2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

A Model-Free Iteration Algorithm for Markov Jump Linear Systems Based on Gauss-Seidel Method

Wenwu Fan[1] and Junlin Xiong[1]

Abstract— **This paper focuses on the linear quadratic regulator problem of discrete-time Markov jump linear systems without knowing the system matrices. A model-free fixed-point iteration algorithm is proposed to learn the optimal state feedback control law without the requirement of an initial admissible control policy. Analogous to the Gauss-Seidel method for linear equations, the model-free algorithm is constantly iterating with the latest information of each mode. It is proved that the algorithm converges monotonically to the optimal solution. In addition, our algorithm is faster than the classical model-based value iteration method. Finally, an example is used to illustrate our results.**

I. Introduction

Markov jump linear systems (MJLSs), containing abrupt changes in their dynamics, are a class of significant stochastic models. The linear quadratic regulator (LQR) problem of MJLSs has been studied since the 1960s [1]. The optimal solution is associated with the coupled algebraic Riccati equations (CARE) [2], which are identical in form to those for a deterministic system except for modal coupling. The controllability and observability of MJLSs are defined in [3], and then a sufficient condition for the existence of a mean-square stabilizing solution to the CARE is presented in [4]. Model-based value iteration (VI) [5] and policy iteration (PI) [6] algorithms are proposed to compute the optimal solution for the LQR problem of discrete-time MJLSs. However, these methods for obtaining the optimal control law require full knowledge of the system dynamics.

Reinforcement learning (RL) is a method of learning optimal actions through interaction with the environment to maximize cumulative expected return [7], [8]. RL has been well applied in control theory, such as state feedback problem [8], [9], output feedback problem [10], optimal tracking problem [11], and H_∞ control problem [12]–[14], to obtain desired control policy without knowing the system matrices. Based on policy iteration, the integral RL method is used for model-free LQR control [15] and optimal tracking control [16] of continuous-time MJLSs. The natural gradient method [17] and policy iteration [18] are applied for the model-free optimal control of discrete-time MJLSs. However, these model-free methods for MJLSs require an initial admissible control policy, which largely relies on the system dynamics.

This paper aims to solve the model-free LQR optimal control problem of discrete-time MJLSs without the requirement of an initial admissible control policy. Unlike deterministic discrete-time linear systems, the optimal solution to the LQR problem of discrete-time MJLSs involves the coupling of multiple modes. The model-free coupled equations for LQR control are established using the system states and control inputs. Then a model-free fixed-point iteration algorithm is proposed to obtain the optimal control law of discrete-time MJLSs, which does not require an initial admissible controller. Similar to the Gauss-Seidel method for solving linear equations, any information obtained is immediately used to iterate for the next mode. We prove that our algorithm converges monotonically to the optimal solution and is at least as fast as the classical model-based value iteration method [5]. Finally, a simulation example demonstrates the effectiveness and good performance of our algorithm.

Notation: In this article, $\mathbb{R}^n$ is the n-dimensional real Euclidean space and $\mathbb{B}\left(\mathbb{R}^n, \mathbb{R}^m\right)$ is the normed bounded linear space of all $m \times n$ real matrices, with $\mathbb{B}\left(\mathbb{R}^n\right) \triangleq \mathbb{B}\left(\mathbb{R}^n, \mathbb{R}^n\right)$. Set $\mathbb{H}^{n,m}$ as the linear space made up of all s-sequences of real matrices $V = (V_1, \ldots, V_s)$ with $V_i \in \mathbb{B}\left(\mathbb{R}^n, \mathbb{R}^m\right), i = 1, \ldots, s$, and, for simplicity, set $\mathbb{H}^n \triangleq \mathbb{H}^{n,n}$. For $X = (X_1, \ldots, X_s)$, $Y = (Y_1, \ldots, Y_s) \in \mathbb{H}^n$, $X > 0$ means X_i, for each i, is a symmetric positive definite matrix; $X \geq 0$ means X_i, for each i, is a symmetric positive semi-definite matrix; $X = 0$ means, for each i, every element in X_i equals 0; $X \geq Y$ means $X_i - Y_i \geq 0$ for each i. Set $\mathbb{H}^{n+} \triangleq \{X | X \in \mathbb{H}^n,\ X \geq 0\}$. The superscript $\top$ indicates the transpose of a matrix and $\mathbb{E}(\cdot)$ denotes the mathematical expectation.

II. Problem Formulation and Preliminaries

Consider a discrete-time MJLS in an appropriate probabilistic space $(\Omega, P, \{\mathfrak{F}_k\}, \mathfrak{F})$:

$$x(k+1) = A_{\theta(k)} x(k) + B_{\theta(k)} u(k), \tag{1}$$

where $x(k) \in \mathbb{R}^n$ is the system state, and $u(k) \in \mathbb{R}^m$ is the control input. $\theta(k)$ is a Markov chain taking values in a finite state space $\mathcal{S} = \{1, 2, \ldots, s\}$ with transition probability matrix $P = [p_{ij}]$. Define $A = (A_1, A_2, \ldots, A_s) \in \mathbb{H}^n$ and $B = (B_1, B_2, \ldots, B_s) \in \mathbb{H}^{m,n}$. In this paper, $\theta(k)$ is observed, and the transition probability matrix P is known. The system matrices A_i and B_i are all unknown for $\forall i \in \mathcal{S}$.

Next, we give some definitions of mean square stability.

Definition 1: The system (1) with $u(k) \equiv 0$ is mean square stable if $\lim\limits_{k \to +\infty} \mathbb{E}\left(\|x(k)\|^2\right) = 0$ for any initial condition $x(0)$ and $\theta(0)$.

This work was financially supported by the National Natural Science Foundation of China under Grant 62273320.

