SCIENCE CHINA
Information Sciences

April 2024, Vol. 67, Iss. 4, 140202:1–140202:17
https://doi.org/10.1007/s11432-023-3982-y

• RESEARCH PAPER •
Special Topic: Analysis and Control of Stochastic Systems

Online Pareto optimal control of mean-field stochastic multi-player systems using policy iteration

Xiushan JIANG[1], Yanshuang WANG[1], Dongya ZHAO[1*] & Ling SHI[2]

[1]*College of New Energy, China University of Petroleum (East China), Qingdao* 266580*, China;*
[2]*Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong* 999077*, China*

Received 31 October 2023/Revised 25 January 2024/Accepted 26 February 2024/Published online 27 March 2024

Abstract In this study, the Pareto optimal strategy problem was investigated for multi-player mean-field stochastic systems governed by Itô differential equations using the reinforcement learning (RL) method. A partially model-free solution for Pareto-optimal control was derived. First, by applying the convexity of cost functions, the Pareto optimal control problem was solved using a weighted-sum optimal control problem. Subsequently, using on-policy RL, we present a novel policy iteration (PI) algorithm based on the $\mathcal{H}$-representation technique. In particular, by alternating between the policy evaluation and policy update steps, the Pareto optimal control policy is obtained when no further improvement occurs in system performance, which eliminates directly solving complicated cross-coupled generalized algebraic Riccati equations (GAREs). Practical numerical examples are presented to demonstrate the effectiveness of the proposed algorithm.

Keywords mean-field stochastic systems, Pareto optimal control, policy iteration scheme, $\mathcal{H}$-representation

1 Introduction

Many practical systems involve multiple players, and each player has its own performance objectives to achieve a meaningful trade-off, which typically results in a multi-objective optimization (MOO) problem. In addition, because many real systems have states and actions in continuous spaces, they are analytically and numerically solvable by modeling the MOO process as a set of linear differential equations and the performance objectives as quadratic utility functions, that is linear quadratic (LQ) differential games. In the field of multiple optimizations with the LQ structure, the Pareto solution, which means that not all players' costs can be improved upon simultaneously, plays a crucial role owing to its excellent ability to achieve a maximum allocation for the limited resources. It has been applied in various fields such as electric power systems [1], economics [2], and gun turret-barrel systems [3]. The published results on Pareto optimal control are theoretically elegant and practically reliable, e.g., the finite and infinite horizon cooperative Pareto efficient equilibria for regular indefinite differential games [4], Pareto-based guaranteed cost control problems [5], and Pareto optimal solutions and their application to the network security model [6]. The mean-field terms, particularly for that governed by Itô-type stochastic differential equations, have been extensively studied [7–9]. LQ Pareto-optimal feedback-control on mean-field stochastic systems (MFSS) has received significant attention. Lin et al. [10] discussed the Pareto game for nominal and uncertain MFSSs. In [11], for the LQ mean-field stochastic game with non-homogeneous terms, the necessary conditions for the existence of Pareto solutions were obtained by employing the equivalent description of Pareto efficiency. In [7], we considered the linear MFSS affected by external disturbances and derived the necessary and sufficient conditions for the Pareto optimal strategies under a H_∞ constraint. These aforementioned LQ Pareto strategies are used for solving related cross-coupled implicative generalized algebraic Riccati equations (GAREs). Although some approximate dynamic programming techniques have been developed to solve some special stochastic optimal control problems, such as the stochastic LQ optimal control based on Q-learning algorithm [12], the stochastic

* Corresponding author (email: dyzhao@upc.edu.cn)

two-player zero-sum and nonzero-sum games based on on-policy and off-policy reinforcement learning (RL) [13], and the stochastic H_∞ control by a data-driven approach [14]. To the best of our knowledge, no effective RL algorithm has been developed for solving the Pareto optimal control problem for MFSS with multiple cooperative players.

Motivated by the above research gap, a partially model-free algorithm based on a policy iteration (PI) technique which is important for theoretical and practical applications, is presented. The PI algorithm was first proposed under the framework of the Markov decision process [15], which is a class of RL methods based on Actor-Critic structure [16]. This structure involves two steps: the policy evaluation and the policy improvement. Because the Pareto optimal control theory with the LQ form is used to study cooperative feedback control with guaranteed performance, the PI method is an approach to learning optimal behaviours by first choosing an admissible control policy for the environment and then assessing the value of the controller. This is essentially an adaptive control technique, where the control policy is modified based on the responses from the environment, as discussed in [17], and it has been broadly applied to solve LQ optimal control problems in recent years [18, 19].

This study extends the results of [17] to MFSSs with multiple players, where the PI algorithm was first proposed to solve an algebraic Riccati equation online for deterministic systems. Although an extension of [17] to a simple stochastic system has been achieved, we further improved the systems with the mean-field term and multiple cooperative inputs. The main contributions of this study are as follows.

(1) An online algorithm was derived for a multi-player stochastic system. Our results introduce a novel approach to tackle the Pareto optimal control problems instead of solving the GAREs using all the system information. In contrast, in the related references [20, 21], in which employed online RL methods in systems with the same multiplicative stochastic noise, the studies were constrained to one-player, one-objective optimal control problems.

(2) In contrast to the algebraic Riccati equation discussed in [17], it is crucial to emphasize that the GAREs derived in this study have distinct structures for the reason that there are two coupled equations instead of one. Therefore, the considered algorithm is not a simple extension of deterministic counterparts.

(3) Different from the conventional online algorithms presented in [17, 19] which just used the straightening operator without removing the repeated elements from the symmetric matrix, we developed a novel algorithm for the related PI scheme using the $\mathcal{H}$-representation technique. This technique, initially introduced in [22], allows us to transform a symmetric matrix-valued equation into a vector-valued equation, effectively eliminating redundant elements.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the problem formulation and some useful preliminaries. We establish the equivalence between the Pareto optimal control problem and joint optimal control problem using the weighted-sum cost function for MFSSs. Section 3 presents the developed PI method and proves its convergence. The online implementation of the proposed adaptive algorithm is provided based on the $\mathcal{H}$-representation technique. Section 4 gives two examples that show the effectiveness of the PI scheme when applying the obtained algorithm. Finally, Section 5 presents the conclusions of the study with potential future extensions.

Notation. Throughout this paper, let $\mathcal{R}$ and $\mathcal{R}^{m\times n}$ denote the set of all real numbers and all $m\times n$ real matrices, respectively. $S>0$ (resp., $S\geqslant 0$) means that S is a real symmetric positive definite (resp. positive semi-definite) matrix. S' denotes the transpose of the matrix or vector S. $\mathbb{E}(x)$ indicates the mathematical expectation of a random variable x. $\mathcal{L}^2_{\mathcal{F}}([0,\infty),\mathcal{R}^{m_l})$ signifies the space of nonanticipative stochastic process $m(t)\in\mathcal{R}^{m_l}$ with respect to (w.r.t.) an increasing σ-algebra $\mathcal{F}_t$ $(t\geqslant 0)$ that satisfies $\mathbb{E}\int_0^\infty \|m(t)\|^2\mathrm{d}t<\infty$. $\mathcal{N}:=\{1,2,\ldots,N\}$. $\mathrm{Col}(A_1\ A_2\ \cdots\ A_n):=[\,A_1'\ A_2'\ \cdots\ A_n'\,]'$. $\mathrm{diag}\{A_1\ A_2\cdots\ A_n\}$ means the diagonal matrix with the main diagonal elements being $A_1, A_2, \ldots, A_n$.

2 Problem formulations and preliminaries

2.1 Pareto optimal control problem

Consider the following MFSS with multi-player (MFSS-MP):

$$\begin{aligned}&\mathrm{d}x_t=\left[Ax_t+\bar{A}\mathbb{E}x_t+\sum_{l=1}^{N}(B^lu_t^l+\bar{B}^l\mathbb{E}u_t^l)\right]\mathrm{d}t+\left[Cx_t+\bar{C}\mathbb{E}x_t+\sum_{l=1}^{N}(D^lu_t^l+\bar{D}^l\mathbb{E}u_t^l)\right]\mathrm{d}w_t,\\&x_0\in\mathcal{R}^n,\end{aligned}\tag{1}$$

where $x_t \in \mathcal{R}^n$ is the system state vector, $\{u_t^l \in \mathcal{R}^{m_l}\}_{l\in\mathcal{N}}$ is the set of the system control input of player l and w_t is a one-dimensional standard Wiener process defined on a complete filtered probability space $(\Omega, \mathcal{F}, \mathcal{P}, \{\mathcal{F}_t\}_{t\geqslant 0})$, where $\mathcal{F}_t = \sigma\{w_s : 0 \leqslant s \leqslant t\}$ is the natural filtration. $\{A, \bar{A}\}$ represents the internal system parameters which are assumed to be unknown in this paper. The coefficients $\{B^l, \bar{B}^l, C, \bar{C}, D^l, \bar{D}^l\}_{l\in\mathcal{N}}$ are known matrices with appropriate dimensions. The initial state $x_0 \in \mathcal{R}^n$ is assumed to be a deterministic vector.

For the online Pareto optimal control problem, the system matrices A and $\bar{A}$ are not necessary to be known in advance. The players $\{u_t^l\}_{l\in\mathcal{N}}$ cooperate with each other to minimize its cost function

$$J^l(x_0; u^1, u^2, \ldots, u^N) := \mathbb{E}\left\{\int_0^\infty \left[x_t' Q^l x_t + (\mathbb{E}x_t)'\bar{Q}^l \mathbb{E}x_t \sum_{i=1}^N ((u_t^i)' R^{li} u_t^i + (\mathbb{E}u_t^i)'\bar{R}^{li}\mathbb{E}u_t^i)\right] \mathrm{d}t\right\},\ l \in \mathcal{N}. \tag{2}$$

The state and control weighted matrices in cost function (2) are deterministic symmetric matrices that satisfy Assumption 1.

Assumption 1. $Q^l, \bar{Q}^l \geqslant 0$, $R^{li}, \bar{R}^{li} > 0$ with $i, l = 1, 2, \ldots, N$.

The admissible control set is given as follows:

$$\mathcal{U}_{[0,\infty)} := \left\{u_t^l \in \mathcal{L}^2_{\mathcal{F}}([0,\infty), \mathcal{R}^{m_l}) \middle| \mathbb{E}\int_0^\infty |u_s^l|^2 \mathrm{d}s < \infty, l \in \mathcal{N}\right\}.$$

This paper focuses on Pareto optimal solutions that cannot be simultaneously improved by all players. The mathematical definition for Pareto optimality can be found in Definition 1.1 in [4]. The objective is expressed as follows.

Problem 1. For system (1) with associated cost functions (2), all players collaborate to discover a collection of $\mathcal{F}_t$-adapted Pareto optimal controllers $u^l \in \mathcal{U}[0,\infty)$. The objective is to minimize each player's individual cost function (2) and stabilize the MFSS-MP (1).

For the Pareto optimal control problem, the weighted-sum method plays an important role.

Lemma 1 ([23]). Let $\rho_l \in (0,1)$ with $\sum_{l=0}^N \rho_l = 1$. Assume $u^* = \mathrm{Col}(u^{1*}\ u^{2*}\ \cdots\ u^{N*}) \in \mathcal{U}_{[0,\infty)}$, such that

$$u^* \in \arg\min_{u\in\mathcal{U}_{[0,\infty)}} \left\{\sum_{l=1}^N \rho_l J^l\left(x_0; u^{1*}, u^{2*}, \ldots, u^{N*}\right)\right\}. \tag{3}$$

Then u^* is Pareto optimal.

