SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2024, Vol. 67, Iss. 2, 122204:1–122204:11
https://doi.org/10.1007/s11432-023-3820-3

Value iteration algorithm for continuous-time linear quadratic stochastic optimal control problems

Guangchen WANG & Heng ZHANG*

School of Control Science and Engineering, Shandong University, Jinan 250061*, China*

Received 20 February 2023/Revised 26 April 2023/Accepted 8 June 2023/Published online 25 January 2024

Abstract In this study, we investigate a continuous-time infinite-horizon linear quadratic stochastic optimal control problem with multiplicative noise in control and state variables. Using the techniques of stochastic stability, exact observability, and stochastic approximation, a value iteration algorithm is developed to solve the corresponding generalized algebraic Riccati equation. Unlike the existing policy iteration algorithm, this algorithm does not rely on an initial stabilizing control. Further, this algorithm can also be used to compute policy evaluation steps that arise in the policy iteration algorithm. Herein, a simulation example is provided to validate the obtained results.

Keywords stochastic systems, optimal control, linear quadratic stochastic problem, generalized algebraic Riccati equation, value iteration algorithm

1 Introduction

The theory of optimal control is a major branch of modern control theory, which focuses on the methods for finding an optimal control to optimize the performance of a controlled system. Pioneered by Wonham [1], the linear quadratic stochastic optimal control (LQSOC) problem and its generalization are among the most significant topics of control problems, which frequently appear in applications such as finance and engineering [2–8]. Because of its significance, this paper considers an infinite-horizon LQSOC problem in the following manner:

$$\inf_{u(\cdot)} J(u(\cdot)) = \mathbb{E}\int_0^{+\infty} \big[u(s)^{\mathrm{T}} R u(s) + x(s)^{\mathrm{T}} Q x(s)\big]\mathrm{d}s$$

subject to

$$\begin{cases} \mathrm{d}x(s) = \big[Ax(s) + Bu(s)\big]\mathrm{d}s + \big[Cx(s) + Du(s)\big]\mathrm{d}w(s), \\ x(0) = \xi. \end{cases}$$

When solving the LQSOC problem, it is natural to encounter a generalized algebraic Riccati equation (GARE):

$$\begin{cases} PA + Q + A^{\mathrm{T}}P + C^{\mathrm{T}}PC - (C^{\mathrm{T}}PD + PB)(D^{\mathrm{T}}PD + R)^{-1}(D^{\mathrm{T}}PC + B^{\mathrm{T}}P) = 0, \\ D^{\mathrm{T}}PD + R > 0. \end{cases} \tag{1}$$

See [9–11] for more information. However, because GARE (1) is nonlinear, obtaining an analytical solution of (1) is difficult. Even with the simpler algebraic Riccati equation found in linear quadratic deterministic optimal control (LQDOC) problems (i.e., $C = D = 0$ in the system dynamics), it is still a difficult problem. Therefore, numerical methods should be used to approximate the solution of GARE (1).

An efficient methodology for obtaining a numerical solution of GARE (1) is Newton's method developed by Damm and Hinrichsen [12]. Using resolvent positive operators, the authors proposed a Newton's

* Corresponding author (email: zhangheng2828@mail.sdu.edu.cn)

method to solve the GARE appearing in multidimensional LQSOC problems. Later, Refs. [11, 13] independently proposed the policy iteration (PI) algorithm to approximate the unique stabilizing solution of GARE (1), which was later shown to be equivalent to the Newton's method in terms of matrices. Initiated by a stabilizing control, the PI algorithm approximates the stabilizing solution of GARE (1) by alternately implementing two procedures: policy evaluation and policy improvement. However, finding a stabilizing control necessitates the adoption of other techniques, such as solving a Lyapunov-type inequality [14] or a linear matrix inequality [9, 15], which may be computationally expensive for large-scale systems. The PI and value iteration (VI) are two well-known main types of reinforcement learning methods. Over the last two decades, the methods have been widely adopted to solve control problems. See [11, 16–21] for more information. One of the primary motivations for developing a continuous-time VI algorithm is to eliminate the need for an initial stabilizing control. However, to the authors' best knowledge, no VI algorithm that addresses the aforementioned LQSOC problem exists.

To address this gap, this study aims to abandon the assumption of an initial stabilizing control and develop a VI algorithm to address the LQSOC problem. In Bian and Jiang [19], a VI algorithm is developed to solve a continuous-time LQDOC problem. Noteworthily, their algorithm is based on the fact that a solution to a differential Riccati equation with any positive semidefinite terminal condition converges to the unique solution to the corresponding algebraic Riccati equation as time passes. Nonetheless, since GARE (1) has a more complex structure, obtaining a similar asymptotic behavior for GARE (1) and the corresponding generalized differential Riccati equation (GDRE) is difficult. To address this issue, herein, we employ the theory of exact observability and stochastic stabilizability. We show that the stabilizing solution to GARE (1) is a locally asymptotically stable equilibrium point for the GDRE under some stabilizability and exact observability assumptions (see Theorem 1 below). Consequently, we develop an iterative VI algorithm by borrowing the stochastic approximation idea [22–24].

The main contributions of this paper are outlined below.

• In contrast to the PI algorithm described in [11, 13], the developed VI algorithm can initiate from any positive definite matrix. Consequently, the need to search for a stabilizing control is eliminated.

• We show that the VI algorithm is effective at solving policy evaluation steps (see (5) below) arising in the PI algorithm. A simulation example shows that the VI algorithm may outperform the strategy developed by Kleinman [25] in solving the policy evaluation step.

• In summary monograph [20], the authors presented a VI algorithm for an ergodic control problem under invariant probability measure (i.e., system state $x(s), \forall s > 0$, of (1), has the same probability distribution as the initial state $x(0)$). Since this paper does not rely on the invariant probability measure assumption, obtaining the asymptotic property is more difficult. By overcoming these difficulties, we develop the VI algorithm for solving the LQSOC problem. Thus, the results obtained herein may serve as a useful generalization of their findings.

This article is organized as follows. Section 2 formulates the LQSOC problem and provides some preliminary information. Section 3 introduces the VI algorithm and demonstrates its utility in solving policy evaluation steps. Section 4 validates the obtained results using a numerical example. Section 5 outlines the conclusion and outlook of this paper.