[1] Wenwu Fan and Junlin Xiong are with the Department of Automation, University of Science and Technology of China, Hefei 230026, China. E-mail: fanwenwu@mail.ustc.edu.cn (Wenwu Fan), junlin.xiong@gmail.com (Junlin Xiong)

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

Definition 2: [4] $F = (F_1, \ldots, F_N) \in \mathbb{H}^{n,m}$ stabilizes (A, B) in the mean square sense if the system (1) with $u(k) = F_{\theta(k)}x(k)$ is mean square stable.

Definition 3: The control $u(k)$ is said to be admissible if the system (1) with $u(k)$ is mean square stable.

Definition 4: (A, B) is mean square stabilizable if there exists $F = (F_1, \ldots, F_N) \in \mathbb{H}^{n,m}$ such that F stabilizes (A, B) in the mean square sense.

The following assumptions hold throughout this paper.

Assumption 1: (A, B) is mean square stabilizable.

Assumption 2: (A_i, B_i) is controllable for each $i \in \mathcal{S}$.

Define the infinite horizon quadratic cost function as

$$\begin{aligned} &J\left(x(0), \theta(0), u\right) \\ \triangleq&\mathbb{E}\left\{\sum_{k=0}^{\infty}\left(x(k)^\top Q_{\theta(k)}x(k) + u(k)^\top R_{\theta(k)}u(k)\right)\right\}, \end{aligned} \tag{2}$$

where $Q = (Q_1, Q_2, \ldots, Q_s) \in \mathbb{H}^n$, $Q > 0$, $R = (R_1, R_2, \ldots, R_s) \in \mathbb{H}^m$, $R > 0$, and $u = (u(0), u(1), \ldots)$ is the control input sequence. The objective of LQR control is to find the optimal state feedback control law, which minimizes the cost (2).

For $X = (X_1, X_2, \ldots, X_s) \in \mathbb{H}^n$, define the following operator $\mathscr{E}(\cdot) = (\mathscr{E}_1(\cdot), \ldots, \mathscr{E}_s(\cdot)) \in \mathbb{B}\left(\mathbb{H}^n\right)$ as

$$\mathscr{E}_i(X) \triangleq \sum_{j=1}^{s} p_{ij}X_j,\ i \in \mathcal{S}. \tag{3}$$

The coupled algebraic Riccati equations (CARE) for discrete-time MJLSs are given as

$$\begin{aligned} X_i =&A_i^\top \mathscr{E}_i(X)A_i + Q_i - A_i^\top \mathscr{E}_i(X)B_i \\ &\times\left(R_i + B_i^\top \mathscr{E}_i(X)B_i\right)^{-1} B_i^\top \mathscr{E}_i(X)A_i,\ i \in \mathcal{S}. \end{aligned} \tag{4}$$

The definition of the mean square stabilizing solution for the CARE (4) is given in the following.

Definition 5: [4] $X = (X_1, X_2, \ldots, X_s) \in \mathbb{H}^n$ is a mean square stabilizing solution for the CARE (4) if X satisfies CARE (4) and $F = (F_1, \ldots, F_N) \in \mathbb{H}^{n,m}$ stabilizes (A, B) in the mean square sense with $F_i = -\left(R_i + B_i^\top \mathscr{E}_i(X)B_i\right)^{-1} B_i^\top \mathscr{E}_i(X)A_i$.

Lemma 1: [4] Under Assumption 1, there exists a unique mean square stabilizing solution $X^* = (X_1^*, X_2^*, \ldots, X_s^*)$ for the CARE (4). Moreover, $X^* > 0$.

The following lemma gives the optimal control law and the optimal cost for the LQR problem of discrete-time MJLSs.

Lemma 2: [2] The optimal control law for the LQR problem is given by

$$u^*(k) = F_{\theta(k)}^* x(k), \tag{5}$$

where $F^* = (F_1^*, \ldots, F_s^*) \in \mathbb{H}^{n,m}$ is

$$F_i^* = -\left(R_i + B_i^\top \mathscr{E}_i(X^*)B_i\right)^{-1} B_i^\top \mathscr{E}_i(X^*)A_i, \tag{6}$$

and the optimal cost is

$$J^*\left(x(0), \theta(0)\right) = \mathbb{E}\left\{x(0)^\top X_{\theta(0)}^* x(0)\right\}. \tag{7}$$

This paper focuses on solving the model-free LQR control problem directly from the system states and control inputs.

III. Model-Free Optimal Control

In this section, a model-free fixed-point iteration algorithm is proposed to obtain the optimal control law for the LQR problem of discrete-time MJLSs. The algorithm does not require an initial admissible control policy and is constantly iterating with the latest information similar to the Gauss-Seidel method. We prove that our algorithm converges monotonically to the mean square stabilizing solution of the CARE (4) and is at least as fast as the classical model-based value iteration algorithm [5].

A. A Model-free Algorithm Based on Gauss-Seidel Method

In this subsection, the model-free equations for LQR control are established using the system states and control inputs of discrete-time MJLSs. Then a model-free fixed-point iteration algorithm based on the Gauss-Seidel method is proposed to solve the LQR problem without the requirement of an initial admissible control policy.