Lemma 2 ([24]). Assume that the admissible control set $\mathcal{U}_{[0,\infty)}$ is convex and the cost functions $J^l(x_0; u^1, u^2, \ldots, u^N)$, $l \in \mathcal{N}$ are convex w.r.t. u^l. Then, if $u^* = \mathrm{Col}(u^{1*}\ u^{2*}\ \cdots\ u^{N*})$ is Pareto efficient, there exists $\rho \in \mathcal{A} := \left\{\rho = (\rho_1, \ldots, \rho_N) \middle| \rho_l \geqslant 0 \text{ and } \sum_{l=1}^N \rho_l = 1\right\}$ such that u^* satisfies condition (3).

Since all players choose to cooperate to minimize the cost function, the joint action is represented as $u := \mathrm{Col}(u^1\ u^2\ \cdots\ u^N) \in \mathcal{R}^m, m = \sum_{i=1}^N$. As a result, system (1) and cost functions in (2) can be reformulated as follows:

$$\mathrm{d}x_t = [Ax_t + \bar{A}\mathbb{E}x_t + Bu_t + \bar{B}\mathbb{E}u_t]\mathrm{d}t + [Cx_t + \bar{C}\mathbb{E}x_t + Du_t + \bar{D}\mathbb{E}u_t]\mathrm{d}w_t,\ x_0 \in \mathcal{R}^n, \tag{4}$$

and

$$J^l(x_0; u) := \mathbb{E}\left\{\int_0^\infty [x_t' Q^l x_t + (\mathbb{E}x_t)'\bar{Q}^l\mathbb{E}x_t + (u_t)' R^l u_t + (\mathbb{E}u_t)'\bar{R}^l\mathbb{E}u_t]\mathrm{d}t\right\} \tag{5}$$

with

$$B = \left[\, B^1 \cdots B^N \,\right],\ \bar{B} = \left[\, \bar{B}^1 \cdots \bar{B}^N \,\right],\ D = \left[\, D^1 \cdots D^N \,\right],\ \bar{D} = \left[\, \bar{D}^1 \cdots \bar{D}^N \,\right],$$
$$R^l = \mathrm{diag}\{\, R^{l1} \cdots R^{lN} \,\},\ \bar{R}^l = \mathrm{diag}\{\, \bar{R}^{l1} \cdots \bar{R}^{lN} \,\}.$$

While Lemma 1 offers a sufficient condition for obtaining Pareto solutions, it is crucial to note that Lemmas 1 and 2 ensure that all Pareto solutions can be derived by solving the minimization problem (3) only if both the admissible control set $\mathcal{U}[0,\infty)$ and the cost functions $J^l(x_0; u^1, u^2, \ldots, u^N)$, $l \in \mathcal{N}$, are convex. Lemma 3 is essential in harnessing this property.

Lemma 3 ([10]). Consider the rewritten system (4) with the cost functions in (5). Set

$$J_\rho(x_0,u):=\sum_{l=1}^{N}\rho_l J^l(x_0,u),\ \rho\in\mathcal{A}.$$

Then $J_\rho(x_0,u)$ is convex w.r.t. $u\in\mathcal{U}_{[0,\infty)}$ with $\mathcal{U}_{[0,\infty)}$ being any convex subset of $\mathcal{R}^m$, if and only if (iff) $J_\rho(0,u)\geqslant 0$ for all $u\in\mathcal{U}_{[0,\infty)}$.

In Lemma 3, the weighted-sum objective function is defined as

$$J_\rho(x_0;u):=\sum_{l=1}^{N}\rho_l J^l(x_0;u)=\mathbb{E}\left\{\int_0^\infty[x_t'Q_\rho x_t+(\mathbb{E}x_t)'\bar{Q}_\rho\mathbb{E}x_t+(u_t)'R_\rho u_t+(\mathbb{E}u_t)'\bar{R}_\rho\mathbb{E}u_t]\mathrm{d}t\right\}. \quad (6)$$

Here, $Q_\rho=\sum_{l=1}^N\rho_l Q^l$, $\bar{Q}_\rho=\sum_{l=1}^N\rho_l\bar{Q}^l$, $R_\rho=\sum_{l=1}^N\rho_l R^l$, and $\bar{R}_\rho=\sum_{l=1}^N\rho_l\bar{R}^l$ for any $\rho\in\mathcal{A}$. Obviously, under Assumption 1, $J^l(x_0,u)\geqslant 0$ holds for all $l\in\mathcal{N}$. According to Lemmas 1–3, solving Problem 1 is to find a joint control u^* that minimizes the weighted-sum cost function $J_\rho(x_0,u)$ and stabilizes system (4). This can be stated as Problem 2.

Problem 2. For any player $u^l, l\in\mathcal{N}$, find a joint admissible controller $u^*\in\mathcal{U}_{[0,\infty)}$ to solve

$$\begin{cases}\inf\limits_{u\in\mathcal{U}_{[0,\infty)}} J_\rho(x_0;u),\\ \text{s.t.}\quad \mathrm{d}x_t=[Ax_t+\bar{A}\mathbb{E}x_t+Bu_t+\bar{B}\mathbb{E}u_t]\mathrm{d}t+[Cx_t+\bar{C}\mathbb{E}x_t+Du_t+\bar{D}\mathbb{E}u_t]\mathrm{d}w_t,\ x_0\in\mathcal{R}^n.\end{cases}$$

The related stabilizability definitions are given as follows.

Definition 1. System (4) is called stabilizable in the mean square sense if there exists controller $u\in\mathcal{U}_{[0,\infty)}$ with $u_t=K(x_t-\mathbb{E}x_t)+\bar{K}\mathbb{E}x_t$ and $\mathbb{E}u_t=\bar{K}\mathbb{E}x_t$ such that for any initial state $x_0\in\mathcal{R}^n$, the closed-loop system

$$\mathrm{d}x_t=[(A+BK)x_t+(\bar{A}+B(\bar{K}-K)+\bar{B}\bar{K})\mathbb{E}x_t]\mathrm{d}t+[(C+DK)x_t+(\bar{C}+D(\bar{K}-K)+\bar{D}\bar{K})\mathbb{E}x_t]\mathrm{d}w_t \quad (7)$$

is asymptotically mean-square stable (ASMS) with $x_0\in\mathcal{R}^n$.

Since $Q^l,\bar{Q}^l\geqslant 0$ are given in Assumption 1, exact observability is an essential condition in dealing with the stabilization problem.

Definition 2. Consider the following MFSS:

$$\begin{cases}\mathrm{d}x_t=(Ax_t+\bar{A}\mathbb{E}x_t)\mathrm{d}t+(Cx_t+\bar{C}\mathbb{E}x_t)\mathrm{d}w_t,\\ y_t=\mathbb{Q}^{\frac{1}{2}}\mathbb{X}_t\end{cases} \quad (8)$$

with

$$\mathbb{X}_t=\begin{bmatrix}\mathbb{E}x_t\\ x_t-\mathbb{E}x_t\end{bmatrix},\quad \mathbb{Q}=\begin{bmatrix}Q & 0\\ 0 & Q+\bar{Q}\end{bmatrix}.$$

If for any $T\geqslant 0$,

$$y_t=0,\text{a.s.},\forall t\in[0,T]\Longrightarrow x_0=0, \quad (9)$$

then system (8), or $(A,\bar{A},C,\bar{C}\mid\mathbb{Q}^{\frac{1}{2}})$ for simplicity, is said to be exactly observable.

For system (8), Assumption 2 is made.

Assumption 2. Assume $(A,\bar{A},C,\bar{C}\mid\mathbb{Q}^{\frac{1}{2}})$ is exactly observable.

2.2 Basic results in mean-field stochastic optimal control

In this paper, we assume that the considered system (1) is mean-square stabilizable and satisfies Assumption 2. Under Theorem 20 in [10] and Theorems 1 and 2 in [8], we have the following result.

Lemma 4 ([7])**.** Consider the rearranged mean-field stochastic system (4) as well as cost (5), MFSSs-MP (1) is mean-square stabilizable with the joint feedback control $u_t = Kx_t + (\bar{K} - K)\mathbb{E}x_t$ iff the coupled GAREs (10) admit a unique positive solution $(P, W > 0)$.

$$\begin{cases} PA + A'P + C'PC + Q_\rho - \mathcal{M}\mathcal{H}^{-1}\mathcal{M}' = 0, \\ W(A+\bar{A}) + (A+\bar{A})'W + (C+\bar{C})'P(C+\bar{C}) + Q_\rho + \bar{Q}_\rho - \widetilde{\mathcal{M}}\widetilde{\mathcal{H}}^{-1}\widetilde{\mathcal{M}}' = 0, \end{cases} \tag{10}$$

where $\mathcal{M}$, $\mathcal{H}$, $\widetilde{\mathcal{M}}$, and $\widetilde{\mathcal{H}}$ are given by

$$\mathcal{M} = PB + C'PD,\ \mathcal{H} = D'PD + R_\rho, \tag{11}$$

$$\widetilde{\mathcal{M}} = W(B+\bar{B}) + (C+\bar{C})'P(D+\bar{D}), \tag{12}$$

$$\widetilde{\mathcal{H}} = (D+\bar{D})'P(D+\bar{D}) + R_\rho + \bar{R}_\rho. \tag{13}$$

In this case, the optimal cooperative cost function is given as

$$J_\rho(x_0, u^*) = \inf_{u \in \mathcal{U}_{[0,\infty\}}} J(x_0; u) = x_0' W x_0 \tag{14}$$

with the joint stabilizing controller

$$u_t^* - \mathbb{E}u_t^* = -\mathcal{H}^{-1}\mathcal{M}'(x_t - \mathbb{E}x_t),\ \mathbb{E}u_t^* = -\widetilde{\mathcal{H}}^{-1}\widetilde{\mathcal{M}}'\mathbb{E}x_t.$$

For $K = -\mathcal{H}^{-1}\mathcal{M}'$, $\bar{K} = -\widetilde{\mathcal{H}}^{-1}\widetilde{\mathcal{M}}'$, there are

$$\mathcal{M}\mathcal{H}^{-1}\mathcal{M}' = K'\mathcal{M}' + \mathcal{M}K + K'\mathcal{H}K, \tag{15}$$

$$\widetilde{\mathcal{M}}\widetilde{\mathcal{H}}^{-1}\widetilde{\mathcal{M}}' = \bar{K}'\widetilde{\mathcal{M}}' + \widetilde{\mathcal{M}}\bar{K} + \bar{K}'\widetilde{\mathcal{H}}\bar{K}. \tag{16}$$

Remark 1. Under Assumption 2, the unique positive solution of GAREs (10) determines the stabilizing controller u_t^*. The cost weighting matrices $R^{li}, l, i \in \mathcal{N}$ in each player's cost function $J^l(x_0; u^1, u^2, \ldots, u^N)$ is assumed to be positive, which can be generalized to the indefinite case [25].

Remark 2. Lemma 4 represents a traditional method for solving the Pareto optimal control problem, employing a model-based approach to derive the controller by solving the GAREs (10). In contrast, the PI algorithm is implemented based on partial information about the system without relying on system matrices A and $\bar{A}$. Furthermore, the knowledge of state-related matrix parameters $(A, \bar{A}, C, \bar{C})$ is not required when implementing the online algorithm for $D = \boldsymbol{O}$ and $\bar{D} = \boldsymbol{O}$. This demonstrates the practicality and convenience of our method, as it eliminates the restrictive requirement concerning the knowledge of system dynamics.

The mean-square stabilizability of system (1) can also be obtained by the generalized algebraic Riccati inequalities (GARIs).

Lemma 5 ([7])**.** Consider the rearranged mean-field stochastic system (4). If there exist matrices $S > 0$, $\bar{S} > 0$, such that the following GARIs hold:

$$\begin{cases} (A+BK)'S + S(A+BK) + (C+DK)'S(C+DK) < 0, \\ (A+\bar{A}+(B+\bar{B})\bar{K})'\bar{S} + \bar{S}(A+\bar{A}+(B+\bar{B})\bar{K}) + (C+\bar{C}+(D+\bar{D})\bar{K})'S(C+\bar{C}+(D+\bar{D})\bar{K}) < 0, \end{cases} \tag{17}$$

then system (4) is mean-square stabilizable, and $u_t = Kx_t + (\bar{K} - K)\mathbb{E}x_t$ is a mean-square stabilizing controller.