2 Problem formulation and some preliminaries

Let us begin with $\mathbb{Z}$ and $\mathbb{R}$, which are collections of non-negative integers and real numbers, respectively. Let $\mathbb{R}^{l\times m}$ denote the collection of all $l \times m$ real matrices. $\mathbb{R}^l$ denotes the l-dimensional Euclidean space. $|\cdot|_F$ is the Frobenius norm for matrices. $|\cdot|$ is the Euclidean norm for vectors, or induced matrix norm for matrices. Zero matrix (or vector) is defined as 0. $\mathrm{diag}\{r\}$ is a square diagonal matrix whose main diagonal is made up of the elements of vector r. I_l is the l-dimensional identity matrix. The transpose of a vector or matrix G is indicated by G^{T}. $\boldsymbol{S}^l_+$, $\boldsymbol{S}^l_{++}$, and $\boldsymbol{S}^l$ indicate the collections of all positive semidefinite matrices, positive definite matrices, and symmetric matrices in $\mathbb{R}^{l\times l}$, respectively. For any matrix $G \in \boldsymbol{S}^l_{++}$ (resp. $G \in \boldsymbol{S}^l_+$), let $G > 0$ (resp. $G \geqslant 0$). For matrices $G \in \boldsymbol{S}^l$, $L \in \boldsymbol{S}^l$, we write $G > L$ (resp. $G \geqslant L$) if $G - L > 0$ (resp. $G - L \geqslant 0$). For any $G \in \boldsymbol{S}^m$, let g_{ji} be the (j,i)th element of matrix G, and let $\mathrm{vecs}(G) := [g_{11}, \sqrt{2}g_{12}, \ldots, \sqrt{2}g_{1m}, g_{22}, \sqrt{2}g_{23}, \ldots, \sqrt{2}g_{m-1m}, g_{mm}]^{\mathrm{T}}$. Furthermore, $w(\cdot)$ is a one-dimensional standard Brownian motion defined on a complete filtered probability space

$(\Omega, \mathcal{F}, \{\mathcal{F}_s\}_{s\geqslant 0}, \mathbb{P})$. We define a Hilbert space as

$$\mathcal{L}^2_{\mathcal{F}}(\mathbb{R}^l) := \left\{ \Psi(\cdot) : [0,+\infty)\times\Omega \ \to \mathbb{R}^l \,\middle|\, \Psi(\cdot) \text{ is } \mathcal{F}_s\text{-adapted and } \mathbb{E}\int_0^{+\infty} |\psi(s,\omega)|^2 \mathrm{d}s < +\infty \right\}.$$

We consider a stochastic system described by

$$\begin{cases} \mathrm{d}x(s) = \big[Ax(s)+Bu(s)\big]\mathrm{d}s + \big[Cx(s)+Du(s)\big]\mathrm{d}w(s), \\ x(0)=\xi, \end{cases} \tag{2}$$

where $B, D \in \mathbb{R}^{n\times m}$, $A, C \in \mathbb{R}^{n\times n}$ are constant matrices and the initial state $\xi \in \mathbb{R}^n$ is a constant vector. The corresponding cost functional is in the form of

$$J(u(\cdot)) = \mathbb{E}\int_0^{+\infty} \big[u(s)^{\mathrm{T}} R u(s) + x(s)^{\mathrm{T}} Q x(s)\big]\mathrm{d}s, \tag{3}$$

where R and Q are constant matrices of proper sizes.

We now define mean-square stabilizability and exact observability, which are commonly used in infinite-horizon LQSOC problems.

Definition 1 ([9, 10]). System (2) is called mean-square stabilizable if for every $\xi \in \mathbb{R}^n$, there is a matrix $K \in \mathbb{R}^{m\times n}$ such that $\lim_{s\to+\infty} \mathbb{E}[x(s)^{\mathrm{T}} x(s)] = 0$, where $x(\cdot)$ is governed by

$$\begin{cases} \mathrm{d}x(s) = (A+BK)x(s)\mathrm{d}s + (C+DK)x(s)\mathrm{d}w(s), \\ x(0)=\xi. \end{cases} \tag{4}$$

In this case, we call the control $u(\cdot) = Kx(\cdot)$ a (mean-square) stabilizing control.

Definition 2 ([26]). $[A, C|Q]$ is called exactly observable, if for any $S > 0$, the system

$$\begin{cases} \mathrm{d}x(s) = Ax(s)\mathrm{d}s + Cx(s)\mathrm{d}w(s), \\ h(s) = Qx(s), \\ x(0)=\xi \end{cases}$$

satisfies

$$h(s) \equiv 0 \,\text{a.s.}, \forall s \in [0,S] \Rightarrow x(0) = \xi = 0.$$

Assumption 1. System (2) is mean-square stabilizable.

Assumption 2. $Q \geqslant 0$, $R > 0$, and $[A, C|Q]$ is exactly observable.

Define an admissible control set as

$$\mathcal{U}_{ad} := \{u(\cdot) \in \mathcal{L}^2_{\mathcal{F}}(\mathbb{R}^m) \,|\, u(\cdot) \text{ is stabilizing}\}.$$

The LQSOC problem is given below.

Problem (LQSOC). Given $\xi \in \mathbb{R}^n$, our task is to find a suitable control $u^*(\cdot) \in \mathcal{U}_{ad}$ such that

$$J(u^*(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}_{ad}} J(u(\cdot)).$$

For any $\xi \in \mathbb{R}^n$, Problem (LQSOC) is said to be well-posed if $-\infty < \inf_{u(\cdot)\in\mathcal{U}_{ad}} J(u(\cdot)) < +\infty$. Under Assumptions 1 and 2, it is simple to show that $0 \leqslant \inf_{u(\cdot)\in\mathcal{U}_{ad}} J(u(\cdot)) < +\infty$, implying that Problem (LQSOC) is well-posed.

The stabilizing solution to GARE (1) is then defined, which is closely related to the optimal control of Problem (LQSOC).

Definition 3 ([10]). If GARE (1) has a solution P such that

$$u(\cdot) = -(D^{\mathrm{T}} P D + R)^{-1}(D^{\mathrm{T}} P C + B^{\mathrm{T}} P)x(\cdot)$$

is stabilizing for system (2), then P is referred to as a stabilizing solution to GARE (1).

To conclude this section, we cite the existing PI method, which not only presents a strategy for approximating the solution of GARE (1) but also establishes the relationship between the stabilizing solution to GARE (1) and the optimal control of Problem (LQSOC).

Lemma 1 ([11, 13])**.** Suppose Assumptions 1 and 2 hold and a stabilizing control $u(\cdot) = K_0 x(\cdot)$ is known. If P_k is solved by

$$P_k(A + BK_k) + Q + (A + BK_k)^{\mathrm{T}} P_k + K_k^{\mathrm{T}} R K_k + (C + DK_k)^{\mathrm{T}} P_k (C + DK_k) = 0, \tag{5}$$

and

$$K_{k+1} = -(D^{\mathrm{T}} P_k D + R)^{-1}(D^{\mathrm{T}} P_k C + B^{\mathrm{T}} P_k), \tag{6}$$

then we have

(i) K_k and P_k, $k = 0, 1, 2, \ldots$ can be uniquely determined at each iteration step;

(ii) $u(\cdot) = K_k x(\cdot)$, $k = 0, 1, 2, \ldots$ are stabilizing;

(iii) $\{P_k\}_{k=0}^{+\infty}$ converges to the unique stabilizing solution P^* to GARE (1), $\{K_k\}_{k=1}^{+\infty}$ converges to $K^* = -(D^{\mathrm{T}} P^* D + R)^{-1}(D^{\mathrm{T}} P^* C + B^{\mathrm{T}} P^*)$, and the optimal control of Problem (LQSOC) is $u^*(\cdot) = K^* x^*(\cdot)$.