For $H^* = (H_1^*, H_2^*, \ldots, H_s^*) \in \mathbb{H}^l$, where $l = m + n$, define H_i^*, $i \in \mathcal{S}$, as

$$\begin{aligned} H_i^* \triangleq& \begin{bmatrix} H_{i,xx}^* & H_{i,xu}^* \\ H_{i,ux}^* & H_{i,uu}^* \end{bmatrix} \\ \triangleq& \begin{bmatrix} Q_i + A_i^\top \mathscr{E}_i(X^*)A_i & A_i^\top \mathscr{E}_i(X^*)B_i \\ B_i^\top \mathscr{E}_i(X^*)A_i & R_i + B_i^\top \mathscr{E}_i(X^*)B_i \end{bmatrix}. \end{aligned} \tag{8}$$

For each $i \in \mathcal{S}$, H_i^* can be expressed as

$$H_i^* = \begin{bmatrix} Q_i & 0 \\ 0 & R_i \end{bmatrix} + \begin{bmatrix} A_i & B_i \end{bmatrix}^\top \mathscr{E}_i(X^*)\begin{bmatrix} A_i & B_i \end{bmatrix}. \tag{9}$$

Then for any $x(k)$ and $u(k)$ when $\theta(k) = i$, $i \in \mathcal{S}$, we obtain

$$\begin{aligned} \begin{bmatrix} x(k) \\ u(k) \end{bmatrix}^\top H_i^* \begin{bmatrix} x(k) \\ u(k) \end{bmatrix} =&x(k)^\top Q_i x(k) + u(k)^\top R_i u(k) \\ &+ x(k+1)^\top \mathscr{E}_i(X^*)x(k+1). \end{aligned} \tag{10}$$

where $x(k+1) = A_i x(k) + B_i u(k)$.

The optimal control gain matrices can be expressed as

$$F_i^* = -(H_{i,uu}^*)^{-1} H_{i,ux}^*,\ i \in \mathcal{S}. \tag{11}$$

According to CARE (4), equations (8) and (11), the relationship between X^* and H^* also satisfies

$$X_i^* = \begin{bmatrix} I \\ F_i^* \end{bmatrix}^\top H_i^* \begin{bmatrix} I \\ F_i^* \end{bmatrix},\ i \in \mathcal{S}. \tag{12}$$

Let $u(k) = F_{\theta(k)}x(k) + e(k)$, where $e(k)$ is probing noise. Define

$$z(k) \triangleq \begin{bmatrix} x(k)^\top & u(k)^\top \end{bmatrix}^\top,$$

and

$$r_{\theta(k)}(x(k), u(k)) \triangleq x(k)^\top Q_{\theta(k)}x(k) + u(k)^\top R_{\theta(k)}u(k).$$

Equation (10) becomes

$$\begin{aligned} z(k)^\top H_i^* z(k) =&r_i(x(k), u(k)) \\ &+ x(k+1)^\top \mathscr{E}_i(X^*)x(k+1). \end{aligned} \tag{13}$$

For a symmetric matrix $H \in \mathbb{B}\left(\mathbb{R}^l\right)$, define $\operatorname{svec}(H) = [h_{11}, \sqrt{2}h_{12}, \cdots, \sqrt{2}h_{1l}, h_{22}, \sqrt{2}h_{23}, \cdots, \sqrt{2}h_{2l}, \cdots, h_{ll}]^\top$ with h_{ij} being an entry. For a vector x, define $\tilde{x} = \operatorname{svec}(xx^\top)$. Then equation (13) can be parameterized as

$$\begin{aligned} \tilde{z}(k)^\top \operatorname{svec}\left(H_i^*\right) =& r_i(x(k), u(k)) \\ &+ \tilde{x}(k+1)^\top \operatorname{svec}\left(\mathscr{E}_i(X^*)\right). \end{aligned} \tag{14}$$

For each $i \in \mathcal{S}$, define Ξ_i as a matrix stacking by all $\tilde{z}(k)^\top$, Ω_i as a matrix stacking by all $r_i(x(k), u(k))$, and Φ_i as a matrix stacking by all $\tilde{x}(k+1)^\top$ when $\theta(k) = i$ from $k = 0$ to $k = N - 1$. It follows from equation (14) that

$$\Xi_i \operatorname{svec}(H_i^*) = \Omega_i + \Phi_i \operatorname{svec}\left(\mathscr{E}_i(X^*)\right). \tag{15}$$

Remark 1: Unlike the model-free optimal control problem of discrete-time linear systems, the data matrices stacking needs to distinguish the operation mode when generating the system state of MJLSs. This is because the optimal control gain matrices are different under different modes.

The over-determined linear equations (15) can be solved by the least squares method. The following assumption is required to ensure that equation (15) has a unique solution.

Assumption 3: $\operatorname{rank}\left(\Xi_i\right) = \dfrac{l(l+1)}{2}, \forall i \in \mathcal{S}$.

Remark 2: Assumption 3 is a common rank condition for model-free control [10], [13]. The probing noise added into the control input ensures the data set is linearly independent. Thus, Assumption 3 holds.

Under Assumption 3, equation (15) can be solved as

$$\operatorname{svec}(H_i^*) = \left(\Xi_i^\top \Xi_i\right)^{-1} \Xi_i^\top \left(\Omega_i + \Phi_i \operatorname{svec}\left(\mathscr{E}_i(X^*)\right)\right). \tag{16}$$

Assumption 3 guarantees that H^* in (16) is the unique solution of (15). Thus, the model-free equation (16) is the same as the model-based equation (9).

The detailed model-free algorithm for solving the LQR problem of discrete-time MJLSs is proposed in Algorithm 1.

Remark 3: System (1) with $u(k) = F_{\theta(k)}^\eta x(k) + e(k)$ is probably unstable. In Algorithm 1, the system is restarted when $\|x(k)\| \geq c$ to continue collecting data to prevent the system state from being too large.

It can be seen that Algorithm 1 does not require an initial admissible control policy. The convergence of Algorithm 1 will be proved in III-B, which shows our algorithm can obtain the optimal state feedback control law. Similar to the Gauss-Seidel method for solving linear equations, the information of each mode is applied to the next iteration immediately after the update in Algorithm 1. This brings advantages in convergence rate to our algorithm, which is shown in III-C.

B. Convergence of Algorithm 1

In this subsection, we prove that Algorithm 1 converges monotonically to the mean square stabilizing solution of the CARE (4).

First, we give the following lemmas.