Pareto solutions can be identified through the following result, which can be obtained using the same method as presented in Theorem 4.3 in [7]. We will omit the proof here for brevity.

Lemma 6. Consider system (1) under the Pareto efficient strategy u^*, then the set of all Pareto solutions is

$$\{J^1(x_0; K(x_t - \mathbb{E}x_t) + \bar{K}\mathbb{E}x_t), \ldots, J^N(x_0; K(x_t - \mathbb{E}x_t) + \bar{K}\mathbb{E}x_t) \mid \alpha \in \mathcal{A}\}$$

with $J^l(x_0;K(x_t-\mathbb{E}x_t)+\bar{K}\mathbb{E}x_t)=x_0'\bar{Z}^lx_0$, $l=1,2,\ldots,N$ satisfying the following generalized Lyapunov equation:

$$\begin{cases}\left(A+\sum\limits_{l=1}^{N}B^lK^l\right)'Z^l+Z^l\left(A+\sum\limits_{l=1}^{N}B^lK^l\right)+\left(C+\sum\limits_{l=1}^{N}D^lK^l\right)'Z^l\left(C+\sum\limits_{l=1}^{N}D^lK^l\right)\\ +\left(Q^l+\sum\limits_{i=1}^{N}(K^i)'R^{li}K^i\right)=0,\\ \left[C+\bar{C}+\sum\limits_{l=1}^{N}(D^l\bar{K}^l+\bar{D}^l\bar{K}^l)\right]'Z^l\left[C+\bar{C}+\sum\limits_{l=1}^{N}(D^l\bar{K}^l+\bar{D}^l\bar{K}^l)\right]+\left[A+\bar{A}+\sum\limits_{l=1}^{N}(B^l+\bar{B}^l\bar{K}^l)\right]'\bar{Z}^l\\ +\bar{Z}^l\left[A+\bar{A}+\sum\limits_{l=1}^{N}(B^l+\bar{B}^l\bar{K}^l)\right]+\bar{Q}^l-Q^l+\sum\limits_{i=1}^{N}[(\bar{K}^i)'(R^{li}+\bar{R}^{li})\bar{K}^i]=0.\end{cases}\tag{18}$$

3 Main results

3.1 PI algorithm

In this subsection, we present a novel PI algorithm for solving Problem 2 in an online fashion. The results demonstrate that we can achieve the optimal solution without requiring prior knowledge of the system matrices A and $\bar{A}$.

Under Assumption 2, K and $\bar{K}$ are the stabilizing gains for system (4) and the corresponding weighted-sum objective function is given by

$$\begin{aligned}\mathcal{C}(x_t)&=J_\rho(x_t;Kx_t+(\bar{K}-K)\mathbb{E}x_t)\\&=\mathbb{E}\int_t^\infty(x_\tau-\mathbb{E}x_\tau)'(Q_\rho+K'R_\rho K)(x_\tau-\mathbb{E}x_\tau)\mathrm{d}\tau+\int_t^\infty(\mathbb{E}x_\tau)'[Q_\rho+\bar{Q}_\rho+\bar{K}'(R_\rho+\bar{R}_\rho)\bar{K}](\mathbb{E}x_\tau)\mathrm{d}\tau\\&=\mathbb{E}[(x_t-\mathbb{E}x_t)'P(x_t-\mathbb{E}x_t)]+\mathbb{E}x_tW\mathbb{E}x_t.\end{aligned}\tag{19}$$

In (19), P and W are the real symmetric positive solutions of the following coupled Lyapunov matrix equations (CLMEs):

$$\begin{cases}K'\mathcal{M}'+\mathcal{M}K+K'\mathcal{H}K=-(PA+A'P+C'PC+Q_\rho),\\ \bar{K}'\widetilde{\mathcal{M}}'+\widetilde{\mathcal{M}}\bar{K}+\bar{K}'\widetilde{\mathcal{H}}\bar{K}=-[W(A+\bar{A})+(A+\bar{A})'W+(C+\bar{C})'P(C+\bar{C})+Q_\rho+\bar{Q}_\rho],\end{cases}$$

where $\mathcal{M}$, $\mathcal{H}$, $\widetilde{\mathcal{M}}$, and $\widetilde{\mathcal{H}}$ are given in (11)–(13).

Next, based on the Actor-Critic structure, we set the Critic update iteration scheme (20) and the Actor update iteration schemes (21) and (22) as follows:

Policy evaluation:

$$\begin{aligned}x_t'W_ix_t=&\mathbb{E}^{\mathcal{F}_t}\int_t^{t+T}(x_\tau-\mathbb{E}^{\mathcal{F}_t}x_\tau)'(Q_\rho+K_i'R_\rho K_i)(x_\tau-\mathbb{E}^{\mathcal{F}_t}x_\tau)\mathrm{d}\tau\\&+\int_t^{t+T}(\mathbb{E}^{\mathcal{F}_t}x_\tau)'[Q_\rho+\bar{Q}_\rho+\bar{K}_i'(R_\rho+\bar{R}_\rho)\bar{K}_i](\mathbb{E}^{\mathcal{F}_t}x_\tau)\mathrm{d}\tau+(\mathbb{E}^{\mathcal{F}_t}x_{t+T})'W_i(\mathbb{E}^{\mathcal{F}_t}x_{t+T})\\&+\mathbb{E}^{\mathcal{F}_t}[(x_{t+T}-\mathbb{E}^{\mathcal{F}_t}x_{t+T})'P_i(x_{t+T}-\mathbb{E}^{\mathcal{F}_t}x_{t+T})].\end{aligned}\tag{20}$$

Policy improvement:

$$K_{i+1}=-(D'P_iD+R_\rho)^{-1}(B'P_i+D'P_iC),\tag{21}$$

$$\bar{K}_{i+1}=-[(D+\bar{D})'P_i(D+\bar{D})+R_\rho+\bar{R}_\rho]^{-1}[(B+\bar{B})'W_i+(D+\bar{D})'P_i(C+\bar{C})].\tag{22}$$

Remark 3. The Actor/Critic scheme consists of a two-step iteration. One step involves updating the Critic structure (20), which is used to calculate the cost when the stabilizing controller is active. The other step involves updating the feedback gains for the controller, denoted as K_i and $\bar{K}_i$, with the goal of further reducing the cost compared with the current control policy.

3.2 The convergence proof

In the following, the convergence results for PIs (20)–(22) are given. For notational simplicity, we write

$$A_i = A + BK_i,\ \hat{A}_i = (A + \bar{A}) + (B + \bar{B})\bar{K}_i,\ C_i = C + DK_i,\ \hat{C}_i = (C + \bar{C}) + (D + \bar{D})\bar{K}_i,$$

and

$$\begin{aligned} &A_{i+1} = A + BK_{i+1},\ \hat{A}_{i+1} = (A + \bar{A}) + (B + \bar{B})\bar{K}_{i+1},\\ &C_i = C + DK_{i+1},\ \hat{C}_{i+1} = (C + \bar{C}) + (D + \bar{D})\bar{K}_{i+1}. \end{aligned}$$

Lemma 7 gives another equivalent description for policy evaluation (20).

Lemma 7. Assume u_t is a stabilizing controller with the control gains K_i and $\bar{K}_i$ and Assumption 2 holds. Then the solvability of policy evaluation (20) is equivalent to the solvability of the following coupled Lyapunov equations:

$$\begin{cases} A_i'P_i + P_iA_i + C_i'P_iC_i = -(Q_\rho + K_i'R_\rho K_i),\\ \hat{C}_i'P_i\hat{C}_i + \hat{A}_i'W_i + W_i\hat{A}_i = -[Q_\rho + \bar{Q}_\rho + \bar{K}_i'(R_\rho + \bar{R}_\rho)\bar{K}_i]. \end{cases} \tag{23}$$

Proof. Since the controller $u_t = K_ix_t + (\bar{K}_i - K_i)\mathbb{E}x_t$ is a stabilizable controller, according to Lemma 4, there exists a unique positive solution for the coupled Lyapunov equations (23). For system (4), the Lyapunov function is chosen as $\mathcal{C}_i(x_t) = (x_t - \mathbb{E}x_t)'P_i(x_t - \mathbb{E}x_t) + \mathbb{E}x_t'W_i\mathbb{E}x_t$. Applying the Itô formula to $\mathcal{C}_i(x_t)$, one has

$$\begin{aligned} \mathcal{L}(\mathcal{C}_i(x_t)) &= \mathbb{E}\{(x_t - \mathbb{E}x_t)'(P_iA_i + A_i'P_i + C_i'P_iC_i)(x_t - \mathbb{E}x_t)\} + (\mathbb{E}x_t)'(\hat{C}_i'P_i\hat{C}_i + \hat{A}_i'W_i + W_i\hat{A}_i)(\mathbb{E}x_t)\\ &= -\mathbb{E}\{(x_t - \mathbb{E}x_t)'(Q_\rho + K_i'R_\rho K_i)(x_t - \mathbb{E}x_t)\} - (\mathbb{E}x_t)'[Q_\rho + \bar{Q}_\rho + \bar{K}_i'(R_\rho + \bar{R}_\rho)\bar{K}_i](\mathbb{E}x_t). \end{aligned} \tag{24}$$

By Dynkin's formula (see Theorem 7.4.1 in [26]), from (24), we obtain

$$\mathbb{E}^{\mathcal{F}_t}\int_t^{t+T}\mathcal{L}(\mathcal{C}_i(x_\tau))\mathrm{d}t = \mathbb{E}^{\mathcal{F}_t}[(x_{t+T} - \mathbb{E}^{\mathcal{F}_t}x_{t+T})'P_i(x_{t+T} - \mathbb{E}^{\mathcal{F}_t}x_{t+T})] - x_t'W_ix_t + \mathbb{E}^{\mathcal{F}_t}x_{t+T}'W_i\mathbb{E}^{\mathcal{F}_t}x_{t+T}.$$

Hence, Eq. (20) is obtained directly. That is, from the unique solvability of the coupled equations (23), the solution of policy evaluation (20) is equivalent to the solution of (23). The proof is completed.

Next, we prove the convergence of the PI algorithm by assuming that the excitation condition is satisfied.

Theorem 1. Assume the feedback controller u_t is a stabilizing controller with the control policies K_i and $\bar{K}_i$ and Assumption 2 holds. If the policy improvement equations (21) and (22) are applied to update the control policy, then the new control policy $u_t = K_{i+1}x_t + (\bar{K}_{i+1} - K_{i+1})\mathbb{E}x_t$ is a stabilizing controller.

