# Monte Carlo $TD(\lambda)$-Methods for the Optimal Control of Discrete-Time Markovian Jump Linear Systems ⋆

Oswaldo L. V. Costa [a], Julio C.C. Aya [a],

[a] *Departamento de Engenharia de Telecomunicações e Controle*
*Escola Politécnica da Universidade de São Paulo*
*CEP: 05508 900 São Paulo SP Brazil*

*Monte Carlo $TD(\lambda)$-Methods for Jump Systems*

**Abstract**

In this paper we present an iterative technique based on Monte Carlo simulations for deriving the optimal control of the infinite horizon linear regulator problem of discrete-time Markovian jump linear systems for the case in which the transition probability matrix of the Markov chain is not known. We trace a parallel with the theory of $TD(\lambda)$ algorithms for Markovian decision processes to develop a $TD(\lambda)$ like algorithm for the optimal control associated to the maximal solution of a set of coupled algebraic Riccati equations (CARE). It is assumed that either there is a sample of past observations of the Markov chain that can be used for the iterative algorithm, or it can be generated through a computer program. Our proofs rely on the spectral radius of the closed loop operators associated to the mean square stability of the system being less than 1.

*Key words:* $TD(\lambda)$ methods, jump systems, Markov parameters, optimal control, Monte Carlo simulations.

## 1 Introduction

In this paper we consider discrete-time Markovian jump linear systems (MJLS) described as below, in an appropriate probabilistic space $(\Omega, P, \{\mathcal{F}_k\}, \mathcal{F})$:

$$\begin{aligned} x(k+1) &= A_{\theta(k)}x(k) + B_{\theta(k)}u(k) \\ x(0) &= x_0 \quad \theta(0) = \theta_0. \end{aligned} \tag{1}$$

Here $\theta(k)$ is a Markov chain taking values in $\{1, \ldots, N\}$ with transition probability matrix $\mathcal{P} = (p_{ij})$ and $u = (u(0), \ldots)$ is the control input sequence assumed to belong to $\mathcal{U} = \{u = (u(0), \ldots); u(k)$ is $\mathcal{F}_k$-measurable for each $k\}$. It is assumed here that both $x(k)$ and $\theta(k)$ are directly accessible at each time $k$. Several results regarding applications, stability conditions and optimal control problems of MJLS, can be found in the current literature (see, for instance, [1], [2], [4], [6]-[12], [14], [15], [17], [18]).

For $u \in \mathcal{U}$, one considers the following quadratic cost for system (1)

$$J_{(x_0,\theta_0)}(u) = E_{(x_0,\theta_0)}\left( \sum_{k=0}^{\infty} \left( \left\|C_{\theta(k)}x(k)\right\|^2 + \left\|D_{\theta(k)}u(k)\right\|^2 \right) \right) \tag{2}$$

and it is desired to minimize (2) over $u \in \mathcal{U}$. It has

⋆ This work was suported in part by FAPESP (Research Council of the State of São Paulo), grant 97/04668-1, PRONEX, grant 015/98, and CNPq (Brazilian National Research Council), grants 305173/88 and 142230/97-9.

*Email addresses:* `oswaldo@lac.usp.br` (Oswaldo L. V. Costa), `julio@lac.usp.br` (Julio C.C. Aya).

been shown in the literature (cf. [7], [11], [12]) that the solution of this problem is associated to the existence of a solution $P = (P_1, \ldots, P_N)$, $P_i \geq 0$, $i = 1, \ldots, N$, to the following set of coupled algebraic Riccati equations (CARE), for $i = 1, \ldots, N$

$$\begin{aligned} X_i &= A_i^* \mathcal{E}_i(X) A_i + C_i^* C_i \\ &- A_i^* \mathcal{E}_i(X) B_i (B_i^* \mathcal{E}_i(X) B_i + D_i^* D_i)^{-1} B_i^* \mathcal{E}_i(X) A_i \end{aligned} \tag{3}$$