Algorithm 1 Model-Free Fixed-point Iteration

```
Input: $X^0 = \left(X_1^0, X_2^0, \ldots, X_s^0\right) = 0$; $\bar{X} = X^0$; $F^0 = \left(F_1^0, F_2^0 \ldots, F_N^0\right) = 0$; tolerable convergence error $\epsilon$; a
    large positive constant $c$; data length $N$; the transition
    probability matrix $P$.
Output: $X^\eta$, $H^\eta$, $F^\eta$.
 1: for $\eta = 0, 1, 2, \ldots$ do
 2:     Let the control input $u(k) = F_{\theta(k)}^\eta x(k) + e(k)$. Collect
        data for obtaining $\Xi_i$, $\Omega_i$, and $\Phi_i$, $i \in \mathcal{S}$. If $\|x(k)\| \geq$
        $c$, restart the system to continue collecting data until
        data length is $N$.
 3:     if $\operatorname{rank}\left(\Xi_i\right) < \frac{l(l+1)}{2}$ for some $i \in \mathcal{S}$ then
 4:         go to step 2
 5:     end if
 6:     for $i = 1, 2, \ldots, s$ do
 7:         $\operatorname{svec}(H_i^\eta) = (\Xi_i^\top \Xi_i)^{-1} \Xi_i^\top \left(\Omega_i + \Phi_i \operatorname{svec}(\mathscr{E}_i(\bar{X}))\right)$,
 8:         $F_i^\eta = -(H_{i,uu}^\eta)^{-1} H_{i,ux}^\eta$,
 9:         $X_i^{\eta+1} = \begin{bmatrix} I \\ F_i^\eta \end{bmatrix}^\top H_i^\eta \begin{bmatrix} I \\ F_i^\eta \end{bmatrix}$,
10:         $\bar{X}_i = X_i^{\eta+1}$,
11:     end for
12:     if $\left\|X^{\eta+1} - X^\eta\right\| < \epsilon$ then
13:         break
14:     end if
15: end for
```

Lemma 3: Define the following operators: $\mathscr{E}_i^1(X) = \sum_{j=1}^{i-1} p_{ij} X_j$ and $\mathscr{E}_i^2(X) = \sum_{j=i}^{s} p_{ij} X_j$, $i \in \mathcal{S}$. Then the iteration of X^η in Algorithm 1 is the same as

$$\begin{aligned} X_i^{\eta+1} =& A_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) A_i + Q_i \\ &- A_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) B_i \\ &\times \left(R_i + B_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) B_i\right)^{-1} \\ &\times B_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) A_i, \; i \in \mathcal{S}. \end{aligned} \tag{17}$$

Proof: See Appendix V-A. ■

Lemma 4: The iteration equation (17) is the same as

$$\begin{aligned} X_i^{\eta+1} =& \left(A_i + BF_i^\eta\right)^\top \left(\mathscr{E}_i^1(X^{\eta+1}) + \mathscr{E}_i^2(X^\eta)\right) \\ &\times \left(A_i + BF_i^\eta\right) + Q_i + \left(F_i^\eta\right)^\top R_i F_i^\eta, \; i \in \mathcal{S}, \end{aligned} \tag{18}$$

where

$$\begin{aligned} F_i^\eta =& -\left(R_i + B_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) B_i\right)^{-1} \\ &\times B_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) A_i. \end{aligned} \tag{19}$$

Proof: Substituting (19) into (18) yields (17). This completes the proof. ■

Lemma 5: Let X^η and F^η, $\eta = 0, 1, \ldots$, satisfy (18) and (19). Then $X^\eta \leq X^{\eta+1} \leq X^*$.

Proof: See Appendix V-B. ■

Algorithm 1 is convergent as shown in the following theorem.

Theorem 1: Let X^η, H^η, and F^η, $\eta = 0, 1, \ldots$, be generated in Algorithm 1. Then

(1) $X^\eta \leq X^{\eta+1} \leq X^*$;

(2) $\lim\limits_{\eta\rightarrow+\infty} X^{\eta} = X^{*}$, $\lim\limits_{\eta\rightarrow+\infty} H^{\eta} = H^{*}$, and $\lim\limits_{\eta\rightarrow+\infty} F^{\eta} = F^{*}$.

Proof: According to Lemma 3-5, it is straightforward to show that $X^{\eta} \leq X^{\eta+1} \leq X^{*}$. Because X^{η}, $\eta = 0, 1, \ldots$, is monotonous and upper-bounded, $\lim\limits_{\eta\rightarrow+\infty} X^{\eta} = X^{\infty}$ exists. It implies that $\lim\limits_{\eta\rightarrow+\infty} H^{\eta} = H^{\infty}$ and $\lim\limits_{\eta\rightarrow+\infty} F^{\eta} = F^{\infty}$ exist.

For each $i \in \mathcal{S}$, we have

$$\begin{aligned} F_i^{\infty} =& - \left(H_{i,uu}^{\infty}\right)^{-1} H_{i,ux}^{\infty}, \\ X_i^{\infty} =& \begin{bmatrix} I \\ F_i^{\infty} \end{bmatrix}^{\top} H_i^{\infty} \begin{bmatrix} I \\ F_i^{\infty} \end{bmatrix}, \\ H_i^{\infty} =& \begin{bmatrix} Q_i & 0 \\ 0 & R_i \end{bmatrix} + \begin{bmatrix} A_i & B_i \end{bmatrix}^{\top} \mathscr{E}_i(X^{\infty}) \begin{bmatrix} A_i & B_i \end{bmatrix}. \end{aligned}$$

Combining the above equations, we can obtain that X^{∞} satisfies the CARE (4) and $X^{\infty} \geq 0$. Because CARE (4) has a unique solution X^{*} in $\mathbb{H}^{n+}$ [4], one has $X^{\infty} = X^{*}$. Then $H^{\infty} = H^{*}$ and $F^{\infty} = F^{*}$. ■

Theorem 1 shows that Algorithm 1 converges monotonically to the optimal solution of the LQR problem for discrete-time MJLSs. Thus, Algorithm 1 can be used to learn the optimal state feedback control policy without knowing the system matrices. Moreover, it can be used to obtain a stabilizing control policy for discrete-time MJLSs.