Proof. Similarly to Lemma 7, choose the Lyapunov function as $\mathcal{C}_i(x_t)$. When the updated controller gains K_{i+1} and $\bar{K}_{i+1}$ work, we applying Itô formula to $\mathcal{C}_i(x_t)$ along the trajectories generated by K_{i+1} and $\bar{K}_{i+1}$. Then,

$$\begin{aligned} \mathcal{L}(\mathcal{C}_i(x_t)) =& \mathbb{E}\{(x_t - \mathbb{E}x_t)'\mathcal{P}_{i+1}(x_t - \mathbb{E}x_t)\} + (\mathbb{E}x_t)'\mathcal{W}_{i+1}(\mathbb{E}x_t)\\ =& \mathbb{E}\{(x_t - \mathbb{E}x_t)'[P_iA_i + A_i'P_i + C_i'P_iC_i + P_i(A_{i+1} - A_i) + (A_{i+1} - A_i)'P_i + C_{i+1}'P_iC_{i+1}\\ &- C_i'P_iC_i](x_t - \mathbb{E}x_t)\} + (\mathbb{E}x_t)'[\hat{C}_i'P_i\hat{C}_i + \hat{A}_i'W_i + W_i\hat{A}_i + \hat{C}_{i+1}'P_i\hat{C}_{i+1} - \hat{C}_i'P_i\hat{C}_i\\ &+ (\hat{A}_{i+1} - \hat{A}_i)W_i + W_i(\hat{A}_{i+1} - \hat{A}_i)](\mathbb{E}x_t), \end{aligned} \tag{25}$$

where $\mathcal{P}_{i+1} = P_iA_{i+1} + A_{i+1}'P_i + C_{i+1}'P_iC_{i+1}$ and $\mathcal{W}_{i+1} = W_i\hat{A}_{i+1} + \hat{A}_{i+1}'W_i + \hat{C}_{i+1}'P_i\hat{C}_{i+1}$. Since the control policies K_i and $\bar{K}_i$ are stabilizable feedback gains, by the coupled Lyapunov equations (23), we obtain that

$$\begin{aligned} &\mathcal{P}_{i+1} = -(Q_\rho + K_i'R_\rho K_i) + P_i(A_{i+1} - A_i) + (A_{i+1} - A_i)'P_i + C_{i+1}'P_iC_{i+1} - C_i'P_iC_i,\\ &\mathcal{W}_{i+1} = -[Q_\rho + \bar{Q}_\rho + \bar{K}_i'(R_\rho + \bar{R}_\rho)\bar{K}_i] + W_i(\hat{A}_{i+1} - \hat{A}_i) + (\hat{A}_{i+1} - \hat{A}_i)'W_i + \hat{C}_{i+1}'P_i\hat{C}_{i+1} - \hat{C}_i'P_i\hat{C}_i. \end{aligned}$$

From the definitions of A_{i+1}, A_i, C_{i+1}, C_i, $\hat{A}_{i+1}$, $\hat{A}_i$, $\hat{C}_{i+1}$ and $\hat{C}_i$ given before, one obtains that

$$\begin{aligned}
&P_i(A_{i+1}-A_i)=P_iB(K_{i+1}-K_i),\\
&W_i(\hat{A}_{i+1}-\hat{A}_i)=W_i(B+\bar{B})(\bar{K}_{i+1}-\bar{K}_i),\\
&C'_{i+1}P_iC_{i+1}-C'_iP_iC_i=C'P_iD(K_{i+1}-K_i)+(K_{i+1}-K_i)'D'P_iC+K'_{i+1}D'P_iDK_{i+1}-K'_iD'P_iDK_i,\\
&\hat{C}'_{i+1}P_i\hat{C}_{i+1}-\hat{C}'_iP_i\hat{C}_i=(C+\bar{C})'P_i(D+\bar{D})(\bar{K}_{i+1}-\bar{K}_i)+(\bar{K}_{i+1}-\bar{K}_i)'(D+\bar{D})'P_i(C+\bar{C})\\
&\qquad\qquad+\bar{K}'_{i+1}(D+\bar{D})'P_i(D+\bar{D})\bar{K}_{i+1}-\bar{K}'_i(D+\bar{D})'P_i(D+\bar{D})\bar{K}_i.
\end{aligned}$$

Based on the policy improvement schemes (21) and (22),

$$P_i(A_{i+1}-A_i)+C'P_iD(K_{i+1}-K_i)=-K'_{i+1}(R_\rho+D'P_iD)(K_{i+1}-K_i),$$

and

$$W_i(\hat{A}_{i+1}-\hat{A}_i)+(C+\bar{C})'P_i(D+\bar{D})(\bar{K}_{i+1}-\bar{K}_i)=-\bar{K}'_{i+1}[R_\rho+\bar{R}_\rho+(D+\bar{D})'P_i(D+\bar{D})](\bar{K}_{i+1}-\bar{K}_i).$$

After some calculations, $\mathcal{P}_{i+1}$ and $\mathcal{W}_{i+1}$ can be simplified as

$$\begin{aligned}
&\mathcal{P}_{i+1}=-Q_\rho-K'_{i+1}R_\rho K_{i+1}-(K_{i+1}-K_i)'(R+D'P_iD)(K_{i+1}-K_i),\\
&\mathcal{W}_{i+1}=-Q_\rho-\bar{Q}_\rho-\bar{K}'_{i+1}(R_\rho+\bar{R}_\rho)\bar{K}_{i+1}-(\bar{K}_{i+1}-\bar{K}_i)'[R_\rho+\bar{R}_\rho+(D+\bar{D})'P_i(D+\bar{D})](\bar{K}_{i+1}-\bar{K}_i).
\end{aligned}$$

Thus, under Assumption 1, substituting $\mathcal{P}_{i+1}$ and $\mathcal{W}_{i+1}$ into (25) leads to $\mathbb{E}d(\mathcal{C}_i(x_t))\leqslant 0$, which shows that the updated control policy $u=K_{i+1}x_t+(\bar{K}_{i+1}-K_{i+1})\mathbb{E}x_t$ with K_{i+1} and $\bar{K}_{i+1}$ given by (21) and (22) is a stabilizing controller. The proof is thus completed.

Theorem 2. The solution sequences $\{P_i\}_{i=1}^\infty$ and $\{W_i\}_{i=1}^\infty$ determined by iteration schemes (21)–(23) satisfy

$$P_1\geqslant P_2\geqslant P_3\geqslant\cdots\geqslant P^*,\ W_1\geqslant W_2\geqslant W_3\geqslant\cdots\geqslant W^*,$$

where P^* and W^* are the unique positive solutions for the coupled GAREs (10) with $P^*,W^*>0$.

Proof. From (23),

$$\begin{cases}
A'_{i-1}P_{i-1}+P_{i-1}A_{i-1}+C'_{i-1}P_{i-1}C_{i-1}=-(Q_\rho+K'_{i-1}R_\rho K_{i-1}),\\
\hat{C}'_iP_{i-1}\hat{C}_{i-1}+\hat{A}'_{i-1}W_{i-1}+W_{i-1}\hat{A}_{i-1}=-[Q_\rho+\bar{Q}_\rho+\bar{K}'_{i-1}(R_\rho+\bar{R}_\rho)\bar{K}_{i-1}].
\end{cases}\tag{26}$$

Considering (23), we obtain

$$\begin{cases}
A'_i(P_{i-1}-P_i)+(P_{i-1}-P_i)A_i+C'_i(P_{i-1}-P_i)C_i=-(K_i-K_{i-1})'(R_\rho+D'P_{i-1}D)(K_i-K_{i-1}),\\
\hat{A}'_i(W_{i-1}-W_i)+(W_{i-1}-W_i)\hat{A}_i+\hat{C}'_i(P_{i-1}-P_i)\hat{C}_i\\
=-(\bar{K}_i-\bar{K}_{i-1})'[R_\rho+\bar{R}_\rho+(D+\bar{D})'P_{i-1}(D+\bar{D})](\bar{K}_i-\bar{K}_{i-1}).
\end{cases}$$

From Assumption 1, we know that

$$R_\rho=\sum_{l=1}^N\rho_l\cdot\left[\begin{array}{c}R^{l1}\cdots R^{lN}\end{array}\right]>0,\ R_\rho+\bar{R}_\rho=\sum_{l=1}^N R_\rho\cdot\left[\begin{array}{c}R^{l1}+\bar{R}^{l1}\cdots R^{lN}+\bar{R}^{lN}\end{array}\right]>0.$$

Therefore, $P_{i-1}-P_i>0$ and $W_{i-1}-W_i>0$, which means that both $\{P_i\}_{i=1}^\infty$ and $\{W_i\}_{i=1}^\infty$ are monotone decreasing sequences. Letting P^* and W^* denote the solution of GARE (10), after computations for (26) subtracting (10), one can obtain that $P_{i-1}>P^*$ and $W_{i-1}>W^*$. Thus sequences $\{P_i\}_{i=1}^\infty$ and $\{W_i\}_{i=1}^\infty$ are lower bounded; i.e., $\lim_{i\to\infty}P_i$ and $\lim_{i\to\infty}W_i$ exist. Letting $i\to\infty$ in (23), and eliminating K_∞ and $\bar{K}_\infty$ by (21) and (22), we have

$$\begin{cases}
P_\infty A+A'P_\infty+C'P_\infty C+Q_\rho-\mathcal{M}_\infty\mathcal{H}_\infty^{-1}\mathcal{M}'_\infty=0,\\
W_\infty(A+\bar{A})+(A+\bar{A})'W_\infty+(C+\bar{C})'P_\infty(C+\bar{C})+Q_\rho+\bar{Q}_\rho-\widetilde{\mathcal{M}}_\infty\widetilde{\mathcal{H}}_\infty^{-1}\widetilde{\mathcal{M}}'_\infty=0,
\end{cases}$$

where

$$\mathcal{M}_\infty = P_\infty B + C' P_\infty D,\ \mathcal{H}_\infty = D' P_\infty D + R_\rho,$$
$$\widetilde{\mathcal{M}}_\infty = W_\infty (B + \bar{B}) + (C + \bar{C})' P_\infty (D + \bar{D}),$$
$$\widetilde{\mathcal{H}}_\infty = (D + \bar{D})' P_\infty (D + \bar{D}) + R_\rho + \bar{R}_\rho.$$

Since the solution of GAREs is unique (see Theorem 1 in [8]), The proof is completed by setting $P_\infty = P^*, W_\infty = W^*$.

3.3 Online implementation of the adaptive optimal control algorithm

In this subsection, the implementation of the iteration schemes (20)–(22) is discussed based on $\mathcal{H}$-representation, where the information regarding the system matrices A and $\bar{A}$ is embedded to be observed online for policy evaluation (20). The $\mathcal{H}$-representation technique is a fundamental tool initially introduced in [22]. It utilizes a vectorized operator $\overrightarrow{X} = (X^1, X^2, \ldots, X^n)'$, where X^i represents the i-th row of an $n \times n$ symmetric matrix X. This technique allows us to express an $n^2 \times 1$ vector $\overrightarrow{X}$ as $\overrightarrow{X} = H\widetilde{X}$, where H is an $n^2 \times \frac{n(n+1)}{2}$ matrix with full column rank, and $\widetilde{X}$ is an $\mathcal{R}^{\frac{n(n+1)}{2} \times 1}$ vector. In the following, we select two bases as

$$\{E_{ij} : 1 \leqslant i \leqslant j \leqslant n\} = \{E_{11}, E_{12}, \ldots, E_{1n}, E_{22}, \ldots, E_{2n}, \ldots, E_{nn}\},$$

and

$$\{F_{ij} : 1 \leqslant i \leqslant j \leqslant n\} = \{F_{11}, F_{12}, \ldots, F_{1n}, F_{22}, \ldots, F_{2n}, \ldots, F_{nn}\},$$

where $E_{ij} = (e_{lk})_{n\times n}$ with $e_{ij} = e_{ji} = 1$ and the other entries being zero, $F_{ij} = (e_{lk})_{n\times n}$ with $f_{ij} = f_{ji} = \frac{1}{2}$ and the other entries being zero. For the above given bases, we define the following two operators.