where $\mathcal{E}(X) = (\mathcal{E}_1(X), \ldots, \mathcal{E}_N(X))$ is defined, for $X = (X_1, \ldots, X_N)$ as $\mathcal{E}_i(X) = \sum_{j=1}^N p_{ij} X_j$. If such a solution $P$ exists, it can be shown (see [7]) that the optimal control law for the problem posed by equations (1), (2) and (3) is given by the feedback control law $u(k) = F_{\theta(k)} x(k)$ where $F = (F_1, \ldots, F_N)$ is given by

$$F_i = -(B_i^* \mathcal{E}_i(P) B_i + D_i^* D_i)^{-1} B_i^* \mathcal{E}_i(P) A_i. \tag{4}$$

Several numerical algorithms for obtaining the maximal solution $P$ of the CARE (3) using standard dynamic programming and optimal control tools, quasi-linearization techniques or LMI (linear matrix inequalities) formulations have been proposed in the current literature (see [1], [2], [4], [7], [8], [9], [10], [11], [12], [15], [18]). All these methods assume that the transition probability matrix $\mathcal{P}$ is available for the computations. The goal of this paper is to present an iterative algorithm for obtaining the feedback matrices (4) associated to the solution $P$ of the CARE (3) for the case in which the transition probability matrix $\mathcal{P}$ is not known. We assume that either past observations of the Markov chain $\theta(k)$ are available or that, like in [3],[16], there exists a computer program that simulates the probabilistic transitions from a given state $i$ to a successor state $j$ for the Markov chain $\theta(k)$. This algorithm traces a close parallel with the $TD(\lambda)$ methods (see [3],[16]) for Markovian decision processes (MDP for short).

As pointed out in [3],[16], the computational methods for MDP require an explicit model for the cost structure and the transition probabilities of the system for all states of the Markov chain. In many situations such model is not available or require complex computations (see [16], pages 112-116). $TD(\lambda)$ methods have been applied to solve problems related to MDP (see for instance [3],[16]) for the case in which the transition probability matrix of the Markov chain is not known. As in our case, it is assumed that (cf. [3],[16]) it is possible to simulate the probabilistic transitions from any given state to a successor state, and the cost-to-go function of a given policy is progressively calculated by generating several sample systems trajectories. In contrast to the computational methods for MDP, generating sample trajectories of the Markov chain required for $TD(\lambda)$ Monte Carlo methods is easy, and particularly attractive when the number of states is large and the value function is required only at a subset of these states. One can generate many sample trajectories starting from these states, ignoring all others ([16], page 115). Therefore our technique would be useful in situations similar to those presented [3],[16], that is, when the transition probabilities of the system are not available or require complex computations, or when the number of states is large but the value function is required only at a some subset of the states. In both cases it is assumed that sample paths of the Markov chain can be generated through a computer program. As mentioned before, another possibility would be for the case in which there are past observations of the Markov chain that could be used for the iterative algorithm. We propose a Monte Carlo policy evaluation like algorithm that incrementally updates the estimates for the feedback optimal control matrices $F$ given by (4). Proof of convergence is obtained by following arguments similar to those in [3]. The main novelty in our approach is that our proofs rely on the fact that the spectral radius of the closed loop operators associated to the mean square stability of the system are less than one, in contrast with the proofs in [3], based on the fact that either the MDP has a discount factor $\alpha$ less than 1 or that it has a cost-free termination state. The paper is presented in the following way. Section 2 deals with the notation and some preliminary results. Section 3 presents the $TD(\lambda)$ method for solving the CARE (3). The main result is Theorem 1, in which the proof of convergence for the value iteration step is established. Section 4 presents the proofs of the results and section 5 presents a numerical example. Section 6 concludes the paper with some final comments.