C. Compared with Value Iteration

In this subsection, we compare Algorithm 1 with a classical model-based value iteration algorithm proposed in [5].

The model-based value iteration is given in the following lemma.

Lemma 6: [5] Let $\widetilde{X}^{0} = \left(\widetilde{X}_1^{0}, \widetilde{X}_2^{0}, \ldots, \widetilde{X}_s^{0}\right) = 0$. For $\eta = 0, 1, 2, \ldots$, and $i \in \mathcal{S}$,

$$\begin{aligned} \widetilde{X}_i^{\eta+1} =& \left(A_i + B\widetilde{F}_i^{\eta}\right)^{\top} \mathscr{E}(\widetilde{X}^{\eta}) \left(A_i + B\widetilde{F}_i^{\eta}\right) \\ & + Q_i + \left(\widetilde{F}_i^{\eta}\right)^{\top} R_i \widetilde{F}_i^{\eta},\ i \in \mathcal{S}, \end{aligned} \tag{20}$$

where

$$\widetilde{F}_i^{\eta} = -\left(R_i + B_i^{\top}\mathscr{E}(\widetilde{X}^{\eta})B_i\right)^{-1} B_i^{\top}\mathscr{E}(\widetilde{X}^{\eta})A_i. \tag{21}$$

Then

(1) $\widetilde{X}^{\eta} \leq \widetilde{X}^{\eta+1} \leq X^{*}$;

(2) $\lim\limits_{\eta\rightarrow+\infty} \widetilde{X}^{\eta} = X^{*}$ and $\lim\limits_{\eta\rightarrow+\infty} \widetilde{F}^{\eta} = F^{*}$.

The model-based value iteration algorithm has the same convergence as Algorithm 1. However, all modes are updated simultaneously before proceeding to the next iteration in the VI algorithm. The following theorem will demonstrate the superiority of Algorithm 1 over the model-based value iteration algorithm.

Theorem 2: Let X^{η}, $\eta = 0, 1, \ldots$, be generated in Algorithm 1, and $\widetilde{X}^{\eta}$, $\eta = 0, 1, \ldots$, be generated in Lemma 6. Then $X^{\eta} \geq \widetilde{X}^{\eta}$ for all η.

Proof: From Lemma 6, for any $i \in \mathcal{S}$, we have $\widetilde{F}_i^{0} = 0$. It follows that

$$\widetilde{X}_i^{1} = Q_i > 0,\ i \in \mathcal{S}.$$

According to Lemma 3, one has that $X_1^1 = Q_1$ and

$$\begin{aligned} X_i^1 =& \left(A_i + BF_i^0\right)^{\top} \left(\mathscr{E}_i^1(X^1) + \mathscr{E}_i^2(X^0)\right) \left(A_i + BF_i^0\right) \\ & + Q_i + \left(F_i^0\right)^{\top} R_i F_i^0 \geq Q_i,\ i = 2, 3, \ldots, s \end{aligned}$$

Thus, we obtain $X^1 \geq \widetilde{X}^1 > 0$.

Assume that $X^{\eta} \geq \widetilde{X}^{\eta} > 0$. According to (20) and (21), $\widetilde{X}_i$, $i \in \mathcal{S}$, can be expressed as

$$\begin{aligned} \widetilde{X}_i^{\eta+1} =& A_i^{\top}\mathscr{E}_i(\widetilde{X}^{\eta})A_i + Q_i - A_i^{\top}\mathscr{E}_i(\widetilde{X}^{\eta})B_i \\ & \times \left(R_i + B_i^{\top}\mathscr{E}_i(\widetilde{X}^{\eta})B_i\right)^{-1} B_i^{\top}\mathscr{E}_i(\widetilde{X}^{\eta})A_i. \end{aligned} \tag{22}$$

Using Woodbury matrix equality, $X_i^{\eta+1}$ and $\widetilde{X}_i^{\eta+1}$, $i \in \mathcal{S}$, also can be written as

$$\begin{aligned} X_i^{\eta+1} =& A_i^{\top} \left(\left(\mathscr{E}_i^1(X^{\eta+1}) + \mathscr{E}_i^2(X^{\eta})\right)^{-1} + B_i R_i^{-1} B_i^{\top} \right)^{-1} \\ & \times A_i + Q_i, \\ \widetilde{X}_i^{\eta+1} =& A_i^{\top} \left(\left(\mathscr{E}_i(\widetilde{X}^{\eta})\right)^{-1} + B_i R_i^{-1} B_i^{\top} \right)^{-1} A_i + Q_i. \end{aligned}$$

According to Theorem 1, we obtain $X^{\eta+1} \geq X^{\eta}$. Under the assumption $X^{\eta} \geq \widetilde{X}^{\eta} > 0$, one has

$$\begin{aligned} \mathscr{E}_i^1(X^{\eta+1}) + \mathscr{E}_i^2(X^{\eta}) =& \sum_{j=1}^{i-1} p_{ij} X_j^{\eta+1} + \sum_{j=i}^{s} p_{ij} X_j^{\eta} \\ \geq& \sum_{j=1}^{i-1} p_{ij} X_j^{\eta} + \sum_{j=i}^{s} p_{ij} X_j^{\eta} \\ \geq& \mathscr{E}_i(\widetilde{X}^{\eta}) > 0. \end{aligned}$$

It follows that

$$0 < \left(\mathscr{E}_i^1(X^{\eta+1}) + \mathscr{E}_i^2(X^{\eta})\right)^{-1} \leq \left(\mathscr{E}_i(\widetilde{X}^{\eta})\right)^{-1}$$