Definition 3. Define $\psi : X_n = (x_{ij})_{n\times n} \in \mathcal{S}_n \mapsto \widetilde{X}_n$, $\varphi : Y_n = (x_{ij})_{n\times n} \in \mathcal{S}_n \mapsto \breve{Y}_n$, where

$$\widetilde{X}_n = (x_{11}, \ldots, x_{1n}, x_{22}, \ldots, x_{2n}, \ldots, x_{n-1,n-1}, x_{n-1,n}, x_{nn})',$$
$$\breve{Y}_n = (y_{11}, 2y_{12}, \ldots, 2y_{1n}, y_{22}, 2y_{23} \ldots, 2y_{2n}, \ldots, y_{n-1,n-1}, 2y_{n-1,n}, y_{nn})'.$$

Set $\nu = \frac{n(n+1)}{2}$. For clarity, we define the H matrix in $\mathcal{H}$-representation corresponding to operator ψ by H_ν^1, while $H_\nu^{\frac{1}{2}}$ refers to the $\mathcal{H}$-representation matrix corresponding to operator φ. Obviously, $(H_\nu^1)' H_\nu^{\frac{1}{2}} = I_\nu$. Define $X_{P_i}^\tau = \mathbb{E}^{\mathcal{F}_t}[(x_\tau - \mathbb{E}^{\mathcal{F}_t} x_\tau)(x_\tau - \mathbb{E}^{\mathcal{F}_t} x_\tau)']$ and $X_{W_i}^\tau = (\mathbb{E}^{\mathcal{F}_t} x_\tau)(\mathbb{E}^{\mathcal{F}_t} x_\tau)'$, then

$$\breve{X}_{P_i} = ((H_\nu^{\frac{1}{2}})' H_\nu^{\frac{1}{2}})^{-1} (H_\nu^{\frac{1}{2}})' \overrightarrow{X}_{P_i},\ \breve{X}_{W_i} = ((H_\nu^{\frac{1}{2}})' H_\nu^{\frac{1}{2}})^{-1} (H_\nu^{\frac{1}{2}})' \overrightarrow{X}_{W_i},$$
$$\widetilde{P}_i = ((H_\nu^1)' H_\nu^1)^{-1} (H_\nu^1)' \overrightarrow{P}_i,\ \widetilde{W}_i = ((H_\nu^1)' H_\nu^1)^{-1} (H_\nu^1)' \overrightarrow{W}_i.$$

Using $\mathcal{H}$-representation technique and vectorized operator $\overrightarrow{\cdot}$, the policy evaluation (20) is written as

$$\widetilde{W}_i'(\breve{X}_{W_i}^t - \breve{X}_{W_i}^{t+T}) - \widetilde{P}_i' \breve{X}_{P_i}^{t+T} = \int_t^{t+T} \widetilde{\mathcal{Q}}' \breve{X}_{P_i}^\tau \mathrm{d}\tau + \int_t^{t+T} \widetilde{\bar{\mathcal{Q}}}' \breve{X}_{W_i}^\tau \mathrm{d}\tau \tag{27}$$

with $\mathcal{Q} = Q_\rho + K_i' R_\rho K_i$ and $\bar{\mathcal{Q}} = Q_\rho + \bar{Q}_\rho + \bar{K}_i'(R_\rho + \bar{R}_\rho)\bar{K}_i$. Since $\widetilde{P}_i \in \mathcal{R}^\nu$ and $\widetilde{W}_i \in \mathcal{R}^\nu$, there are $n(n+1)$ unknown parameters in $\widetilde{P}_i$ and $\widetilde{W}_i$. In the least-square sense, Eq. (27) must be computed at $N \geqslant \nu$ points at each $[t_j, t_j + \Delta t_j]$ over time intervals $[0, T]$ with $j = 1, 2, \ldots, N$. Let

$$\Delta X_{P_i}^j = -\mathbb{E}^{\mathcal{F}_{t_j}}[(x_{t_j+\Delta t_j} - \mathbb{E}^{\mathcal{F}_{t_j}} x_{t_j+\Delta t_j})(x_{t_j+\Delta t_j} - \mathbb{E}^{\mathcal{F}_{t_j}} x_{t_j+\Delta t_j})'],$$
$$\Delta X_{W_i}^j = x_{t_j}(x_{t_j})' - (\mathbb{E}^{\mathcal{F}_{t_j}} x_{t_j+\Delta t_j})(\mathbb{E}^{\mathcal{F}_{t_j}} x_{t_j+\Delta t_j})'.$$

After calculations with $\overrightarrow{X}_{P_i} \to \breve{X}_{P_i}$ or $\overrightarrow{X}_{W_i} \to \breve{X}_{W_i}$ based on $\mathcal{H}$-representation technique, we get $\Delta\breve{X}_{P_i}$ and $\Delta\breve{X}_{W_i}$. The solution of the unknown parameters $\widetilde{P}_i$ and $\widetilde{W}_i$ are obtained as

$$\begin{bmatrix} \widetilde{P}_i \\ \widetilde{W}_i \end{bmatrix} = \mathcal{X}_{P_i,W_i}^{-1} (\mathcal{Y}_{P_i} + \mathcal{Y}_{W_i}), \tag{28}$$

where

$$\mathcal{X}_{P_i,W_i}=\begin{bmatrix}(\Delta\breve{X}^1_{P_i})' & (\Delta\breve{X}^1_{W_i})'\\ (\Delta\breve{X}^2_{P_i})' & (\Delta\breve{X}^2_{W_i})'\\ \vdots & \vdots\\ (\Delta\breve{X}^N_{P_i})' & (\Delta\breve{X}^N_{W_i})'\end{bmatrix},\ \mathcal{Y}_{P_i}=\begin{bmatrix}\left(\int_{t_1}^{t_1+\Delta t_1}\widetilde{\mathcal{Q}}'\breve{X}^\tau_{P_i}\mathrm{d}\tau\right)^1\\ \vdots\\ \left(\int_{t_N}^{t_N+\Delta t_N}\widetilde{\mathcal{Q}}'\breve{X}^\tau_{P_i}\mathrm{d}\tau\right)^N\end{bmatrix},$$

$$\mathcal{Y}_{W_i}=\begin{bmatrix}\left(\int_{t_1}^{t_1+\Delta t_1}\widetilde{\bar{\mathcal{Q}}}'\breve{X}^\tau_{W_i}\mathrm{d}\tau\right)^1\\ \vdots\\ \left(\int_{t_N}^{t_N+\Delta t_N}\widetilde{\bar{\mathcal{Q}}}'\breve{X}^\tau_{W_i}\mathrm{d}\tau\right)^N\end{bmatrix}.$$

Remark 4. Since $\widetilde{P}_i\in\mathcal{R}^\nu$ and $\widetilde{W}_i\in\mathcal{R}^\nu$, there are $n(n+1)$ unknown parameters in $\widetilde{P}_i$ and $\widetilde{W}_i$. The $\mathcal{H}$-representation technique helps to transform the two linear $n\times n$ symmetric matrix equations into two $\frac{n(n+1)}{2}$-dimensional vector equations, respectively, which can reduce computational complexity and transform the matrix equations into vector equations.

Remark 5. In (27), if we choose $\overrightarrow{X}_{P_i}=H^1_\nu\widetilde{X}_{P_i}$ and $\overrightarrow{X}_{W_i}=H^1_\nu\widetilde{X}_{W_i}$, we can only use a standard basis H^1_ν to eliminate the repeated elements in X_{P_i}, X_{W_i}, P_i and W_i. The solutions of $\widetilde{P}_i$ and $\widetilde{W}_i$ can be given as

$$\begin{bmatrix}\widetilde{P}_i\\ \widetilde{W}_i\end{bmatrix}=\bar{\mathcal{X}}^{-1}_{P_i,W_i}(\bar{\mathcal{Y}}_{P_i}+\bar{\mathcal{Y}}_{W_i}),\tag{29}$$

where

$$\bar{\mathcal{X}}_{P_i,W_i}=\begin{bmatrix}\bar{\mathcal{X}}^{11}_{P_i,W_i} & \bar{\mathcal{X}}^{12}_{P_i,W_i}\\ \vdots & \vdots\\ \bar{\mathcal{X}}^{N1}_{P_i,W_i} & \bar{\mathcal{X}}^{N2}_{P_i,W_i}\end{bmatrix},\ \bar{\mathcal{Y}}_{P_i}=\begin{bmatrix}\int_{t_1}^{t_1+\Delta t_1}\widetilde{\mathcal{Q}}'(H^1_\nu)'H^1_\nu\widetilde{X}^\tau_{P_i}\mathrm{d}\tau\\ \vdots\\ \int_{t_N}^{t_N+\Delta t_N}\widetilde{\mathcal{Q}}'(H^1_\nu)'H^1_\nu\widetilde{X}^\tau_{P_i}\mathrm{d}\tau\end{bmatrix}',$$

$$\bar{\mathcal{Y}}_{W_i}=\begin{bmatrix}\int_{t_1}^{t_1+\Delta t_1}\widetilde{\bar{\mathcal{Q}}}'(H^1_\nu)'H^1_\nu\widetilde{X}^\tau_{P_i}\mathrm{d}\tau\\ \vdots\\ \int_{t_N}^{t_N+\Delta t_N}\widetilde{\bar{\mathcal{Q}}}'(H^1_\nu)'H^1_\nu\widetilde{X}^\tau_{P_i}\mathrm{d}\tau\end{bmatrix}'$$

with

$$\bar{\mathcal{X}}^{11}_{P_i,W_i}=(-\widetilde{X}^{t_1+\Delta t_1}_{P_i}(H^1_\nu)'H^1_\nu)',\ \bar{\mathcal{X}}^{12}_{P_i,W_i}=((\widetilde{X}^{t_1}_{W_i}-\widetilde{X}^{t_1+\Delta t_1}_{W_i})(H^1_\nu)'H^1_\nu)',$$
$$\bar{\mathcal{X}}^{N1}_{P_i,W_i}=(-\widetilde{X}^{t_N+\Delta t_N}_{P_i}(H^1_\nu)'H^1_\nu)',\ \bar{\mathcal{X}}^{N2}_{P_i,W_i}=((\widetilde{X}^{t_N}_{W_i}-\widetilde{X}^{t_N+\Delta t_N}_{W_i})(H^1_\nu)'H^1_\nu)'.$$

Remark 6. In the new policy evaluation, either (28) or (29), based on $\mathcal{H}$-representation, the persistence condition can be described as follows: for given constants $\beta_1>0$, $\beta_2>0$, and $T>0$, it holds that

$$\beta_1 I\leqslant\int_t^{t+T}(\widetilde{X}^\tau_{P_i}-\widetilde{X}^{\tau+T}_{P_i})(\widetilde{X}^\tau_{P_i}-\widetilde{X}^{\tau+T}_{P_i})'+(\widetilde{X}^\tau_{W_i}-\widetilde{X}^{\tau+T}_{W_i})(\widetilde{X}^\tau_{W_i}-\widetilde{X}^{\tau+T}_{W_i})'\mathrm{d}\tau\leqslant\beta_2 I$$

for all values of t.

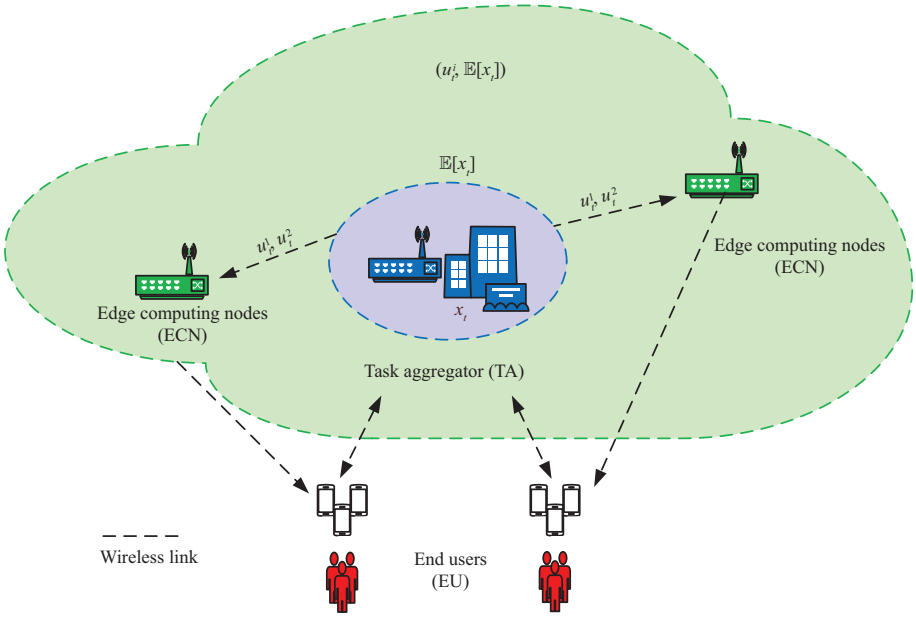

Figure 1 (Color online) Computation offloading with two modes in the MECN.