## 2 Notation and Preliminary Results

For $\mathbb{X}$ and $\mathbb{Y}$ complex Banach spaces we set $\mathbb{B}(\mathbb{X}, \mathbb{Y})$ the Banach space of all bounded linear operators of $\mathbb{X}$ into

$\mathbb{Y}$, with the uniform induced norm represented by $\|\cdot\|$. For simplicity we shall set $\mathbb{B}(\mathbb{X}) := \mathbb{B}(\mathbb{X}, \mathbb{X})$. The spectral radius of an operator $\mathcal{T} \in \mathbb{B}(\mathbb{X})$ will be denoted by $r_\sigma(\mathcal{T})$. If $\mathbb{X}$ is a Hilbert space then the inner product will be denoted by $\langle \cdot;\cdot \rangle$, and for $\mathcal{T} \in \mathbb{B}(\mathbb{X})$, $\mathcal{T}^*$ will denote the adjoint operator of $\mathcal{T}$. As usual, $\mathcal{T} \geq 0$ ($\mathcal{T} > 0$ respectively) will denote that the operator $\mathcal{T} \in \mathbb{B}(\mathbb{X})$ will be positive semi-definite (positive definite). In particular we shall denote by $\mathbb{C}^n$ the $n$ dimensional complex Euclidean spaces and by $\mathbb{B}(\mathbb{C}^n, \mathbb{C}^m)$ the normed bounded linear space of all $m \times n$ complex matrices, with $\mathbb{B}(\mathbb{C}^n) := \mathbb{B}(\mathbb{C}^n, \mathbb{C}^n)$ and the inner product in $\mathbb{B}(\mathbb{C}^n)$ given by $\langle H; V \rangle = tr\{H^* V\}$ for $H, V \in \mathbb{B}(\mathbb{C}^n)$ (and $\|H\|^2 = tr\{H^* H\}$). The superscript $'$ will denote transpose of a matrix. We shall write $\|\cdot\|_2$ for the norm in $\mathbb{B}(\mathbb{C}^n)$ induced by the Euclidean norm in $\mathbb{C}^n$.

Set $\mathbb{H}^{n,m}$ the linear space made up of all $N$-sequences of complex matrices $V = (V_1, \ldots, V_N)$ with $V_i \in \mathbb{B}(\mathbb{C}^n, \mathbb{C}^m)$, $i = 1, \ldots, N$ and, for simplicity, set $\mathbb{H}^n := \mathbb{H}^{n,n}$. Throughout the paper we shall consider $\mathbb{H}^n$ equipped with the following inner product. For $H = (H_1, \ldots, H_N), V = (V_1, \ldots, V_N) \in \mathbb{H}^n$ we shall define $\langle \cdot;\cdot \rangle$ in $\mathbb{H}^n$ as follows: $\langle H; V \rangle := \sum_{i=1}^N tr\{H_i^* V_i\}$, and $\|H\|^2 := \langle H; H \rangle$ (so that $\|H\|^2 = \sum_{i=1}^N \|H_i\|^2$). Therefore with the above inner product, $\mathbb{H}^n$ is a Hilbert space. We shall also set the following $\|H\|_{max}$ and $\|H\|_1$ norms in $\mathbb{H}^n$. For $H = (H_1, \ldots, H_N) \in \mathbb{H}^n$, set $\|H\|_{max} := max\{\|H_i\|_2 ; i = 1, \ldots, N\}$ and $\|H\|_1 := \sum_{i=1}^N \|H_i\|_2$. We set $\mathbb{H}^{n+} := \{V = (V_1, \ldots, V_N) \in \mathbb{H}^n; V_i \geq 0, \ i = 1, \ldots, N\}$ and shall write, for $V = (V_1, ..., V_N) \in \mathbb{H}^n$ and $S = (S_1, \ldots, S_N) \in \mathbb{H}^n$, that $V > S$ if $V_i - S_i > 0$ for $i = 1, \ldots, N$ (similarly for $\geq$). For $\Gamma = (\Gamma_1, \ldots, \Gamma_N) \in \mathbb{H}^n$ we define the following operators $\mathcal{L}(\cdot) = (\mathcal{L}_1(\cdot), \ldots, \mathcal{L}_N(\cdot))$, $\mathcal{T}(\cdot) = (\mathcal{T}_1(\cdot), \ldots, \mathcal{T}_N(\cdot)) \in \mathbb{B}(\mathbb{H}^n)$ for $V = (V_1, \ldots, V_N) \in \mathbb{H}^n$