Then we have

$$\begin{aligned} & \left(\left(\mathscr{E}_i^1(X^{\eta+1}) + \mathscr{E}_i^2(X^{\eta})\right)^{-1} + B_i R_i^{-1} B_i^{\top} \right)^{-1} \\ \geq & \left(\left(\mathscr{E}_i(\widetilde{X}^{\eta})\right)^{-1} + B_i R_i^{-1} B_i^{\top} \right)^{-1} \end{aligned}$$

It is straightforward to show that

$$X_i^{\eta+1} \geq \widetilde{X}_i^{\eta+1} > 0.$$

Therefore, $X^{\eta} \geq \widetilde{X}^{\eta}$ for all η. ■

According to Theorem 1 and Lemma 6, both Algorithm 1 and the model-based VI algorithm converge monotonically to the optimal solution. Theorem 2 shows that X^{η} in Algorithm 1 is always closer to the optimal solution than VI. Thus, Algorithm 1 is at least as fast as the model-based value iteration algorithm.

IV. Illustrative Example

Consider a classical discrete-time MJLS [5] with three modes. The system matrices are provided as $A_1 = \begin{bmatrix} 0 & 1 \\ -2.5 & 3.2 \end{bmatrix}$, $A_2 = \begin{bmatrix} 0 & 1 \\ -4.3 & 4.5 \end{bmatrix}$, $A_3 = \begin{bmatrix} 0 & 1 \\ 5.3 & -5.2 \end{bmatrix}$, and $B_1 = B_2 = B_3 = \begin{bmatrix} 0 & 1 \end{bmatrix}^{\top}$. The discrete state transition

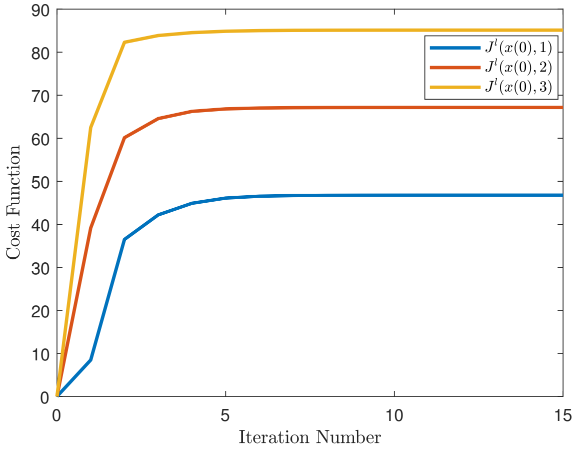

Fig. 1. Cost $J^\eta(x(0),\theta(0))$ vs. iteration number η.

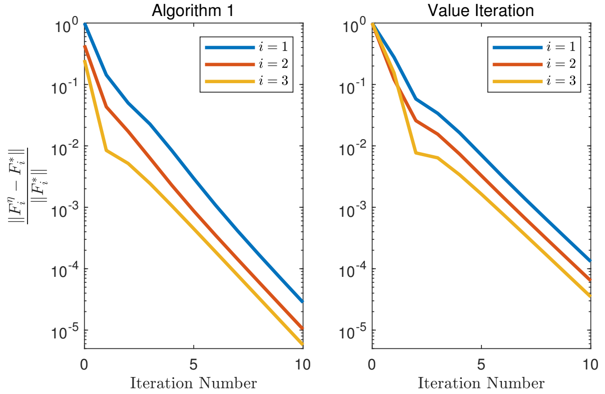

Fig. 2. Relative error $\dfrac{\|F_i^\eta - F_i^*\|}{\|F_i^*\|}$ vs. iteration number η.

probability matrix is $P = \begin{bmatrix} 0.67 & 0.17 & 0.16 \\ 0.3 & 0.47 & 0.23 \\ 0.26 & 0.1 & 0.64 \end{bmatrix}$. The weight matrices of quadratic cost are given as $Q_1 = \begin{bmatrix} 3.6 & -3.8 \\ -3.8 & 4.87 \end{bmatrix}$, $Q_2 = \begin{bmatrix} 10 & -3 \\ -3 & 8 \end{bmatrix}$, $Q_3 = \begin{bmatrix} 5 & -4.5 \\ -4.5 & 4.5 \end{bmatrix}$, and $R_1 = 2.6$, $R_2 = 1.165$, $R_3 = 1.111$.

The initial state x_0 is a random vector with standard normal distribution. The data length is set as $N = 100$. The probing noise is given by $e_k = \operatorname{randn}(m,1)$, where $\operatorname{randn}(m,1)$ is an $m \times 1$ matrix with standard normal distribution.

Define the cost function in the iterative process as $J^\eta(x(0),\theta(0)) = \mathbb{E}\left\{x(0)^\top X^\eta_{\theta(0)} x(0)\right\}$. The change of cost function in the iterative process is shown in Fig. 1. Our algorithm converges to the optimal solution with $J^*(x(0),1) = 46.77$, $J^*(x(0),2) = 67.17$, and $J^*(x(0),3) = 85.13$. The optimal state feedback gain F^* is given as $F_1^* = \begin{bmatrix} 2.3172 & -2.3317 \end{bmatrix}$, $F_2^* = \begin{bmatrix} 4.1684 & -3.7131 \end{bmatrix}$, $F_3^* = \begin{bmatrix} -5.1657 & 5.7933 \end{bmatrix}$.

Fig. 2 shows the changes in the relative errors of F^η during the iteration process. Consider the case of the first iteration step. The relative error of our algorithm is smaller than the VI method for modes 2 and 3. The reason is that our algorithm immediately uses the just generated information of mode 1. As shown in Fig. 2, the relative error of gain matrices decreases faster than the VI algorithm for each mode, which means our algorithm has a faster convergence rate than VI.