In the theory of reinforcement learning, a policy indicates an agent's behavior at a given time. According to this concept, in the context of control theory, Eq. (5) is usually referred to as policy evaluation and Eq. (6) is referred to as policy improvement. By employing these two steps alternately, the PI algorithm converges to the optimal values K^* and P^*. However, as previously stated, a stabilizing control should be known prior to running the PI algorithm.

3 VI algorithm for problem (LQSOC)

In this section, we propose an iterative VI algorithm to numerically approximate P^* without the prerequisite of an initial stabilizing control. Furthermore, we also present some related results to demonstrate that the obtained VI algorithm can be used to calculate step (5) arising in the PI algorithm.

To that end, we first perform an asymptotic analysis for GARE (1), which is a key component and motivation for the VI algorithm.

Theorem 1. Let Assumptions 1 and 2 hold. Given a GDRE,

$$\begin{cases} \dot{P}(t) - (C^{\mathrm{T}} P(t) D + P(t) B)(D^{\mathrm{T}} P(t) D + R)^{-1}(D^{\mathrm{T}} P(t) C + B^{\mathrm{T}} P(t)) \\ \qquad + P(t) A + Q + C^{\mathrm{T}} P(t) C + A^{\mathrm{T}} P(t) = 0, \\ P(T) = N, \\ D^{\mathrm{T}} P(t) D + R > 0 \end{cases} \tag{7}$$

with $N \in \boldsymbol{S}_+^n$, then the solution $P(t)$ to GDRE (7) is monotonically nondecreasing as t decreases and satisfies $\lim_{t \to -\infty} P(t) = P^*$, where P^* is the unique stabilizing solution to GARE (1).

Before concluding the proof of Theorem 1, we present the following lemma to demonstrate some asymptotic properties of GDRE (7).

Lemma 2. Suppose Assumption 1 holds, $Q \geqslant 0$, and $R > 0$.

(i) For any $P(T) = N \geqslant 0$, the solution $P(t)$ of GDRE (7) exists. In this case, $P(t)$ is monotonically nondecreasing as $t \to -\infty$, and $P(t)$ converges to a positive semidefinite solution to GARE (1) as t declines;

(ii) If a solution of GDRE (7) with $P(T) = 0$ converges to P^* as t drops, then any solution to GDRE (7) with a terminal condition $P(T) \geqslant 0$ converges to P^*.

Proof. (i) is a special case of Theorem 4.1 in Ait Rami et al. [10]. (ii) can be derived by applying Theorems 4.2 and 4.6 in Ait Rami et al. [10]. The proof is finished.

With the help of Lemma 2, we can now prove Theorem 1.

Proof of Theorem 1. According to (ii) of Lemma 2, it is sufficient to demonstrate that $\lim_{t \to -\infty} P(t) = P^*$ for $P(T) = 0$. Indeed, (i) of Lemma 2 shows that $\lim_{t \to -\infty} P(t) = \hat{P} \geqslant 0$, where $P(\cdot)$ is the solution to GDRE (7) with $P(T) = 0$ and $\hat{P}$ is a solution to GARE (1). It then suffices to prove that $\hat{P} = P^*$.

To that end, we demonstrate that $\hat{P} > 0$. Inserting $\hat{K} = -(D^{\mathrm{T}} \hat{P} D + R)^{-1}(D^{\mathrm{T}} \hat{P} C + B^{\mathrm{T}} \hat{P})$ into (4)

and applying Ito's formula to $x^{\mathrm{T}}(s)\hat{P}x(s)$, we get

$$\begin{aligned}&\mathrm{d}(x(s)^{\mathrm{T}}\hat{P}x(s))\\&\quad=\{x(s)^{\mathrm{T}}((C+D\hat{K})^{\mathrm{T}}\hat{P}(C+D\hat{K})+(A+B\hat{K})^{\mathrm{T}}\hat{P}+\hat{P}(A+B\hat{K}))x(s)\}\mathrm{d}s+\{\cdots\}\mathrm{d}w(s)\\&\quad=\{x(s)^{\mathrm{T}}(A^{\mathrm{T}}\hat{P}+C^{\mathrm{T}}\hat{P}C+\hat{P}A-(C^{\mathrm{T}}\hat{P}D+\hat{P}B)(D^{\mathrm{T}}\hat{P}D+R)^{-1}(D^{\mathrm{T}}\hat{P}C+B^{\mathrm{T}}\hat{P})\\&\qquad-\hat{K}^{\mathrm{T}}R\hat{K})x(s)\}\mathrm{d}s+\{\cdots\}\mathrm{d}w(s).\end{aligned}\tag{8}$$

For any fixed $S>0$, Eq. (8) and GARE (1) imply

$$0\leqslant\mathbb{E}\int_0^S\big[x(s)^{\mathrm{T}}(\hat{K}^{\mathrm{T}}R\hat{K}+Q)x(s)\big]\mathrm{d}s=\mathbb{E}\big[\xi^{\mathrm{T}}\hat{P}\xi-x(S)^{\mathrm{T}}\hat{P}x(S)\big].\tag{9}$$

Suppose $\hat{P}>0$ does not hold; then there exists a $\xi\neq0$ such that $\hat{P}\xi=0$. In view of (9), we arrive at

$$0\leqslant\mathbb{E}\int_0^S\big[x(s)^{\mathrm{T}}(\hat{K}^{\mathrm{T}}R\hat{K}+Q)x(s)\big]\mathrm{d}s=\mathbb{E}\big[\xi^{\mathrm{T}}\hat{P}\xi-x(S)^{\mathrm{T}}\hat{P}x(S)\big]=-\mathbb{E}\big[x(S)^{\mathrm{T}}\hat{P}x(S)\big]\leqslant0,$$

which implies

$$(Q+\hat{K}^{\mathrm{T}}R\hat{K})x(s)=0\text{ a.s., }\forall s\in[0,S].\tag{10}$$

According to the stochastic Popov-Belevith-Hautus criterion for exact observability (e.g., Theorem 4 in Zhang and Chen [26]), we know that the exact observability of $\big[A,C|Q\big]$ implies that of $\big[A+B\hat{K},C+D\hat{K}|Q+\hat{K}^{\mathrm{T}}R\hat{K}\big]$. Clearly, Eq. (10) contradicts with the exact observability of $[A+B\hat{K},C+D\hat{K}|Q+\hat{K}^{\mathrm{T}}R\hat{K}]$, so we have $\hat{P}>0$.