An online PI algorithm based on $\mathcal{H}$-representation technique for solving the Pareto optimal control problem is given in Algorithm 1.

Algorithm 1. Online PI algorithm.

Step 1. Initialization: set the iteration number $i = 1$ and start with a stabilizing control policy $u_t^i = K^i x_t + (\bar{K}^i - K^i)\mathbb{E}x_t$.

Step 2. Online solve for P^i and W^i using the policy evaluation cost via least-square (28) or (29).

Step 3. Update the control policy u_t^{i+1} by (21) and (22).

Step 4. Stop the iteration if $|P^{i+1} - P^i| \leqslant \varepsilon$ and $|W^{i+1} - W^i| \leqslant \varepsilon$ for a small positive value ε. Otherwise, set $i = i + 1$ and go to Step 1.

4 Example

In this section, we give two examples with dimensions of 1 and 4 to verify the effectiveness of our result by applying the PI algorithm designed in Algorithm 1. Example 1 is a practical illustration aimed at showcasing the validity of our algorithm in the presence of stochastic noise in the system. Three sets of experiments, each with different initial states and initial feedback gains K^0 and $\bar{K}^0$, are conducted to learn the state trajectories and determine the optimal control gains K^* and $\bar{K}^*$. Example 2 makes a comparison between our method and a model-based approach.

Example 1. Consider an improved optimal computation offloading control problem in multi-access edge computing networks (MECNs) [27]. Since the finite number of computing nodes in MECNs and the fact that the computing capability of each node can influence the workload for computation tasks, the utilization of a mean-field term is employed for problem formulation. In this example, a cell in an MECN consists of one task aggregator (TA) and two sets of edge computing nodes (ECNs), as shown in Figure 1. The updated network state dynamics for computation offloading in a mean-field game setting are expressed as follows:

$$\mathrm{d}x_t = \left[r^0 x_t + \bar{r}^0 \mathbb{E}x_t + \sum_{l=1}^{2}(r^l u_t^l + \bar{r}^l \mathbb{E}u_t^l)\right]\mathrm{d}t + \left[\sigma^0 x_t + \bar{\sigma}^0 \mathbb{E}x_t + \sum_{l=1}^{2}(\sigma^l u_t^l + \bar{\sigma}^l \mathbb{E}u_t^l)\right]\mathrm{d}w_t, \tag{30}$$

where the network state x_t denotes the amount of aggregate computation task that needs to be offloaded by the TA to the ECNs. The network state dynamics $\mathrm{d}x_t/\mathrm{d}t$ refers to the evolution of the state of the TA. Let the admissible computation offloading controllers u_t^l, $l = 1, 2$ denote the portion of the aggregate task offloaded by ECN l from the TA at time t. Each ECN cooperates to minimize its cost by controlling

Table 1 Parameters of the MECNs in system (30) and cost function (31)

	$l=0$	$l=1$	$l=2$		$l=0$	$l=1$	$l=2$		$l=0$	$l=1$	$l=2$
r^l	-1.0199	-0.8840	0.0904	$\bar{\rho}^l$	$\times$	0.4000	0.1500	$\bar{\tau}_1^l$	$\times$	2.0104	0.1610
$\bar{r}^l$	-0.2683	-0.5039	-0.8454	τ_1^l	$\times$	0.3300	2.4003	$\bar{\tau}_2^l$	$\times$	0.0405	0.0010
σ^l	1.1804	0.8490	-1.0158	τ_2^l	$\times$	0.0420	1.1030	$\bar{e}_1^l$	$\times$	0.2080	0.1111
$\bar{\sigma}^l$	-0.5474	0.9710	-0.7736	e_1^l	$\times$	0.0096	0.0880	$\bar{e}_2^l$	$\times$	0.0414	0.1021
ρ^l	$\times$	2.0000	0.7500	e_2^l	$\times$	0.0161	1.0603				

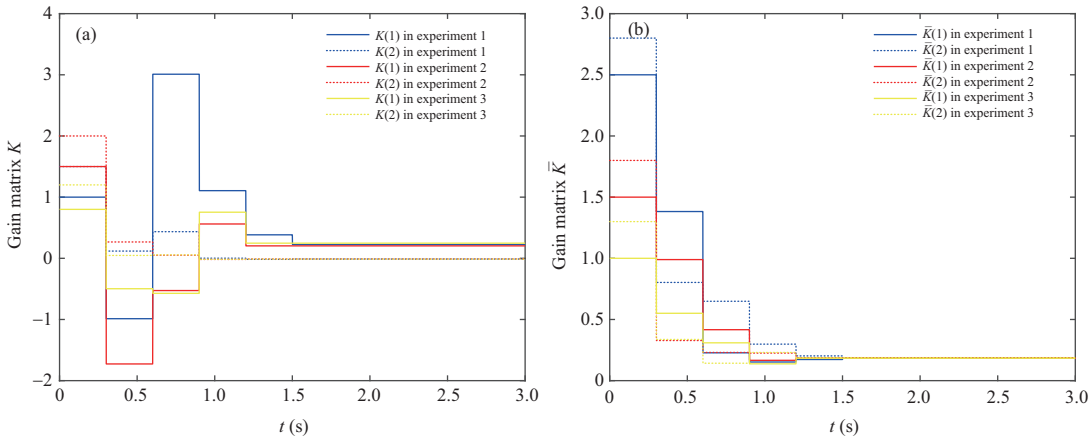

Figure 2 (Color online) Evolution of the control gain (a) K and (b) $\bar{K}$.

its own offloading strategy. In engineering practice, one often takes

$$J^l(u^1,u^2,x_0)=\mathbb{E}\int_0^\infty\left\{\rho^l x_t^2+\bar{\rho}^l\mathbb{E}x_t^2+\sum_{i=1}^2[(\tau_i^l+e_i^l)(u_i)_t^2+(\bar{\tau}_i^l+\bar{e}_i^l)\mathbb{E}(u_i)_t^2]\right\}\mathrm{d}t,\quad l=1,2. \tag{31}$$

The physical meaning of the parameters in dynamic equation (30) and cost functions of (31) can be referred to [27]. All parameters are chosen as in Table 1.

For $\rho=0.3$, we get the weighted-sum objective function

$$\begin{aligned}J_\rho=\mathbb{E}\int_0^\infty&\left[1.0635x_t^2-0.0431(\mathbb{E}x_t)^2+\begin{bmatrix}u_t^1 & u_t^2\end{bmatrix}\begin{bmatrix}2.0586 & 0\\ 0 & 1.7423\end{bmatrix}\begin{bmatrix}u_t^1\\ u_t^2\end{bmatrix}\right.\\&\left.+\begin{bmatrix}\mathbb{E}u_t^1 & \mathbb{E}u_t^2\end{bmatrix}\begin{bmatrix}0.6614 & 0\\ 0 & 0.0989\end{bmatrix}\begin{bmatrix}\mathbb{E}u_t^1\\ \mathbb{E}u_t^2\end{bmatrix}\right]\mathrm{d}t.\end{aligned}$$

Based on Algorithm 1, the sampling interval of state information is 0.001 s. Three experiments are implemented by setting different initial states and stabilizing control gains. Three different initial variables are chosen as $(x_0)^1=2$, $(x_0)^2=3$ and $(x_0)^3=4$. The initial stabilizing control gains are chosen as $(K)^1=[1\ 1.5]$, $(\bar{K})^1=[2.5\ 2.8]$, $(K)^2=[0.8\ 1.2]$, $(\bar{K})^2=[1\ 1.3]$, $(K)^3=[1.5\ 2]$, $(\bar{K})^3=[1.5\ 1.8]$. The convergences of control gains K and $\bar{K}$ in different experiments are shown in Figures 2(a) and (b), where K arrives at optimal value at $(K)^{1*}=[0.1851\ -0.0104]$, $(K)^{2*}=[0.1726\ -0.0098]$, $(K)^{3*}=[0.1719\ -0.0098]$, and $\bar{K}$ at $(\bar{K})^{1*}=[0.1864\ 0.1809]$, $(\bar{K})^{2*}=[0.1877\ 0.1798]$, $(\bar{K})^{3*}=[0.1877\ 0.1797]$ at time 1.5 s. The initial value $P(0)$ and $W(0)$ are chosen to be 0. Using Algorithm 1, we obtain P and W online which are shown in Figures 3(a) and (b), and find the learned optimal value $P^{1*}=0.4853, P^{2*}=0.4519, P^{3*}=0.4501$ and $W^{1*}=0.4015, W^{2*}=0.4016, W^{3*}=0.4014$. While the solutions that are obtained by directly solving the algebraic Riccati equation (10) are $P^*=0.9228$ and $W^*=0.4392$. Figure 3(b) shows the convergence of P and W to their optimal values in 10^{-2} accuracy at time $t=1.5$ s. To illustrate that the final iterative control gain satisfies ASMS, the state curves under the obtained control gain are depicted in Figure 4(a). From the simulation results, it is obvious that the system states tend to zero at $t=3$ s, which indicates that the system is mean-square stabilizable. Figure 4(b) is the system state in the second experiment which is used to show that we have adopted the Monte Carlo experiment in each sampled interval $[t_j,t_j+\Delta t_j]$ ($t_j=0.3$ s). The expectation $\mathbb{E}^{\mathcal{F}_{t_j}}[\cdot]$ is obtained by

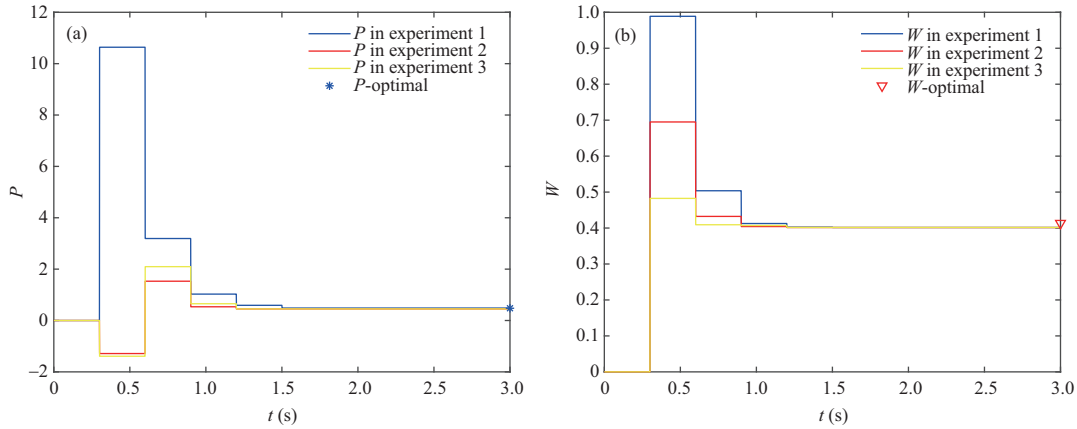

Figure 3 (Color online) Evolution of (a) P and (b) W.