$$\mathcal{L}_i(V) := \Gamma_i^* \mathcal{E}_i(\cdot) \Gamma_i \tag{5}$$

$$\mathcal{T}_j(V) := \sum_{i=1}^N p_{ij} \Gamma_i V_i \Gamma_i^* \tag{6}$$

where the operator $\mathcal{E}(\cdot) = (\mathcal{E}_1(\cdot), \ldots, \mathcal{E}_N(\cdot)) \in \mathbb{B}(\mathbb{H}^n)$ is defined as in section 1. It is easy to verify that the operators $\mathcal{E}$, $\mathcal{L}$, and $\mathcal{T}$ map $\mathbb{H}^{n+}$ into $\mathbb{H}^{n+}$, and it has been proved in [6] that $r_\sigma(\mathcal{L}) = r_\sigma(\mathcal{T})$ (in fact $\mathcal{L} = \mathcal{T}^*$ according to the inner product defined above). We also define the operator $\mathcal{G} \in \mathbb{B}(\mathbb{H}^n)$ as

$$\mathcal{G}_i(V) := \sum_{j=1}^N p_{ij} \Gamma_j^* V_j \Gamma_j \tag{7}$$

where again $V = (V_1, ..., V_N) \in \mathbb{H}^n$ and $\mathcal{G}(V) = (\mathcal{G}_1(V), \ldots, \mathcal{G}_1(V))$. It has been shown in [6] that $r_\sigma(\mathcal{G}) = r_\sigma(\mathcal{L}) = r_\sigma(\mathcal{T})$.

We assume in (1) and (2) that $A = (A_1, ..., A_N) \in \mathbb{H}^n$, $B = (B_1, ..., B_N) \in \mathbb{H}^{m,n}$, $C = (C_1, ..., C_N) \in \mathbb{H}^{n,p}$ and $D = (D_1, ..., D_N) \in \mathbb{H}^{m,p}$. It has been shown in [6], [14], that model (1) with $u(k) = F_{\theta(k)} x(k)$, and $V_i(k) = E\big(x(k)x(k)^* 1_{\{\theta(k)=i\}}\big)$, $V(k) = (V_1(k), \ldots, V_N(k)) \in \mathbb{H}^{n+}$ leads to $V(k+1) = \mathcal{T}(V(k))$, $k = 0, 1, \ldots$, where $\Gamma_i = A_i + B_i F_i$ in (6), and $E\left(\|x(k)\|^2\right) = \sum_{i=1}^N tr\{V_i(k)\}$. In what follows we shall need also the operator for the $4^{th}$ moment of $x(t)$. With the Kronecker product $L \otimes K \in \mathbb{B}(\mathbb{C}^{n^2})$ for $L, K \in \mathbb{B}(\mathbb{C}^n)$ and the operator $vec\{\cdot\} : \mathbb{B}(\mathbb{C}^n) \mapsto \mathbb{C}^{n^2}$ defined in the usual way (see [5]), let the operator $\mathcal{H} \in \mathbb{B}(\mathbb{H}^{n^2})$ be as follows: for $S = (S_1, \ldots, S_N) \in \mathbb{H}^{n^2}$, $\mathcal{H}(S) = (\mathcal{H}_1(S), \ldots, \mathcal{H}_N(S))$ is defined as

$$\mathcal{H}_j(S) := \sum_{i=1}^N p_{ij} (\overline{\Gamma}_i \otimes \Gamma_i) S_i (\Gamma_i' \otimes \Gamma_i^*). \tag{8}$$

Let $S_i(k) = E\big(vec\{x(k)x(k)^*\} vec\{x(k)x(k)^*\}^* 1_{\{\theta(k)=i\}}\big)$, $S(k) = (S_1(k), \ldots, S_N(k)) \in \mathbb{H}^{n^2+}$. Define $\overline{\Gamma} \otimes \Gamma = (\overline{\Gamma}_1 \otimes \Gamma_1, \ldots, \overline{\Gamma}_N \otimes \Gamma_N)$. We have the following results, proved in section 4.