Since $\hat{P}$ is a solution to GARE (1), GARE (1) can be transformed to

$$\begin{cases}\hat{P}(A+B\hat{K})+(C+D\hat{K})^{\mathrm{T}}P(C+D\hat{K})+(A+B\hat{K})^{\mathrm{T}}\hat{P}=-\hat{K}^{\mathrm{T}}R\hat{K}-Q,\\D^{\mathrm{T}}\hat{P}D+R>0.\end{cases}$$

Using Theorem 6 in Zhang and Chen [26], we can deduce that $\hat{P}=P^*>0$ is the stabilizing solution of GARE (1). Then the proof is complete.

Remark 1. Since GDRE (7) is a nonlinear backward differential matrix equation, we can reverse the timeline in (7) to get a forward differential matrix equation:

$$\begin{cases}\dot{P}(t)=A^{\mathrm{T}}P(t)+Q-(C^{\mathrm{T}}P(t)D+P(t)B)(D^{\mathrm{T}}P(t)D+R)^{-1}(D^{\mathrm{T}}P(t)C+B^{\mathrm{T}}P(t))\\\qquad\quad+C^{\mathrm{T}}P(t)C+P(t)A,\\P(0)=N,\\D^{\mathrm{T}}P(t)D+R>0.\end{cases}\tag{11}$$

Obviously, Theorem 1 indicates that $\lim_{t\to+\infty}P(t)=P^*$, where $P(\cdot)$ denotes the solution to (11) with $P(0)=N\in\boldsymbol{S}_+^n$.

Theorem 1 and Remark 1 mean that the stabilizing solution P^* of GARE (1) can be solved by the limit of the solution to GDRE (7). However, due to the nonlinear structure of the GDRE, obtaining the analytical solution of GDRE (7) remains difficult. Inspired by [11, 13, 19] hereinafter, we will propose a novel VI algorithm to approximate P^* using this asymptotic property.

We are now going to define some symbols that will be used in the proposed algorithm. $\{\mathcal{H}_d\}_{d=0}^{+\infty}$ is defined as a sequence of bounded collections with nonempty interiors that satisfy

$$\lim_{d\to+\infty}\mathcal{H}_d=\boldsymbol{S}_+^n,\ \mathcal{H}_d\subseteq\mathcal{H}_{d+1},\ \forall d\in\mathbb{Z}.$$

$0<\gamma_k\in\mathbb{R}$, $k=0,1,2,\ldots$, and $\{\gamma_k\}_{k=0}^{+\infty}$ satisfies

$$\sum_{k=0}^{+\infty}\gamma_k=+\infty,\lim_{k\to+\infty}\gamma_k=0.$$

The VI algorithm is summarized in Algorithm 1, and its convergence proof is presented in the following theorem.

Algorithm 1 VI algorithm

Select $P_0 > 0$. $d \leftarrow 0, k \leftarrow 0$.
repeat
  $\widetilde{P}_{k+1} \leftarrow P_k + \gamma_k[A^{\mathrm{T}}P_k + P_kA + C^{\mathrm{T}}P_kC + Q - (P_kB + C^{\mathrm{T}}P_kD)(R + D^{\mathrm{T}}P_kD)^{-1}(B^{\mathrm{T}}P_k + D^{\mathrm{T}}P_kC)]$;
  if $\widetilde{P}_{k+1} \in \mathcal{H}_d$ **then**
    $P_{k+1} \leftarrow \widetilde{P}_{k+1}$;
  else
    $P_{k+1} \leftarrow P_0, d \leftarrow d + 1$;
  end if
  $k \leftarrow k + 1$;
until $|\widetilde{P}_{k+1} - P_k|/\gamma_k < \varepsilon$, where ε is a small constant threshold.

Theorem 2. Suppose Assumptions 1 and 2 hold; then we obtain

(i) There exists a compact set $\mathcal{W} \in \boldsymbol{S}_+^n$ and an integer $\hat{h} \in \mathbb{Z}$ such that $P^* \in \mathcal{W}$ and $\{P_k\}_{k=\hat{h}}^{+\infty} \subset \mathcal{W}$;

(ii) $\{P_k\}_{k=0}^{+\infty}$ obtained by Algorithm 1 satisfies $\lim_{k\to+\infty} P_k = P^*$.

Proof. (i) First, we convert (11) into an ordinary differential equation (t will be suppressed for illustrating simplicity):

$$\dot{p} = f(p), \tag{12}$$

where $p = \mathrm{vecs}(P)$, $\mathscr{P} := \{P \in \boldsymbol{S}^n \mid R + D^{\mathrm{T}}PD > 0\}$, and $f(\cdot) : \mathrm{vecs}(\mathscr{P}) \to \mathbb{R}^{n(n+1)/2}$ is

$$f(p) := \mathrm{vecs}(PA + Q + C^{\mathrm{T}}PC + A^{\mathrm{T}}P - (C^{\mathrm{T}}PD + PB)(D^{\mathrm{T}}PD + R)^{-1}(D^{\mathrm{T}}PC + B^{\mathrm{T}}P)). \tag{13}$$

According to Theorem 1 and Remark 1, if $P(0) \in \boldsymbol{S}_+^n$, the solution $P(\cdot)$ of (11) converges to P^*. Thus, we know that $p^* = \mathrm{vecs}(P^*)$ is a locally asymptotically stable equilibrium point (see Khalil [27]) of (12).