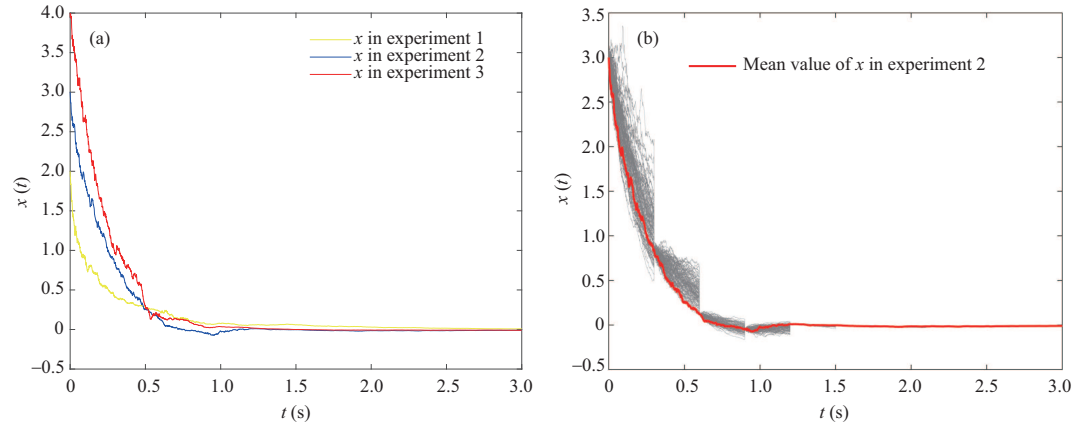

Figure 4 (Color online) (a) Trajectory of the state variables during the simulation; (b) trajectory of the state variables in experiment 2 with initial state $x_0 = 3$.

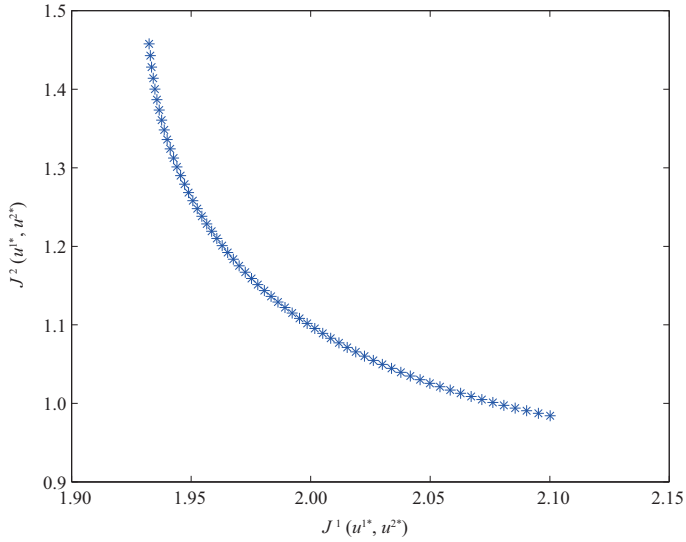

Figure 5 (Color online) Pareto solutions using the learned parameters P^* and W^*.

calculating the mean value of M sample paths, that all start from the same initial state x_{t_j} in each interval. In addition, letting the weighted-sum parameter ρ vary in $[0, 1]$, we obtain all Pareto solutions according to Lemma 6 by using the learned parameters P^* and W^*, i.e., the whole Pareto frontier, which is shown in Figure 5.

Example 2. In this example, we consider the system (1) with $N = 2, n = 4, m_1 = m_2 = 2$. The coefficients in system (1) are given as

$$A = \begin{bmatrix} -0.1689 & -0.9555 & 1.9846 & 0.5897 \\ 0.8321 & 0.8016 & 1.4732 & -0.0663 \\ -2.2296 & 0.4935 & -0.4115 & -0.2851 \\ -0.6265 & 2.8890 & -2.7135 & 1.6935 \end{bmatrix}, \bar{A} = \begin{bmatrix} -3.0529 & 0.4068 & 1.6493 & 1.3409 \\ -0.5029 & 1.0017 & 0.4665 & 2.5729 \\ -2.0731 & -0.2981 & 0.9305 & 0.7523 \\ -1.2731 & -0.8347 & 1.1100 & 0.7080 \end{bmatrix},$$

$$C = \begin{bmatrix} -0.4225 & -1.0813 & -0.7491 & -0.1229 \\ -0.4123 & 0.6064 & -1.4851 & 0.2258 \\ 1.8622 & -0.4776 & 1.1905 & -1.6586 \\ 1.4065 & -0.3794 & -1.9303 & 0.8444 \end{bmatrix}, \bar{C} = \begin{bmatrix} 0.1953 & 0.6438 & -0.1070 & 0.3069 \\ -0.1812 & -1.8825 & 1.1141 & -1.2781 \\ -0.5152 & 0.4614 & 1.0383 & -0.0710 \\ -0.2958 & -1.4320 & -0.8909 & 1.2713 \end{bmatrix},$$

$$B^1 = \begin{bmatrix} 0.3691 & 0.2448 \\ 0.5968 & -2.2757 \\ -0.0200 & 1.2097 \\ 0.0086 & 0.7000 \end{bmatrix}, B^2 = \begin{bmatrix} 0.3136 & 0.6027 \\ 0.8221 & 1.1538 \\ -0.5459 & 0.5757 \\ -0.9668 & -0.2977 \end{bmatrix}, D^1 = \begin{bmatrix} -1.8636 & 1.7800 \\ 1.7430 & 0.6499 \\ 0.7267 & 0.9024 \\ -1.3273 & 0.3036 \end{bmatrix},$$

$$D^2 = \begin{bmatrix} -0.0442 & -0.1975 \\ 1.1640 & -0.0459 \\ -1.1468 & -1.0010 \\ -1.7076 & 0.2269 \end{bmatrix}, \bar{B}^1 = \begin{bmatrix} -1.5261 & 0.5294 \\ 0.9229 & -0.3809 \\ 0.5890 & 0.4791 \\ 1.3635 & 0.2310 \end{bmatrix}, \bar{B}^2 = \begin{bmatrix} -1.4222 & -0.2706 \\ -0.3939 & -0.9341 \\ -0.6285 & 0.2448 \\ -0.5118 & 0.3309 \end{bmatrix},$$

$$\bar{D}^1 = \begin{bmatrix} 0.6985 & -0.4665 \\ 0.9654 & -0.1408 \\ 1.3141 & -0.4145 \\ -0.1657 & -0.0853 \end{bmatrix}, \bar{D}^2 = \begin{bmatrix} 1.0180 & -0.9200 \\ -0.0886 & 1.9582 \\ -0.0027 & -0.4527 \\ -1.4240 & -2.2467 \end{bmatrix},$$

$$R_{11} = \begin{bmatrix} 3.3717 & -0.3033 \\ -0.3033 & 5.9656 \end{bmatrix}, R_{12} = \begin{bmatrix} 2.0240 & 0.5210 \\ 0.5210 & 2.9568 \end{bmatrix}, R_{21} = \begin{bmatrix} 5.7230 & 1.0468 \\ 1.0468 & 0.8585 \end{bmatrix}, R_{22} = \begin{bmatrix} 3.9549 & 0.7870 \\ 0.7870 & 1.8030 \end{bmatrix},$$

$$\bar{R}_{11} = \begin{bmatrix} 7.6088 & 5.2130 \\ 5.2130 & 3.5978 \end{bmatrix}, \bar{R}_{12} = \begin{bmatrix} 0.9714 & 0.4549 \\ 0.4549 & 0.3259 \end{bmatrix}, \bar{R}_{21} = \begin{bmatrix} 6.0424 & -1.3582 \\ -1.3582 & 4.6730 \end{bmatrix}, \bar{R}_{22} = \begin{bmatrix} 0.9251 & 0.6773 \\ 0.6773 & 1.8981 \end{bmatrix}.$$

The other matrices in cost functions $J^1(x_0; u^1, u^2)$ and $J^2(x_0; u^1, u^2)$ are $Q^1 = 0.0599$, $Q^2 = 1.4966$, $\bar{Q}^1 = 0.5427$, and $\bar{Q}^2 = 0.2081$. The weighted-sum variable is taken as $\rho = 0.3$. Now, we adopt the same method as described in Section III-B in [25] to compare our method with the model-based approach. Similarly, the estimations of $\hat{A}$ of A and $\hat{\bar{A}}$ of $\bar{A}$ should be obtained, which are used to solve GAREs (10) with the estimation matrices by the generalized semidefinite programming (SDP) method given in [28]. The estimation procedure is given as follows:

(1) Select $K = -(D'D)^{-1}D'C$ and $\bar{K} = -[(D + \bar{D})'(D + \bar{D})]^{-1}(D + \bar{D})'(C + \bar{D})$.

(2) Read the data $\{\mathbb{E}x_k\}_{k=0}^{n^2}$ and $\{x_k - \mathbb{E}x_k\}_{k=0}^{n^2}$ at time $t_k = \frac{k}{n^2}$, $k = 0, 1, 2, \ldots, n^2$, respectively.

(3) Define $\bar{y}_k = (\mathbb{E}x_{k+1} - \mathbb{E}x_k)/(t_{k+1} - t_k)$, $y_k = (x_{k+1} - \mathbb{E}x_{k+1} - x_k + \mathbb{E}x_k)/(t_{k+1} - t_k)$. Note that $\bar{Y} = (\bar{y}_0, \ldots, \bar{y}_{n^2-1})$, $Y = (y_0, \ldots, y_{n^2-1})$, $\bar{X} = (\mathbb{E}x_0, \ldots, \mathbb{E}x_{n^2-1})$, and $X = (x_0 - \mathbb{E}x_0, \ldots, x_{n^2-1} - \mathbb{E}x_{n^2-1})$.

(4) Estimate $\hat{A} = YX'(XX')^{-1} - BK$ and $\hat{\bar{A}} = \bar{Y}\bar{X}'(\bar{X}\bar{X}')^{-1} - (B + \bar{B})\bar{K} - \hat{A}$.

After several iterations, we obtain

$$\hat{A} = \begin{bmatrix} -0.1855 & -0.9550 & 2.0027 & 0.5970 \\ 0.8235 & 0.8074 & 1.4792 & -0.0660 \\ -2.2363 & 0.4935 & -0.3982 & -0.2828 \\ -0.6570 & 2.9004 & -2.7028 & 1.6817 \end{bmatrix}, \hat{\bar{A}} = \begin{bmatrix} -2.5339 & -0.9028 & 2.7393 & -1.8365 \\ 0.9867 & 8.1649 & 4.3827 & -4.1414 \\ -1.6550 & 2.0944 & 2.2579 & -1.7475 \\ -1.3086 & -2.5040 & -0.4206 & 0.3496 \end{bmatrix}.$$

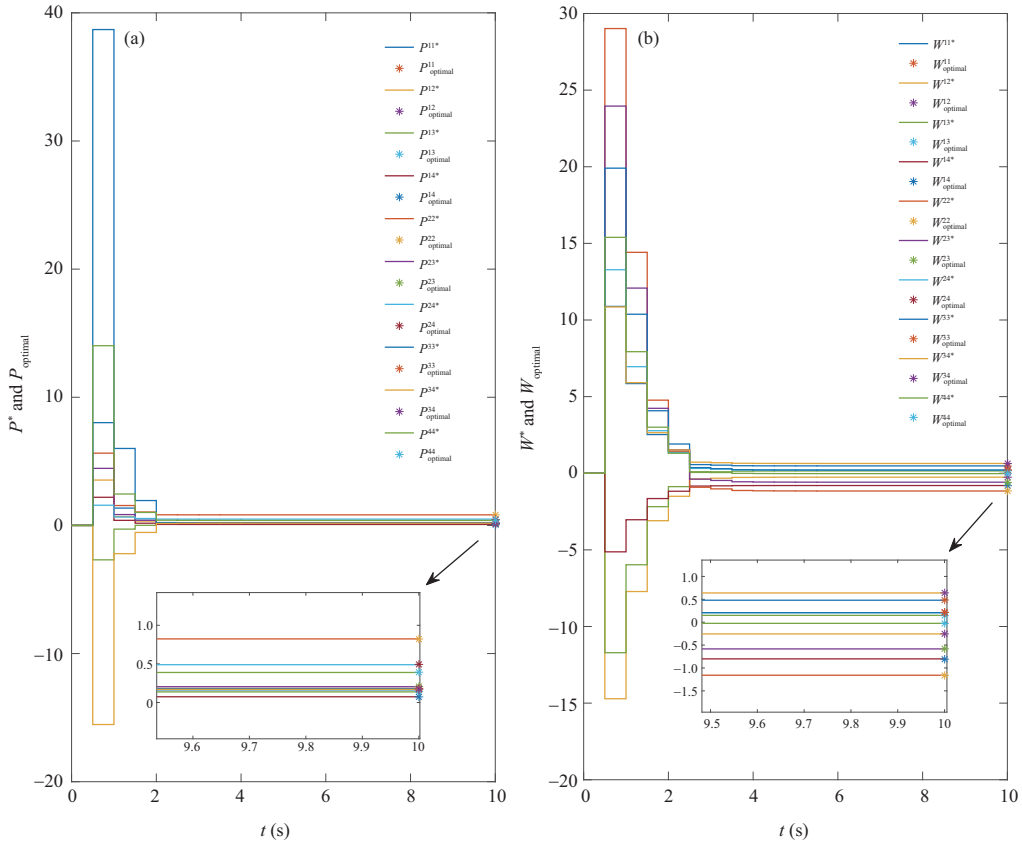

Figure 6 (Color online) Evolution of (a) P^* and (b) W^*.