**Lemma 1** *$S(k+1) = \mathcal{H}(S(k))$ and $E\left(\|x(k)\|^4\right) = \sum_{i=1}^N tr\{S_i(k)\}$.*

**Lemma 2** *If $\lambda^2 < \frac{1}{\|\overline{\Gamma} \otimes \Gamma\|_{max}}$ then $\lambda^2 r_\sigma(\mathcal{H}) < 1$.*

Next we define the stability concept that we shall consider in the following sections.

**Definition 1** *We say that $F = (F_1, ..., F_N) \in \mathbb{H}^{n,m}$ stabilizes $(A, B)$ in the mean square sense if, when we make $u(k) = F_{\theta(k)} x(k)$ in system (1), we have that $E(\|x(k)\|^2) \to 0$ as $k \to \infty$ for any initial condition $x(0)$ and $\theta(0)$. We say that $(A, B)$ is mean square stabilizable if for some $F = (F_1, \ldots, F_N) \in \mathbb{H}^{n,m}$, we have that $F$ stabilizes $(A, B)$ in the mean square sense.*

The following result, proved in [6], shows that $F = (F_1, \ldots, F_N)$ stabilizes system (1) in the mean square sense if and only if the expectral radius of the operator (6) (or (5), (7)) in closed loop is less than one.

**Lemma 3** *$F = (F_1, \ldots, F_N) \in \mathbb{H}^{n,m}$ stabilizes $(A, B)$ in the mean square sense if and only if $r_\sigma(\mathcal{T}) < 1$, where $\mathcal{T}$ is as in* (6) *with $\Gamma_i = A_i + B_i F_i$.*

We make the following definition:

**Definition 2** *We define $\mathcal{F}(\cdot) = (\mathcal{F}_1(\cdot), \ldots, \mathcal{F}_N(\cdot)) : \mathbb{H}^{n+} \to \mathbb{H}^{n,m}$, $\mathcal{V}(\cdot) = (\mathcal{V}_1(\cdot), \ldots, \mathcal{V}_N(\cdot)) : \mathbb{H}^{n,m} \to \mathbb{H}^{n}$ and $\mathcal{R}(\cdot) = (\mathcal{R}_1(\cdot), \ldots, \mathcal{R}_N(\cdot)) : \mathbb{H}^{n+} \to \mathbb{H}^{n+}$ as*

$$
\begin{aligned}
&\mathcal{F}_i(X) := -(B_i^* \mathcal{E}_i(X) B_i + D_i^* D_i)^{-1} B_i^* \mathcal{E}_i(X) A_i \qquad (9)\\
&\mathcal{V}_i(F) := C_i^* C_i + F_i^* D_i^* D_i F_i \\
&\mathcal{R}_i(X) := A_i^* \mathcal{E}_i(X) A_i + C_i^* C_i \\
&\qquad - A_i^* \mathcal{E}_i(X) B_i (B_i^* \mathcal{E}_i(X) B_i + D_i^* D_i)^{-1} B_i^* \mathcal{E}_i(X) A_i
\end{aligned}
$$

*where $X = (X_1, \ldots, X_N) \in \mathbb{H}^{n+}$, $F = (F_1, \ldots, F_N) \in \mathbb{H}^{n,m}$.*

The following identity will be useful in the sequel: for any $F = (F_1, \ldots, F_N) \in \mathbb{H}^{n,m}$, we have that

$$
\begin{aligned}
&(A_i + B_i F_i)^* \mathcal{E}_i(S)(A_i + B_i F_i) + F_i^* D_i^* D_i F_i = \\
&(A_i + B_i \mathcal{F}_i(S))^* \mathcal{E}_i(S)(A_i + B_i \mathcal{F}_i(S)) \\
&+ \mathcal{F}_i(S)^* D_i^* D_i \mathcal{F}_i(S) + \qquad (10)\\
&(F_i - \mathcal{F}_i(S))^* (B_i^* \mathcal{E}_i(S) B_i + D_i^* D_i)(F_i - \mathcal{F}_i(S)).
\end{aligned}
$$

The next Lemma, proved in [8], provides the existence of the maximal solution for (3) whenever $(A, B)$ is mean square stabilizable. It is based on a quasi-linearization technique for the CARE (3), and parallels the policy iteration technique for MDP (see Remark 1 below).