Next, we show that the function $f(p)$ is locally Lipschitz in $\mathrm{vecs}(\mathscr{P})$. To accomplish this, we simply need to demonstrate the local Lipschitz property of

$$g(p) := \mathrm{vecs}((C^{\mathrm{T}}PD + PB)(D^{\mathrm{T}}PD + R)^{-1}(D^{\mathrm{T}}PC + B^{\mathrm{T}}P)).$$

In fact, for any P_1, $P_2 \in \boldsymbol{S}^n$ satisfying $\mathrm{vecs}(P_1)$, $\mathrm{vecs}(P_2) \in \mathrm{vecs}(\mathscr{P})$, we know

$$\begin{aligned}
&|(C^{\mathrm{T}}P_1D + P_1B)(D^{\mathrm{T}}P_1D + R)^{-1}(D^{\mathrm{T}}P_1C + B^{\mathrm{T}}P_1)\\
&\quad - (C^{\mathrm{T}}P_2D + P_2B)(D^{\mathrm{T}}P_2D + R)^{-1}(D^{\mathrm{T}}P_2C + B^{\mathrm{T}}P_2)|_F\\
&= |(C^{\mathrm{T}}P_1D + P_1B)(D^{\mathrm{T}}P_1D + R)^{-1}[D^{\mathrm{T}}(P_1 - P_2)C + B^{\mathrm{T}}(P_1 - P_2)]\\
&\quad + (C^{\mathrm{T}}P_1D + P_1B)(D^{\mathrm{T}}P_1D + R)^{-1}(D^{\mathrm{T}}P_2C + B^{\mathrm{T}}P_2)\\
&\quad - [C^{\mathrm{T}}(P_2 - P_1)D + (P_2 - P_1)B](D^{\mathrm{T}}P_2D + R)^{-1}(D^{\mathrm{T}}P_2C + B^{\mathrm{T}}P_2)\\
&\quad - (C^{\mathrm{T}}P_1D + P_1B)(D^{\mathrm{T}}P_2D + R)^{-1}(D^{\mathrm{T}}P_2C + B^{\mathrm{T}}P_2)|_F\\
&= |(C^{\mathrm{T}}P_1D + P_1B)(D^{\mathrm{T}}P_1D + R)^{-1}[D^{\mathrm{T}}(P_1 - P_2)C + B^{\mathrm{T}}(P_1 - P_2)]\\
&\quad - [C^{\mathrm{T}}(P_2 - P_1)D + (P_2 - P_1)B](D^{\mathrm{T}}P_2D + R)^{-1}(D^{\mathrm{T}}P_2C + B^{\mathrm{T}}P_2)\\
&\quad + (C^{\mathrm{T}}P_1D + P_1B)[(D^{\mathrm{T}}P_1D + R)^{-1}(D^{\mathrm{T}}P_2D + R)(D^{\mathrm{T}}P_2D + R)^{-1}\\
&\quad - (D^{\mathrm{T}}P_1D + R)^{-1}(D^{\mathrm{T}}P_1D + R)(D^{\mathrm{T}}P_2D + R)^{-1}](D^{\mathrm{T}}P_2C + B^{\mathrm{T}}P_2)|_F\\
&= |(C^{\mathrm{T}}P_1D + P_1B)(D^{\mathrm{T}}P_1D + R)^{-1}[D^{\mathrm{T}}(P_1 - P_2)C + B^{\mathrm{T}}(P_1 - P_2)]\\
&\quad - [C^{\mathrm{T}}(P_2 - P_1)D + (P_2 - P_1)B](D^{\mathrm{T}}P_2D + R)^{-1}(D^{\mathrm{T}}P_2C + B^{\mathrm{T}}P_2)\\
&\quad + (C^{\mathrm{T}}P_1D + P_1B)(D^{\mathrm{T}}P_1D + R)^{-1}[D^{\mathrm{T}}(P_2 - P_1)D](D^{\mathrm{T}}P_2D + R)^{-1}(D^{\mathrm{T}}P_2C + B^{\mathrm{T}}P_2)|_F.
\end{aligned} \tag{14}$$

Since $|\mathrm{vecs}(P)| = |P|_F$, $\forall P \in \boldsymbol{S}^n$, Eq. (14) and the property of Frobenius norm imply that $g(\cdot)$ is locally Lipschitz.

Finally, since $f(p)$ is locally Lipschitz, the converse Lyapunov theorem (e.g., Theorem 4.17 in Khalil [27]) can be applied and the set $\mathcal{W}$ and integer $\hat{h}$ can be constructed using procedures similar to the proof of Lemma 3.4 in Bian and Jiang [19].

(ii) First, based on (i) and Algorithm 1, it follows that

$$\begin{aligned}
P_{k+1} = P_k + \gamma_k\big(&A^{\mathrm{T}}P_k + C^{\mathrm{T}}P_kC + P_kA + Q\\
&- (P_kB + C^{\mathrm{T}}P_kD)(D^{\mathrm{T}}P_kD + R)^{-1}(B^{\mathrm{T}}P_k + D^{\mathrm{T}}P_kC)\big) + \mathcal{Z}_k,\ \forall k \geqslant \hat{h},
\end{aligned}$$

where

$$\mathcal{Z}_k := \begin{cases} P_0 - \widetilde{P}_{k+1}, & \text{if } \widetilde{P}_{k+1} \notin \mathcal{W}, \\ 0, & \text{otherwise.} \end{cases}$$

Next, we define an interpolation

$$P^0(t) := \begin{cases} P_k, & t \in [t_k, t_{k+1}), \\ P_0, & t \leqslant 0, \end{cases}$$

and its left-shifted process $P^k(t) := P^0(t + t_k)$, $\forall t \in (-\infty, +\infty)$, where $t_k = \sum_{j=0}^{k-1} \gamma_j$ and $t_0 = 0$.

Following these definitions, we derive

$$\begin{aligned} P^k(t) = P_k + &\sum_{j=k}^{q(t+t_k)-1} \gamma_j \big(A^{\mathrm{T}} P_j + C^{\mathrm{T}} P_j C + P_j A + Q \\ &- (P_j B + C^{\mathrm{T}} P_j D)(D^{\mathrm{T}} P_j D + R)^{-1}(B^{\mathrm{T}} P_j + D^{\mathrm{T}} P_j C)\big) + \sum_{j=k}^{q(t+t_k)-1} \mathcal{Z}_j, \\ = P^k(0) + L^k(t) + Z^k(t) + G^k(t), &\quad \forall k \geqslant \hat{h}, \quad \forall t \geqslant 0, \end{aligned}$$

where

$$\begin{aligned} L^k(t) := \int_0^t &\big(A^{\mathrm{T}} P^k(s) + C^{\mathrm{T}} P^k(s) C + P^k(s) A + Q \\ &- (P^k(s) B + C^{\mathrm{T}} P^k(s) D)(D^{\mathrm{T}} P^k(s) D + R)^{-1}(B^{\mathrm{T}} P^k(s) + D^{\mathrm{T}} P^k(s) C)\big) \mathrm{d}s, \end{aligned}$$

$$Z^k(t) := \sum_{j=k}^{q(t+t_k)-1} \mathcal{Z}_j, \qquad q(t) := \begin{cases} 0, & t < 0, \\ i, & 0 \leqslant t_i \leqslant t < t_{i+1}, \end{cases}$$

and

$$\begin{aligned} G^k(t) := &\sum_{j=k}^{q(t+t_k)-1} \gamma_j \big(A^{\mathrm{T}} P_j + C^{\mathrm{T}} P_j C + P_j A + Q \\ &- (P_j B + C^{\mathrm{T}} P_j D)(D^{\mathrm{T}} P_j D + R)^{-1}(B^{\mathrm{T}} P_j + D^{\mathrm{T}} P_j C)\big) - L^k(t). \end{aligned}$$

In the preceding procedures, we assume that the term $\sum_{j=k}^{q(t+t_k)-1}$ equals zero when $t \in [0, \gamma_k)$.