V. Conclusion

The model-free LQR control problem of discrete-time MJLSs has been studied in this paper. The model-free equations were established without knowing the system matrices. Based on the Gauss-Seidel method, a model-free fixed-point iteration algorithm was proposed for designing the optimal control policy of discrete-time MJLSs. Our model-free algorithm converges monotonically to the optimal solution and is faster than the classical model-based value iteration algorithm for discrete-time MJLSs. Finally, a numerical example was used to verify the feasibility and effectiveness of our algorithm.

APPENDIX

A. Proof of Lemma 3

Proof: Under Assumption 3, line 7 in Algorithm 1 is the same as

$$H_i^\eta = \begin{bmatrix} Q_i & 0 \\ 0 & R_i \end{bmatrix} + \begin{bmatrix} A_i & B_i \end{bmatrix}^\top \mathscr{E}_i(\bar{X}) \begin{bmatrix} A_i & B_i \end{bmatrix}.$$

Then the iteration of lines 7 and 10 in Algorithm 1 can be written as

$$\begin{aligned} H_{i,xx}^\eta =& Q_i + A_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) A_i, \\ H_{i,xu}^\eta =& A_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) B_i, \\ H_{i,ux}^\eta =& B_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) A_i, \\ H_{i,uu}^\eta =& R_i + B_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) B_i. \end{aligned}$$

The iteration of lines 8 and 9 in Algorithm 1 can be expressed as

$$\begin{aligned} X_i^{\eta+1} =& \begin{bmatrix} I \\ -(H_{i,uu}^\eta)^{-1} H_{i,ux}^\eta \end{bmatrix}^\top \begin{bmatrix} H_{i,xx}^\eta & H_{i,xu}^\eta \\ H_{i,ux}^\eta & H_{i,uu}^\eta \end{bmatrix} \\ & \times \begin{bmatrix} I \\ -(H_{i,uu}^\eta)^{-1} H_{i,ux}^\eta \end{bmatrix} \\ =& H_{i,xx}^\eta - H_{i,xu}^\eta (H_{i,uu}^\eta)^{-1} H_{i,ux}^\eta. \end{aligned}$$

For $i \in \mathcal{S}$, we have

$$\begin{aligned} X_i^{\eta+1} =& H_{i,xx}^\eta - H_{i,xu}^\eta (H_{i,uu}^\eta)^{-1} H_{i,ux}^\eta \\ =& A_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) A_i + Q_i \\ & - A_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) B_i \\ & \times \left(R_i + B_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) B_i\right)^{-1} \\ & \times B_i^\top \left(\mathscr{E}_i^1\left(X^{\eta+1}\right) + \mathscr{E}_i^2\left(X^\eta\right)\right) A_i. \end{aligned} \tag{23}$$

This completes the proof. ■

B. Proof of Lemma 5

Proof: When $i = 1$, we have $X_1^1 = Q_1 \geq 0$. When $i > 1$ and $i \in \mathcal{S}$, we have

$$\begin{aligned} X_i^1 = &\left(A_i + BF_i^0\right)^\top \left(\mathscr{E}_i^1(X^1) + \mathscr{E}_i^2(X^0)\right)\left(A_i + BF_i^0\right) \\ &+ Q_i + \left(F_i^0\right)^\top R_i F_i^0 \geq 0, \end{aligned}$$

Hence, $0 = X^0 \leq X^1$.

Assume that $X^\eta \leq X^{\eta+1}$. Define the following operators: $\mathscr{R}_i^1(X) = \sum_{j=1}^{i-1} p_{ij} R_i + B_i^\top \mathscr{E}_i^1(X) B_i$ and $\mathscr{R}_i^2(X) = \sum_{j=i}^{N} p_{ij} R_i + B_i^\top \mathscr{E}_i^2(X) B_i$. From equation (19), we have

$$\begin{aligned} &B_i^\top \left(\mathscr{E}_i^1(X^{\eta+1}) + \mathscr{E}_i^2(X^\eta)\right) A_i \\ &= -\left(\mathscr{R}_i^1(X^{\eta+1}) + \mathscr{R}_i^2(X^\eta)\right) F_i^\eta,\ i \in \mathcal{S}. \end{aligned}$$

After a tedious derivation, $X_i^{\eta+1}$, $i \in \mathcal{S}$, in equation (18) can be written as

$$\begin{aligned} X_i^{\eta+1} = &\left(A_i + BF_i^{\eta+1}\right)^\top \left(\mathscr{E}_i^1(X^{\eta+1}) + \mathscr{E}_i^2(X^\eta)\right) \\ &\times \left(A_i + BF_i^{\eta+1}\right) \\ &+ Q_i + \left(F_i^{\eta+1}\right)^\top R_i F_i^{\eta+1} - M_i^\eta, \end{aligned} \tag{24}$$

where

$$\begin{aligned} M_i^\eta = &(F_i^{\eta+1} - F_i^\eta)^\top \left(\mathscr{R}_i^1(X^{\eta+1}) + \mathscr{R}_i^2(X^\eta)\right) \\ &\times (F_i^{\eta+1} - F_i^\eta). \end{aligned}$$

For each $i \in \mathcal{S}$, $X_i^{\eta+2}$ satisfies the following equation

$$\begin{aligned} X_i^{\eta+2} = &\left(A_i + BF_i^{\eta+1}\right)^\top \left(\mathscr{E}_i^1(X^{\eta+2}) + \mathscr{E}_i^2(X^{\eta+1})\right) \\ &\times \left(A_i + BF_i^{\eta+1}\right) + Q_i + \left(F_i^{\eta+1}\right)^\top R_i F_i^{\eta+1}. \end{aligned} \tag{25}$$

Subtracting equation (24) by equation (25) yields

$$\begin{aligned} &X_i^{\eta+2} - X_i^{\eta+1} \\ &= \left(A_i + BF_i^{\eta+1}\right)^\top \left(\mathscr{E}_i^1(X^{\eta+2} - X^{\eta+1})\right. \\ &\left. + \mathscr{E}_i^2(X^{\eta+1} - X^\eta)\right)\left(A_i + BF_i^{\eta+1}\right) + M_i^\eta,\ i \in \mathcal{S}. \end{aligned}$$

Therefore, $X_i^{\eta+1} \leq X_i^{\eta+2}$, $i \in \mathcal{S}$. It shows that $X^\eta \leq X^{\eta+1}$.