**Lemma 4** *Suppose that $(A, B)$ is mean square stabilizable and considers $F^0 = (F_1^0, \ldots, F_N^0) \in \mathbb{H}^{n,m}$ such that stabilizes $(A, B)$ in the mean square sense. Then for $l = 0, 1, 2, \ldots$, there exists $P^l = (P_1^l, \ldots, P_N^l)$ which satisfies the following properties:*

*a) $P^0 \geq P^1 \geq \cdots \geq P^l \geq X$, for arbitrary $X \in \mathbb{H}^{n+}$ such that $X \geq \mathcal{R}(X)$.*

*b) $r_\sigma(\mathcal{L}^l) < 1$ where $\mathcal{L}^l(\cdot) = (\mathcal{L}_1^l(\cdot), \ldots, \mathcal{L}_N^l(\cdot))$ and for $i = 1, \ldots, N$,*

$$
\begin{aligned}
\mathcal{L}_i^l(\cdot) &:= A_i^{l^*} \mathcal{E}_i(\cdot) A_i^l, \ A_i^l := A_i + B_i F_i^l, \\
F_i^l &:= \mathcal{F}_i(P^{l-1}) \ \text{for} \ l = 1, 2, \ldots.
\end{aligned}
$$

*c) $P^l$ satisfies $P^l = \mathcal{L}^l(P^l) + \mathcal{V}(F^l)$ and is given by $P^l = \sum_{k=0}^{\infty} (\mathcal{L}^l)^k (\mathcal{V}(F^l))$.*

*Moreover there exists $P^+ = (P_1^+, \ldots, P_N^+) \in \mathbb{H}^{n+}$ such that $P^+ = \mathcal{R}(P^+)$, $P^+ \geq X$ for any $X \in \mathbb{H}^{n+}$ such that $X \geq \mathcal{R}(X)$, and $P^l \to P^+$ as $l \to \infty$. Furthermore $r_\sigma(\mathcal{L}^+) \leq 1$, where $\mathcal{L}^+(\cdot) = (\mathcal{L}_1^+(\cdot), \ldots, \mathcal{L}_N^+(\cdot))$ is defined as $\mathcal{L}_i^+(\cdot) = A_i^{+^*} \mathcal{E}_i(\cdot) A_i^+$, for $i = 1, \ldots, N$, and $F_i^+ = \mathcal{F}_i(P^+)$, $A_i^+ = A_i + B_i F_i^+$.*

**Remark 1** *Step c) in Lemma 4, which corresponds the calculation of the solution of the linear system $X^l = \mathcal{L}^l(X^l) + \mathcal{V}(F^l)$, can be seen as the policy evaluation step in the policy iteration technique for MDP, while from identity* (10)*, the choice $F_i^{l+1} = \mathcal{F}_i(P^l)$ can be seen as the policy improvement step.*

We close this section with the following result which will be useful in the sequel. In an appropriate probabilistic space $(\Phi, \mathbb{P}, \{\Sigma_t\}, \Sigma)$ consider two sequences of stochastic processes $\{W(t); t = 0, 1, \ldots\}$, $\{\gamma(t); t = 0, 1, \ldots\}$ such that for each $t = 0, 1, \ldots, W(t)$ is an $n \times n$ matrix and $\gamma(t)$ is a scalar positive variable $\Sigma_t$-adapted. Assume that $\mathbb{P}$-almost surely we have

$$
\sum_{t=0}^{\infty} \gamma(t) = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \gamma(t)^2 < \infty. \qquad (11)
$$