Then, using the stochastic approximation method, which is similar to the proof of Theorem 3.3 in Bian and Jiang [19], we obtain the convergence of Algorithm 1. This brings the proof to a close.

Let us now give a corollary to analyze the convergence rate of the crucial updating equation:

$$P_{k+1} \leftarrow P_k + \gamma_k [A^{\mathrm{T}} P_k + C^{\mathrm{T}} P_k C + P_k A + Q - (P_k B + C^{\mathrm{T}} P_k D)(R + D^{\mathrm{T}} P_k D)^{-1}(B^{\mathrm{T}} P_k + D^{\mathrm{T}} P_k C)], \ \forall k \geqslant \hat{h}.$$

Corollary 1. $\{P_k\}_{k=\hat{h}}^{+\infty}$ converges in a sublinear rate to P^*.

Proof. Given that P^* is a solution of GARE (1), we derive

$$\begin{aligned} P_{k+1} - P^* = &\gamma_k [A^{\mathrm{T}} P_k + C^{\mathrm{T}} P_k C + P_k A + Q - (P_k B + C^{\mathrm{T}} P_k D)(D^{\mathrm{T}} P_k D + R)^{-1}(B^{\mathrm{T}} P_k + D^{\mathrm{T}} P_k C)] \\ &- \gamma_k [A^{\mathrm{T}} P^* + C^{\mathrm{T}} P^* C + P^* A + Q - (P^* B + C^{\mathrm{T}} P^* D)(D^{\mathrm{T}} P^* D + R)^{-1}(B^{\mathrm{T}} P^* + D^{\mathrm{T}} P^* C)] \\ &+ P_k - P^* \\ = &P_k - P^* + \gamma_k [A^{\mathrm{T}}(P_k - P^*) + C^{\mathrm{T}}(P_k - P^*) C + (P_k - P^*) A \\ &+ (C^{\mathrm{T}} P^* D + P^* B)(D^{\mathrm{T}} P^* D + R)^{-1}[D^{\mathrm{T}}(P^* - P_k) C + B^{\mathrm{T}}(P^* - P_k)] \\ &- [C^{\mathrm{T}}(P_k - P^*) D + (P_k - P^*) B](D^{\mathrm{T}} P_k D + R)^{-1}(D^{\mathrm{T}} P_k C + B^{\mathrm{T}} P_k) \\ &+ (C^{\mathrm{T}} P^* D + P^* B)(D^{\mathrm{T}} P^* D + R)^{-1}[D^{\mathrm{T}}(P_k - P^*) D](D^{\mathrm{T}} P_k D + R)^{-1}(D^{\mathrm{T}} P_k C + B^{\mathrm{T}} P_k)], \end{aligned}$$

where the final equality is obtained using procedures similar to (14).

Algorithm 2 PI implemented by Algorithm 1

Choose K_0 such that $u(\cdot) = K_0 x(\cdot)$ is a stabilizing control. Choose $P_0 > 0$ and select two small thresholds $\varepsilon_1 > 0$, $\varepsilon_2 > 0$. Initial $i \leftarrow 0$.
repeat
  $d \leftarrow 0, k \leftarrow 0$;
  loop
    $\widetilde{P}_{k+1} \leftarrow P_k + \gamma_k[(A+BK_i)^{\mathrm{T}}P_k + P_k(A+BK_i) + (C+DK_i)^{\mathrm{T}}P_k(C+DK_i) + Q + K_i^{\mathrm{T}}RK_i]$;
    if $\widetilde{P}_{k+1} \in \mathcal{H}_d$ **then**
      $P_{k+1} \leftarrow \widetilde{P}_{k+1}$;
    else if $|\widetilde{P}_{k+1} - P_k|/\gamma_k < \varepsilon_1$ **then**
      return P_k;
    else
      $P_{k+1} \leftarrow P_0, d \leftarrow d+1$;
    end if
    $k \leftarrow k+1$;
  end loop
  $K_{i+1} = -(D^{\mathrm{T}}P_kD + R)^{-1}(D^{\mathrm{T}}P_kC + B^{\mathrm{T}}P_k)$;
  $i \leftarrow i+1$;
until $|K_{i+1} - K_i| < \varepsilon_2$.

Keeping in mind that $P^* \in \mathcal{W}$, $\{P_k\}_{k=\hat{h}}^{+\infty} \subset \mathcal{W}$, and $\lim_{k\to+\infty}\gamma_k = 0$, the above equation implies

$$\lim_{k\to+\infty}\frac{|P_{k+1}-P^*|}{|P_k-P^*|} = 1,$$

which completes the proof.

Furthermore, in light of the obtained VI algorithm, we summarize the implementation of the PI method in Algorithm 2, whose policy evaluation steps are calculated by Algorithm 1. The following theorem demonstrates the applicability of the VI algorithm in calculating the policy evaluation step and provides the convergence of Algorithm 2.

Theorem 3. Let Assumptions 1 and 2 hold. Then we have

(i) For any K_k obtained by Lemma 1, the solution P_k of (5) can be solved by running Algorithm 1;

(ii) $\{P_i\}_{i=0}^{+\infty}$ and $\{K_i\}_{i=1}^{+\infty}$ defined in Algorithm 2 satisfy $\lim_{i\to+\infty}K_i = K^*$ and $\lim_{i\to+\infty}P_i = P^*$.

Proof. Given $k \in \mathbb{Z}$, introduce a new system and a new cost functional as

$$\begin{cases}\mathrm{d}x(s) = \big[(A+BK_k)x(s) + 0\cdot u(s)\big]\mathrm{d}s + \big[(C+DK_k)x(s) + 0\cdot u(s)\big]\mathrm{d}w(s),\\ x(0) = \xi,\end{cases} \tag{15}$$

$$J(u(\cdot)) = \mathbb{E}\int_0^{+\infty}\big[u(s)^{\mathrm{T}}u(s) + x(s)^{\mathrm{T}}(K_kRK_k + Q)x(s)\big]\mathrm{d}s. \tag{16}$$

It is clear that Eq. (5) is the corresponding GARE to this new problem (15) with (16). According to Lemma 1, the system (15) is mean-square stabilizable. Furthermore, it follows from the stochastic Popov-Belevith-Hautus criterion for exact observability (e.g., Theorem 4 in Zhang and Chen [26]) that $\big[A+BK_k, C+DK_k|Q+K_k^{\mathrm{T}}RK_k\big]$, $\forall k \in \mathbb{Z}$ is exactly observable. Thus, (i) can be obtained by applying Algorithm 1 to problem (15) with (16). (ii) follows directly from (i) and Lemma 1. The proof is thus completed.