The next thing is to prove $X^{\eta+1} \leq X^*$. First, $0 = X^0 \leq X^*$. Assume that $X^\eta \leq X^*$. Similar to equation (24), $X^{\eta+1}$, $i \in \mathcal{S}$, can be expressed as

$$\begin{aligned} X_i^{\eta+1} = &(A_i + BF_i^*)^\top \left(\mathscr{E}_i^1(X^{\eta+1}) + \mathscr{E}_i^2(X^\eta)\right) \\ &\times (A_i + BF_i^*) + Q_i + (F_i^*)^\top R_i F_i^* - M_i^*, \end{aligned} \tag{26}$$

where

$$M_i^* = (F_i^\eta - F_i^*)^\top \left(\mathscr{R}_i^1(X^{\eta+1}) + \mathscr{R}_i^2(X^\eta)\right)(F_i^\eta - F_i^*).$$

The unique mean square stabilizing solution X^*, $i \in \mathcal{S}$, satisfies

$$\begin{aligned} X_i^* = &(A_i + BF_i^*)^\top \mathscr{E}_i(X^*)(A_i + BF_i^*) \\ &+ Q_i + (F_i^*)^\top R_i F_i^*. \end{aligned} \tag{27}$$

Subtracting equation (26) by equation (27) yields

$$\begin{aligned} X_i^* - X_i^{\eta+1} = &(A_i + BF_i^*)^\top \left(\mathscr{E}_i^1(X^* - X^{\eta+1})\right. \\ &\left. + \mathscr{E}_i^2(X^* - X^\eta)\right)(A_i + BF_i^*) + M_i^*,\ i \in \mathcal{S}. \end{aligned}$$

Hence, $X^{\eta+1} \leq X^*$. It is obtained that $X^\eta \leq X^{\eta+1} \leq X^*$. ■

References

[1] D. Sworder, "Feedback control of a class of linear systems with jump parameters," *IEEE Transactions on Automatic Control*, vol. 14, no. 1, pp. 9-14, 1969.

[2] O. L. V. Costa and M. D. Fragoso, "Discrete-time LQ-optimal control problems for infinite Markov jump parameter systems," *IEEE Transactions on Automatic Control*, vol. 40, no. 12, pp. 2076-2088, 1995.

[3] Y. Ji and H. J. Chizeck, "Controllability, observability and discrete-time markovian jump linear quadratic control," *International Journal of Control*, vol. 48, no. 2, pp. 481-498, 1988.

[4] O. L. V. Costa, "Mean-square stabilizing solutions for discrete-time coupled algebraic Riccati equations," *IEEE Transactions on Automatic Control*, vol. 41, no. 4, pp. 593-598, April 1996.

[5] H. Abou-Kandil, G. Freiling and G. Jank, "On the solution of discrete-time Markovian jump linear quadratic control problems," *Automatica*, vol. 31, no. 5, pp. 765-768, 1995.

[6] O. L. V. Costa and R. P. Marques, "Maximal and stabilizing Hermitian solutions for discrete-time coupled algebraic Riccati equations," *Mathematics of Control, Signals and Systems*, vol. 12, no. 2, pp. 167-195, Jun. 1999.

[7] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction (second edition)," *Cambridge, MA: The MIT Press*, 2018.

[8] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32-50, 2009.

[9] S. He, H. Fang, and M. Zhang et al., "Adaptive optimal control for a class of nonlinear systems: The online policy iteration approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 2, pp. 549-558, 2020.

[10] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data" *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 14-25, 2011.

[11] B. Kiumarsi, F. L. Lewis, and M. -B. Naghibi-Sistani et al., "Optimal tracking control of unknown discrete-time linear systems using input-output measured data" *IEEE Transactions on Cybernetics*, vol. 45, no. 12, pp. 2770-2779, 2015.

[12] A. Al-Tamimi, F. L. Lewis and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, no. 3, pp. 473-481, 2007.

[13] B. Kiumarsi, F. L. Lewis, and Z. P. Jiang, "H_∞ control of linear discrete-time systems: Off-policy reinforcement learning" *Automatica*, Vol. 78, pp. 144-152, 2017.

[14] H. Fang, M. Zhang, and S. He et al., "Solving the zero-sum control problem for tidal turbine system: An online reinforcement learning approach," *IEEE Transactions on Cybernetics*, 2022.

[15] J. Song, S. He, and F. Liu et al., "Data-driven policy iteration algorithm for optimal control of continuous-time Itô stochastic systems with Markovian jumps," *IET Control Theory & Applications*, vol. 10, no. 12, pp. 1431-1439, 2016.

[16] K. Zhang, H. Zhang, and Y. Cai et al., "Parallel optimal tracking control schemes for mode-dependent control of coupled Markov jump systems via integral RL method," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 3, pp. 1332-1342, 2020.

[17] J. P. Jansch-Porto, B. Hu and G. Dullerud, "Policy learning of MDPs with mixed continuous/discrete variables: A case study on model-free control of Markovian jump systems," in *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, vol. 120, pp. 947-957, Urbana, Illinois, USA, 2020.

[18] Z. Huang, Y. Tu, and H. Fang et al., "Off-policy reinforcement learning for tracking control of discrete-time Markov jump linear systems with completely unknown dynamics," *Journal of the Franklin Institute*, 2022.