Assume also that the noise matrix terms satisfy

$$
\mathbb{E}(W(t) \big| \Sigma_t) = 0 \quad \text{and} \quad \mathbb{E}(\|W(t)\|^2 \big| \Sigma_t) \leq A(t). \qquad (12)
$$

where $\{A(t); t = 0, 1, \ldots\}$ is a stochastic process such that for each $t = 0, 1, \ldots$, $A(t)$ is a scalar positive variable $\Sigma_t$-adapted. Consider the stochastic process $\{R(t); t = 0, 1, \ldots\}$, $R(t)$ an $n \times n$ matrix, with arbitrary initial value $R(0)$, given by the sequence

$$
R(t+1) = (1 - \gamma(t)) R(t) + \gamma(t) W(t) \qquad (13)
$$

**Lemma 5** *Suppose that* (11)*-*(12) *are satisfied and the sequence of $n \times n$ matrices $\{R(t); t = 0, 1, \ldots\}$ are given by* (13)*. If the sequence $A(t)$ is bounded $\mathbb{P}$-almost surely then $R(t)$ converges to zero $\mathbb{P}$-almost surely.*

**Proof.** See Corollary 4.1, page 161 in [3]. ■

## 3 $TD(\lambda)$ Algorithm

If the transition probability matrix $\mathcal{P}$ were known in advance we could use Lemma 4 to obtain an iterative algorithm for the maximal solution $P$ of the CARE (3) and then from (4) obtain the optimal control law $F$. If the transition probability matrix $\mathcal{P}$ is not known then we could try to establish a Monte Carlo algorithm for obtaining the solution $P^l$ as in c) of Lemma 4. But this would not be enough to obtain $F^{l+1} = \mathcal{F}\left(P^l\right)$ since, as can be seen from (9), it depends on $\mathcal{P}$ through the operator $\mathcal{E}(\cdot)$. An alternative way would be to calculate directly $S^l := \mathcal{E}\left(P^l\right)$, which would lead us to the following equation: $S^l = \mathcal{G}^l\left(S^l\right) + \mathcal{E}(\mathcal{V}\left(F^l\right))$. Let us write for simplicity $Q^l = \mathcal{E}\left(\mathcal{V}\left(F^l\right)\right)$. Note that as seen in Section 2, $r_\sigma(\mathcal{G}^l) = r_\sigma\left(\mathcal{L}^l\right) < 1$, and Theorem 5.17, page 102 in [19] can be applied to say that the equation in $Y$

$$Y = \mathcal{G}^l(Y) + Q^l \tag{14}$$

has a unique solution $S^l$ given by $S^l = \sum_{k=0}^{\infty}\left(\mathcal{G}^l\right)^k(Q^l)$. Once $S^l$ is calculated, $F^{l+1}$ can be obtained from (9) as

$$F_i^{l+1} = -(B_i^* S_i^l B_i + D_i^* D_i)^{-1} B_i^* S_i^l A_i \tag{15}$$

The remaining of this section is now devoted to calculating $S^l$ through Monte Carlo simulations, tracing a parallel with $TD(\lambda)$ methods (see [3]). For simplicity we shall suppress the superscript $l$. For any $\lambda \in [0,1)$ define the operator $\mathcal{J} \in \mathbb{B}(\mathbb{H}^n)$ and $Z \in \mathbb{H}^n$ in the following way:

$$\mathcal{J}(\cdot) := (1-\lambda)\sum_{k=0}^{\infty}\lambda^k \mathcal{G}^{k+1}(\cdot),\ Z := \sum_{k=0}^{\infty}(\lambda\mathcal{G})^k(Q).$$

By iterating (14) with $Y = S$ we have

$$S = \mathcal{G}^{k+1}(S) + \sum_{t=0}^{k}\mathcal{G}^{k-t}(Q), \qquad k = 0,1,\ldots. \tag{16}$$

We have from (16) that

$$\begin{aligned} S &= \sum_{k=0}^{\infty}(1-\lambda)\lambda^k S \\ &= \sum_{t=0}^{\infty}(\lambda\mathcal{G})^t(Q) + (1-\lambda)\ \sum_{k=0}^{\infty}\lambda^k\mathcal{G}^{k+1}(S) \\ &= Z + \mathcal{J}(S) \\ &= \sum_{k=0}^{\infty}\lambda^k\left[\mathcal{G}^k(Q) + \mathcal{G}^{k+1}(S) - \mathcal{G}^k(S)\right] + S. \end{aligned} \tag{17}$$