4 Numerical example

In this section, we implement Algorithms 1 and 2 to solve the corresponding GARE of Problem (LQSOC). All numerical results were generated by MATLAB (version R2020a) on a Windows computer (Windows 7) with an Intel(R) core(TM) i7-7700 CPU running @ 3.60 GHz and 8 GB of main memory. The coefficient matrices of system (2) are chosen according to Ait Rami and Zhou [9]:

$$A = \begin{bmatrix}-0.76165 & 0.25467 & -0.92393 & -0.17284 & -0.70045\\ 1.47398 & -0.68342 & 2.72667 & -0.60199 & -0.82622\\ 0.85298 & 0.81451 & -1.70868 & -1.56195 & 1.13239\\ 0.72233 & -0.18848 & 0.06988 & -0.38887 & -0.23301\\ 0.63808 & -1.03274 & -1.37728 & 0.65430 & -0.23439\end{bmatrix},\ B = \begin{bmatrix}1.40276 & -0.74610\\ 0.32687 & -1.7211\\ 0.06135 & -1.71576\\ -0.18905 & 0.10230\\ 0.42498 & -1.28586\end{bmatrix},$$

$$C=\begin{bmatrix} 0.15269 & 0.00291 & 0.00648 & -0.11443 & 0.46638 \\ -0.09445 & -0.35861 & 0.19884 & -0.14688 & -0.32973 \\ 0.64373 & 0.43718 & -0.34427 & 0.05755 & -0.21438 \\ -0.11438 & 0.04115 & -0.22659 & -0.06408 & 0.07433 \\ -0.16133 & 0.22956 & 0.30741 & 0.29844 & -0.38512 \end{bmatrix},\ D=\begin{bmatrix} 0.70138 & -0.37305 \\ 0.16344 & -0.86055 \\ 0.03068 & -0.85788 \\ -0.09453 & 0.05115 \\ 0.21249 & -0.64293 \end{bmatrix}.$$

We choose $R=I_2$, $Q=\mathrm{diag}\{0.5,1,1,0.1,1\}$, $P_0=0.01I_5$, $\gamma_k=100/(k+1)$, $\forall k\in\mathbb{Z}$, and

$$\mathcal{H}_d=\{P\in\boldsymbol{S}_+^n\,\big|\,|P|\leqslant 10(d+1)\},\ \forall d\in\mathbb{Z}.$$

By applying Algorithm 1, the obtained approximation value $\tilde{P}^*$ is

$$\tilde{P}^*=\begin{bmatrix} 0.55716 & 0.35001 & 0.21927 & 0.04877 & -0.48710 \\ 0.35001 & 1.45536 & 1.41043 & -1.08310 & -0.86988 \\ 0.21927 & 1.41043 & 1.84026 & -1.30250 & -0.88101 \\ 0.04877 & -1.08310 & -1.30250 & 2.26032 & 0.08835 \\ -0.48710 & -0.86988 & -0.88101 & 0.08835 & 1.86882 \end{bmatrix}.$$

To check the error of the proposed algorithm, let

$$\mathcal{R}(P)=\ PA+Q+A^{\mathrm{T}}P+C^{\mathrm{T}}PC-(C^{\mathrm{T}}PD+PB)(D^{\mathrm{T}}PD+R)^{-1}(D^{\mathrm{T}}PC+B^{\mathrm{T}}P).$$

In this case, the difference between $\tilde{P}^*$ and the true value P^* is $|\mathcal{R}(\tilde{P}^*)|=1.8392\times10^{-15}$, and then $\tilde{K}^*=-(D^{\mathrm{T}}\tilde{P}^*D+R)^{-1}(D^{\mathrm{T}}\tilde{P}^*C+B^{\mathrm{T}}\tilde{P}^*)$ is

$$\tilde{K}^*=\begin{bmatrix} -0.50199 & -0.12955 & -0.00208 & 0.11095 & -0.02374 \\ 0.43633 & 0.87418 & 0.97933 & -0.90574 & -0.41587 \end{bmatrix}.$$

By setting $\xi=[0.1,0,-0.5,2,-0.2]^{\mathrm{T}}$, the state trajectories of system (2) subject to $u(\cdot)=0$ and $u(\cdot)=\tilde{K}^*x(\cdot)$ are plotted in Figures 1 and 2. Figures 1 and 2 show that $\tilde{P}^*$ obtained by Algorithm 1 is, in fact, the stabilizing solution to GARE (1). Furthermore, to validate the performance of Algorithm 1 with various initial values, we run the VI algorithm under three different P_0. Table 1 shows the corresponding results.

In addition, to demonstrate the efficacy of Algorithm 1 in determining policy evaluation steps, we solve the original numerical example by implementing the PI algorithm in Lemma 1. For comparison purposes, we solve the policy evaluation step (5) by Algorithm 1 and by the method in Kleinman [25], respectively. Let

$$K_0=\begin{bmatrix} -0.69238 & -0.12700 & 0.03431 & 0.04340 & 0.02159 \\ 0.43765 & 0.89896 & 1.01608 & -0.92002 & -0.44953 \end{bmatrix}$$

and set other parameters as described at the beginning of this section. Table 2 shows the corresponding simulation results, which suggest that Algorithm 1 may be a more powerful method for calculating policy evaluation steps than the algorithm developed by Kleinman [25]. When compared with Table 2, Table 1 also implies that the computation error of Algorithm 1 is less than that of the PI algorithm. The above results indicate that the VI algorithm may outperform the existing PI method.

5 Conclusion and outlook

In this study, we investigated an infinite-horizon LQSOC problem in continuous time. First, we developed a novel VI algorithm to solve the problem. Unlike the PI algorithm established in [11,13], the VI algorithm does not require a stabilizing control to initiate the algorithm. Then, we showed that Algorithm 1 can be used to solve policy evaluation steps arising in the PI algorithm. Finally, we presented a simulation example to illustrate the advantages of the obtained results.

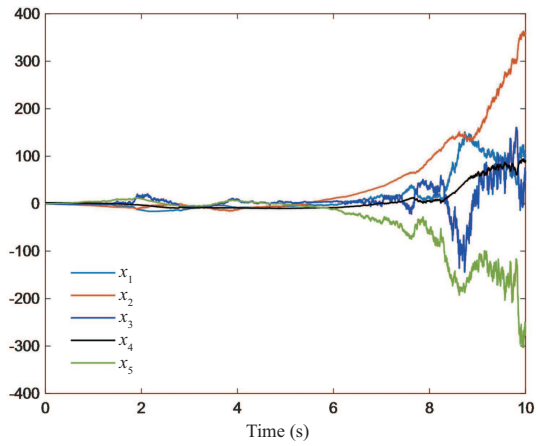

Figure 1 (Color online) State trajectories under $u(\cdot) = 0$.

