal policy. Denote it as $\hat{u}(i,x)=-\hat{D}_ix$ and the corresponding quantity as $\hat{S}_i$ and $\hat{F}_i$, we have

$$\hat{D}_i=(B_i\hat{F}_iB_i+V_i)^{-1}B_i^{\mathrm{T}}\hat{F}_iA_i. \tag{32}$$

Substituting (32) into (31), we have

$$\begin{cases}\hat{S}_i=A_i^{\mathrm{T}}\hat{F}_iA_i-A_i^{\mathrm{T}}\hat{F}_iB_i(V_i+B_i^{\mathrm{T}}\hat{F}_iB_i)^{-1}B_i^{\mathrm{T}}\hat{F}_iA_i+Q_i,\\ \hat{F}_i=\sum_{j\in\mathscr{M}}\hat{S}_jp_{\text{jump}}(j|i).\end{cases} \tag{33}$$

This is the *coupled Riccati equation*.

We can easily understand the policy iteration procedure. At the kth iteration step, we have $u^{\{k\}}(i,x)=-D_i^{\{k\}}x$,

$$D_i^{\{k\}}=(B_i^{\mathrm{T}}S_i^{\{k-1\}}B_i+V_i)^{-1}B_i^{\mathrm{T}}F_i^{\{k-1\}}A,$$

and $C_i^{\{k\}}=A_i-B_iD_i^{\{k\}}$, $W_i^{\{k\}}=Q_i+D_i^{\{k\}^{\mathrm{T}}}V_iD_i^{\{k\}}$. On the other hand, we have

$$S_i^{\{k\}}=C_i^{\{k\}^{\mathrm{T}}}F_i^{\{k-1\}}C_i^{\{k\}}+W_i^{\{k\}}. \tag{34}$$

We can obtain $S_i^{\{k\}}$ from the above equation, then obtain next linear feedback control $D_i^{\{k+1\}}$. When this iterative procedure continues, the feedback coefficient $D_i^{\{k\}}$ converges to (32), which leads to the coupled Riccati equation (33). It should be pointed out that the existence and uniqueness of solution to coupled Riccati equation depend on conditions of stochastic stabilizability and stochastic detectability. We do not discuss them with details (see Costa et al., 1995, Lemma 1–3). The above policy iteration procedure can be viewed as a numerical approach for solving the coupled Riccati equation iteratively. This numerical approach is stable and can reach the optimum (Sutton & Batto, 1998).

To our best knowledge, former research (Costa et al., 2005) also obtained these results, as an extension of finite horizon case. We deal with infinite horizon case directly with our MDP approach and get the same results. Furthermore, we propose performance potentials in (30) for JLQ problem for the first time.

The well-known *Linear Quadratic* problem is a special case of the JLQ problem when there is only one mode. Its optimal control and the corresponding Riccati equation can be derived in a similar way.

5. Discussion and conclusion

In this paper, we apply the potential based policy iteration approach to MDPs with continuous state spaces to solve optimal control problems. We derive the potentials for JLQ problem and show that the solution to this problem can be obtained via coupled Riccati equations.

One of the main advantages of the policy iteration based approach is that it can be implemented on-line and learning based algorithms can be developed when the system structure and parameters are unknown. In addition, this approach can be applied to optimal control of nonlinear systems in the same way as linear systems.

The implementation details of the learning based policy iteration approach to the control problems of nonlinear systems will be discussed in a parallel paper.

Acknowledgments

The author would like to express their gratitude to the four anonymous reviewers for their comments in helping to revise this paper.

References

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control* (Vol. II). Belmont, MA: Athena Scientific.

Cao, X. R. (2000). A unified approach to Markov decision problems and performance sensitivity analysis. *Automatica*, *36*(5), 771–774.

Cao, X. R. (2003). From perturbation analysis to Markov decision processes and reinforcement learning. *Discrete Event Dynamic Systems: Theory and Applications*, *13*(1), 9–39.