Computing $\hat{P}^*$ and $\hat{W}^*$ in GAREs (10) by SDP based on the estimated $\hat{A}$ and $\hat{\bar{A}}$, there are

$$\hat{P}^* = \begin{bmatrix} 0.1651 & 0.1803 & 0.1371 & 0.0717 \\ 0.1803 & 0.8438 & 0.2032 & 0.4980 \\ 0.1371 & 0.2032 & 0.1808 & 0.1707 \\ 0.0717 & 0.4980 & 0.1707 & 0.3937 \end{bmatrix}, \quad \hat{W}^* = \begin{bmatrix} -0.0396 & -0.4780 & 0.6402 & -1.3231 \\ -0.4780 & -0.1593 & -0.1468 & -0.6295 \\ 0.6402 & -0.1468 & -0.3567 & 1.0675 \\ -1.3231 & -0.6295 & 1.0675 & -1.3391 \end{bmatrix}.$$

Then, according to the following two expressions:

$$\begin{aligned} \mathcal{R}(\hat{P}^*) =& \hat{P}^*A + A'\hat{P}^* + C'\hat{P}^*C + \rho Q^1 + (1-\rho)Q^2 - (\hat{P}^*B + C'\hat{P}^*D)(D'\hat{P}^*D + R)^{-1}(\hat{P}^*B + C'\hat{P}^*D)', \\ \mathcal{M}(\hat{W}^*) =& \hat{W}^*(A+\bar{A}) + (A+\bar{A})'\hat{W}^* + (C+\bar{C})'\hat{P}^*(C+\bar{C}) + \rho Q^1 + (1-\rho)Q^2 + \rho\bar{Q}^1 + (1-\rho)\bar{Q}^2 \\ & - [\hat{W}^*(B+\bar{B}) + (C+\bar{C})'\hat{P}^*(D+\bar{D})][(D+\bar{D})'\hat{P}^*(D+\bar{D}) + R]^{-1} \\ & \cdot [\hat{W}^*(B+\bar{B}) + (C+\bar{C})'\hat{P}^*(D+\bar{D})]', \end{aligned}$$

we get $\|\mathcal{R}(\hat{P}^*)\| = 0.0499$ and $\|\mathcal{M}(\hat{W}^*)\| = 1.4857$. By Algorithm 1, we get

$$P^* = \begin{bmatrix} 0.1638 & 0.1821 & 0.1375 & 0.0745 \\ 0.1821 & 0.8240 & 0.2046 & 0.4897 \\ 0.1375 & 0.2045 & 0.1800 & 0.1708 \\ 0.0744 & 0.4897 & 0.1708 & 0.3890 \end{bmatrix}, \quad W^* = \begin{bmatrix} 0.4802 & -0.2560 & 0.1496 & -0.8011 \\ -0.2563 & -1.1581 & -0.5834 & 0.2075 \\ 0.1494 & -0.5835 & 0.2086 & 0.6393 \\ -0.8011 & 0.2073 & 0.6392 & -0.0250 \end{bmatrix}$$

with $\|\mathcal{R}(P^*)\| = 0.0030$ and $\|\mathcal{M}(W^*)\| = 0.0032$. From the comparisons between $\|\mathcal{R}(P^*)\|$ and $\|\mathcal{R}(\hat{P}^*)\|$, $\|\mathcal{M}(W^*)\|$ and $\|\mathcal{M}(\hat{W}^*)\|$, which measure the distance from the experimental value to the real solution of GAREs, we can see that the algorithm obtained in this paper performs better. Moreover, the evolutions

of P^* and W^* solved by Algorithm 1 converge to their optimal values

$$P_{\text{optimal}} = \begin{bmatrix} 0.1637 & 0.1820 & 0.1375 & 0.0745 \\ 0.1820 & 0.8240 & 0.2045 & 0.4896 \\ 0.1375 & 0.2045 & 0.1800 & 0.1708 \\ 0.0745 & 0.4896 & 0.1708 & 0.3890 \end{bmatrix}, \ W_{\text{optimal}} = \begin{bmatrix} 0.4803 & -0.2563 & 0.1495 & -0.8012 \\ -0.2563 & -1.1584 & -0.5835 & 0.2075 \\ 0.1495 & -0.5835 & 0.2086 & 0.6392 \\ -0.8012 & 0.2075 & 0.6392 & -0.0250 \end{bmatrix},$$

which are shown in Figures 6(a) and (b).

5 Conclusion

In this study, a novel online PI algorithm was proposed for the Pareto optimal control of MFSSs with multi-player and multi-objective. It has been shown that coupled GAREs can be solved without knowing complete system information. An extensive convergence analysis of the algorithm was conducted. In addition, a practical algorithm based on the $\mathcal{H}$-representation technique was proposed. The simulation results demonstrated the performance of the PI method. Future efforts should be made to investigate a more general model-free method for Pareto optimal control problems of different types of systems, such as discrete-time stochastic systems [29–32] and stochastic switching systems [33].

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62103442, 12326343, 62373229), Natural Science Foundation of Shandong Province (Grant No. ZR2021QF080), Fundamental Research Funds for the Central Universities (Grant No. 23CX06024A), and Outstanding Youth Innovation Team in Shandong Higher Education Institutions (Grant No. 2023KJ061).

References

1 Mu C, Wang K, Ni Z, et al. Cooperative differential game-based optimal control and its application to power systems. IEEE Trans Ind Inf, 2020, 16: 5169–5179
2 Dockner E J, Jorgensen S, Long N V, et al. Differential Games in Economics and Management Science. Cambridge: Cambridge University Press, 2000
3 Sun Q, Wang X, Yang G, et al. Optimal constraint following for fuzzy mechanical systems based on a time-varying β-measure and cooperative game theory. IEEE Trans Syst Man Cybern Syst, 2022, 52: 7574–7587
4 Engwerda J. The regular convex cooperative linear quadratic control problem. Automatica, 2008, 44: 2453–2457
5 Lin Y, Jiang X, Zhang W. Necessary and sufficient conditions for Pareto optimality of the stochastic systems in finite horizon. Automatica, 2018, 94: 341–348
6 Zhang W, Peng C. Indefinite mean-field stochastic cooperative linear-quadratic dynamic difference game with its application to the network security model. IEEE Trans Cybern, 2022, 52: 11805–11818
7 Jiang X, Su S F, Zhao D. Pareto optimal strategy under H^∞ constraint for the mean-field stochastic systems in infinite horizon. IEEE Trans Cybern, 2023, 53: 6963–6976
8 Qi Q, Zhang H, Wu Z. Stabilization control for linear continuous-time mean-field systems. IEEE Trans Autom Control, 2019, 64: 3461–3468
9 Zhang T, Deng F, Shi P. Nonfragile finite-time stabilization for discrete mean-field stochastic systems. IEEE Trans Autom Control, 2023, 68: 6423–6430
10 Lin Y, Zhang T, Zhang W. Pareto-based guaranteed cost control of the uncertain mean-field stochastic systems in infinite horizon. Automatica, 2018, 92: 197–209
11 Lin Y, Zhang W. Pareto efficiency in the infinite horizon mean-field type cooperative stochastic differential game. J Franklin Inst, 2021, 358: 5532–5551
12 Wang T, Zhang H, Luo Y. Stochastic linear quadratic optimal control for model-free discrete-time systems based on Q-learning algorithm. Neurocomputing, 2018, 312: 1–8
13 Liu M, Wan Y, Lewis F L, et al. Adaptive optimal control for stochastic multiplayer differential games using on-policy and off-policy reinforcement learning. IEEE Trans Neural Netw Learn Syst, 2020, 31: 5522–5533
14 Bian T, Jiang Z P. Stochastic and adaptive optimal control of uncertain interconnected systems: a data-driven approach. Syst Control Lett, 2018, 115: 48–54
15 Howard R A. Dynamic Programming and Markov Processes. Cambridge: MIT Press, 1960
16 Barto A G, Sutton R S, Anderson C W. Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans Syst Man Cybern, 1983, SMC-13: 834–846
17 Vrabie D, Pastravanu O, Abu-Khalaf M, et al. Adaptive optimal control for continuous-time linear systems based on policy iteration. Automatica, 2009, 45: 477–484
18 Kiumarsi B, Vamvoudakis K G, Modares H, et al. Optimal and autonomous control using reinforcement learning: a survey. IEEE Trans Neural Netw Learn Syst, 2018, 29: 2042–2062
19 Pang B, Bian T, Jiang Z P. Robust policy iteration for continuous-time linear quadratic regulation. IEEE Trans Autom Control, 2022, 67: 504–511
20 Li N, Li X, Peng J, et al. Stochastic linear quadratic optimal control problem: a reinforcement learning method. IEEE Trans Autom Control, 2022, 67: 5009–5016
21 Pang B, Jiang Z P. Reinforcement learning for adaptive optimal stationary control of linear stochastic systems. IEEE Trans Autom Control, 2022, 68: 2383–2390
22 Zhang W H, Chen B S. $\mathcal{H}$-representation and applications to generalized Lyapunov equations and linear stochastic systems. IEEE Trans Autom Control, 2012, 57: 3009–3022

23 Leitmann G. Cooperative and Noncooperative Many Player Differential Games. Berlin: Springer-Verlag, 1974
24 Engwerda J C. LQ Dynamic Optimization and Differential Games. Chichester: Wiley, 2005
25 Li N, Li X, Yu Z. Indefinite mean-field type linear-quadratic stochastic optimal control problems. Automatica, 2020, 122: 109267
26 Øksendal B. Stochastic Differential Equations: An Introduction with Applications. New York: Springer, 2013
27 Banez R A, Tembine H, Li L, et al. Mean-field-type game-based computation offloading in multi-access edge computing networks. IEEE Trans Wireless Commun, 2020, 19: 8366–8381
28 Rami M A, Xun Yu, Zhou M A. Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls. IEEE Trans Autom Control, 2000, 45: 1131–1143
29 Hu Z, Shi P, Zhang J, et al. Control of discrete-time stochastic systems with packet loss by event-triggered approach. IEEE Trans Syst Man Cybern Syst, 2021, 51: 755–764
30 Qi W, Yang X, Park J H, et al. Fuzzy SMC for quantized nonlinear stochastic switching systems with semi-Markovian process and application. IEEE Trans Cybern, 2022, 52: 9316–9325
31 Jiang X S, Tian S P, Zhang W H. pth moment exponential stability of general nonlinear discrete-time stochastic systems. Sci China Inf Sci, 2021, 64: 209204
32 Zhang T L, Xu S Y, Zhang W H. Predefined-time stabilization for nonlinear stochastic Itô systems. Sci China Inf Sci, 2023, 66: 182202
33 Qi W, Zhang N, Zong G, et al. Asynchronous sliding-mode control for discrete-time networked hidden stochastic jump systems with cyber attacks. IEEE Trans Cybern, 2024, 54: 1934–1946