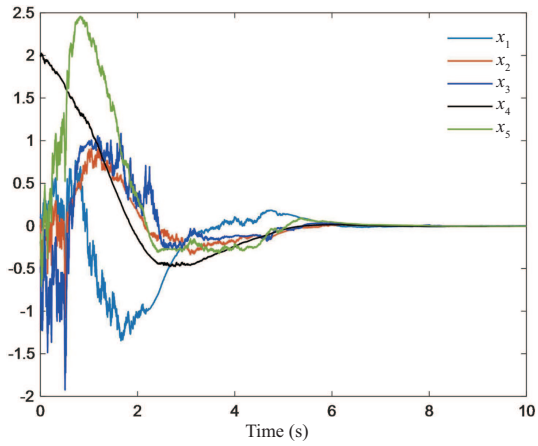

Figure 2 (Color online) State trajectories under $u(\cdot) = \bar{K}^* x(\cdot)$.

Table 1 Algorithm 1 with different P_0

	$P_0 = 0.05I_5$	$P_0 = 0.5I_5$	$P_0 = I_5$
$\lvert\mathcal{R}(\tilde{P}^*)\rvert$	2.2025×10^{-15}	1.5946×10^{-15}	1.6052×10^{-15}
CPU time (s)	0.1018	0.1109	0.4021

Table 2 Results of the PI algorithm

	PI implemented by Algorithm 1	PI implemented by Kleinman [25]
$\lvert\mathcal{R}(\tilde{P}^*)\rvert$	4.2798×10^{-15}	6.1466×10^{-14}
CPU time (s)	0.2480	0.7484

According to Chen et al. [28], one of the key differences between LQDOC problems and LQSOC problems is that the weighting matrix R in performance index (3) can be indefinite. In this paper, we only consider the case where R is positive definite. Otherwise, conducting an asymptotic analysis similar to Theorem 1 is difficult. This problem will be considered in the near future.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61925306, 61821004, 11831010), National Key R&D Program of China (Grant No. 2022YFA1006103), and Natural Science Foundation of Shandong Province (Grant Nos. ZR2019ZD42, ZR2020ZD24). The authors greatly appreciate the efforts of anonymous reviewers, which have improved the quality of this paper.

References

1 Wonham W M. On a matrix Riccati equation of stochastic control. SIAM J Control, 1968, 6: 681–697
2 Kolmanovsky V B, Shaikhet L E. Control of Systems with Aftereffect. Providence: American Mathematical Society, 1996
3 Yong J M, Zhou X Y. Stochastic Control: Hamiltonian Systems and HJB Equations. New York: Springer, 1999
4 Wang B C, Zhang H S, Zhang J F. Linear quadratic mean field social control with common noise: a directly decoupling method. Automatica, 2022, 146: 110619
5 Wang G C, Wu Z. A maximum principle for mean-field stochastic control system with noisy observation. Automatica, 2022, 137: 110135
6 Han Y C, Sun Y F. Stochastic linear quadratic optimal control problem for systems driven by fractional Brownian motions. Optim Control Appl Methods, 2019, 40: 900–913
7 Peng C C, Zhang W H. Multicriteria optimization problems of finite horizon stochastic cooperative linear-quadratic difference games. Sci China Inf Sci, 2022, 65: 172203
8 Hafayed M, Abba A, Abbas S. On partial-information optimal singular control problem for mean-field stochastic differential equations driven by Teugels martingales measures. Int J Control, 2016, 89: 397–410
9 Ait Rami M, Zhou X Y. Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls. IEEE Trans Automat Contr, 2000, 45: 1131–1143
10 Ait Rami M, Chen X, Moore J B, et al. Solvability and asymptotic behavior of generalized Riccati equations arising in indefinite stochastic LQ controls. IEEE Trans Automat Contr, 2001, 46: 428–440
11 Ni Y H, Fang H T. Policy iteration algorithm for singular controlled diffusion processes. SIAM J Control Optim, 2013, 51: 3844–3862
12 Damm T, Hinrichsen D. Newton's method for a rational matrix equation occurring in stochastic control. Linear Algebra Its Appl, 2001, 332–334: 81–109
13 Zhang W H. Study on algebraic Riccati equation arising from infinite horizon stochastic LQ optimal control. Dissertation for Ph.D. Degree. Hangzhou: Zhejiang University, 1998

14 Sun J R, Yong J M. Stochastic linear quadratic optimal control problems in infinite horizon. Appl Math Optim, 2018, 78: 145–183
15 Vandenberghe L, Boyd S. A primal-dual potential reduction method for problems involving matrix inequalities. Math Programming, 1995, 69: 205–236
16 Wei Q L, Liu D R, Lin H Q. Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems. IEEE Trans Cybern, 2016, 46: 840–853
17 Wei Q L, Lewis F L, Liu D R, et al. Discrete-time local value iteration adaptive dynamic programming: convergence analysis. IEEE Trans Syst Man Cybern Syst, 2018, 48: 875–891
18 Kleinman D L. On an iterative technique for Riccati equation computations. IEEE Trans Automat Contr, 1968, 13: 114–115
19 Bian T, Jiang Z P. Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design. Automatica, 2016, 71: 348–360
20 Jiang Z P, Bian T, Gao W N. Learning-based control: a tutorial and some recent results. FNT Syst Control, 2020, 8: 176–284
21 Xie K D, Yu X, Lan W Y. Optimal output regulation for unknown continuous-time linear systems by internal model and adaptive dynamic programming. Automatica, 2022, 146: 110564
22 Du K, Meng Q X, Zhang F. A Q-learning algorithm for discrete-time linear-quadratic control with random parameters of unknown distribution: convergence and stabilization. SIAM J Control Optim, 2022, 60: 1991–2015
23 Ljung L. Analysis of recursive stochastic algorithms. IEEE Trans Automat Contr, 1977, 22: 551–575
24 Kushner H J, Clark D S. Stochastic Approximation Methods for Constrained and Unconstrained Systems. New York: Springer, 1978
25 Kleinman D L. Numerical solution of the state dependent noise problem. IEEE Trans Automat Contr, 1976, 21: 419–420
26 Zhang W H, Chen B S. On stabilizability and exact observability of stochastic systems with their applications. Automatica, 2004, 40: 87–94
27 Khalil H K. Nonlinear Systems. Englewood Cliffs: Prentice Hall, 2002
28 Chen S P, Li X J, Zhou X Y. Stochastic linear quadratic regulators with indefinite control weight costs. SIAM J Control Optim, 1998, 36: 1685–1702