Cao, X. R., & Wan, Y. W. (1998). Algorithms for sensitivity analysis of Markov systems through potentials and perturbation realization. *IEEE Transactions on Control Systems Technology*, *6*(4), 482–494.

Costa, O. L. V., Fragoso, M. D., & Marques, R. P. (1995). Discrete-time LQ-optimal control problems for infinite Markov jump parameter systems. *IEEE Transactions on Automatic Control*, *40*(12), 2076–2088.

Costa, O. L. V., Fragoso, M. D., & Marques, R. P. (2005). *Discrete-time Markov jump linear systems*. New York: Springer.

Durrett, R. (1996). *Probability: Theory and examples*. (2nd ed.), USA: Duxbury Press.

Hernandez-Lerma, O., & Lasserre, J. B. (1996). *Discrete-time Markov control processes: Basic optimality criteria*. New York: Springer.

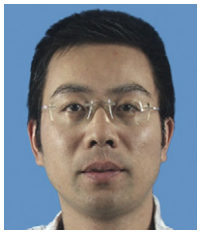
Kushner, H. J. (1977). *Probability methods for approximations in stochastic control and elliptic equations. Mathematics in science and engineering* (Vol. 129). New York: Academic Press.

Kushner, H. J., & Paul, G. (1992). *Numerical methods for stochastic control problems in continuous time*. New York: Springer.

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

Sutton, R. S., & Batto, A. G. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Tsitsiklis, J. N., & Van Roy, B. (1996). Feature-based methods for large scale dynamic programming. *Machine Learning*, *22*(1–3), 59–94.



Kan-Jian Zhang received the B.S. degree in mathematics from Nankai University, China in 1994, and the M.S. and Ph.D. degrees in control theory and control engineering from Southeast University, China in 1997 and 2000. He is currently an associate professor in Research Institute of Automation, Southeast University. His research is in nonlinear control theory and its applications, with particular interest in robust output feedback design and optimization control.

Yan-Kai Xu received his B.E. degree in automatic control in 2003 from Tsinghua University, Beijing, China. He is currently a Ph.D. candidate in the Center for Intelligent and Networked Systems (CFINS), Department of Automation, Tsinghua University. His research interests include optimization and control of stochastic systems, discrete event dynamic systems, and machine learning.

Xi Chen received her B.Sc. and M.Eng. from Nankai University, Tianjin, China, in 1986 and 1989, respectively. After graduation, she worked in the Software Engineering Institute at Beijing University of Aeronautics and Astronautics for seven years. From October 1996 she studied in the Chinese University of Hong Kong and received her Ph.D. in 2000. Then she worked as a post-doctoral fellow in Information Communication Institute of Singapore and in the Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong. Since July 2003, she works in the Center for Intelligent and Networked Systems (CFINS), Department of Automation, Tsinghua University, Beijing, China. Her research interests include wireless sensor networks and stochastic control.

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

Dr. Cao owns three patents in data and tele-communications and published two books: Realization Probabilities—the Dynamics of Queuing Systems, Springer Verlag, 1994, and Perturbation Analysis of Discrete-Event Dynamic Systems, Kluwer Academic Publishers, 1991 (co-authored with Y.C. Ho). He received the Outstanding Transactions Paper Award from the IEEE Control System Society in 1987 and the Outstanding Publication Award from the Institution of Management Science in 1990. He is a fellow of IEEE, Chairman of the IEEE Fellow Evaluation Committee of the IEEE Control Systems Society, Editor-in-Chief of Discrete Event Dynamic Systems: Theory and Applications, Associate Editor at Large of IEEE Transactions of Automatic Control, member of the Board of Governors of IEEE Control Systems Society, member of IFAC Technical Board, and chairman of IFAC Coordinating Committee on Systems and Signals. He has been served as associate editor of a number of international journals and chairman of a few technical committees of international professional societies. His current research areas include discrete event dynamic systems, stochastic learning and optimization theory, performance analysis of communication systems, and signal processing.