Equation (17) suggests the following temporal difference method. Set $\Xi := \{1,\ldots,N\}^\infty$ and for each $i = 1,\ldots,N$ and $t = 1,2,\ldots$ consider random variables $\Theta_i(t) = (\theta_i(t,0),\theta_i(t,1),\ldots) \in \Xi$ such that $\theta_i(t,0) = i$ and $\{\theta_i(t,k)\}$ has the same distribution as $\{\theta(k)\}$. Consider $\{\gamma(t); t = 1,2,\ldots\}$ satisfying (11) with probability 1 and for arbitrary $Y(0) = (Y_1(0),\ldots,Y_N(0)) \in \mathbb{H}^n$ define for $t = 1,2,\ldots$, the sequence $Y(t) = (Y_1(t),\ldots,Y_N(t)) \in \mathbb{H}^n$ in the following way:

$$Y(t+1) = Y(t) + \gamma(t)\sum_{k=0}^{\infty}\lambda^k\mathcal{D}(t,k,Y(t)), \tag{18}$$

where for $k = 0,1,\ldots$ the bounded affine operator $D(t,k,\cdot)$ is defined for $V = (V_1,\ldots,V_N) \in \mathbb{H}^n$ in terms of $\mathcal{B}_i(t,k) \in \mathbb{H}^n$ and $\mathcal{C}_i(t,k,\cdot) \in \mathbb{B}(\mathbb{H}^n)$ as

$$\begin{aligned}
&\Upsilon_i(t,k) := \left(A_{\theta_i(t,k)} + B_{\theta_i(t,k)}F_{\theta_i(t,k)}\right)\Upsilon_i(t,k-1)\\
&\Upsilon_i(t,0) := I\\
&\mathcal{D}_i(t,k,V) := \mathcal{B}_i(t,k) + \mathcal{C}_i(t,k,V)\\
&\mathcal{B}_i(t,k) := \Upsilon_i(t,k)^*\Big[C^*_{\theta_i(t,k+1)}C_{\theta_i(t,k+1)}\\
&\quad + F^*_{\theta_i(t,k+1)}D^*_{\theta_i(t,k+1)}D_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)}\Big]\Upsilon_i(t,k)\\
&\mathcal{C}_i(t,k,V) := \Upsilon_i(t,k)^*\Big[\big(A_{\theta_i(t,k+1)}\\
&\quad + B_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)}\big)^* V_{\theta_i(t,k+1)}\big(A_{\theta_i(t,k+1)}\\
&\quad + B_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)}\big) - V_{\theta_i(t,k)}\Big]\Upsilon_i(t,k),\\
&D(t,k,V) := (\mathcal{D}_1(t,k,V),\ldots,\mathcal{D}_N(t,k,V)).
\end{aligned}$$

Notice that

$$E\left(\mathcal{D}_i(t,k,V)\middle|\theta(0)=i\right) = \mathcal{G}_i^k(Q) + \mathcal{G}_i^{k+1}(V) - \mathcal{G}_i^k(V).$$

We define $W(t) = (W_1(t),\ldots,W_N(t))$ with

$$\begin{aligned} W_i(t) := \sum_{k=0}^{\infty}\lambda^k\Big(&\mathcal{D}_i(t,k,Y(t)) \\ &- \left(\mathcal{G}_i^k(Q) + \mathcal{G}_i^{k+1}(Y(t)) - \mathcal{G}_i^k(Y(t))\right)\Big). \end{aligned}$$

From (17) we can rewrite (18) as follows:

$$\begin{aligned} Y(t+1) = (1-\gamma(t))Y(t)& \\ + \gamma(t)(Z + \mathcal{J}(Y(t)) + W(t))&. \end{aligned} \tag{19}$$

We denote by $\Sigma_t$ the history of the algorithm until time $t$, which can be defined as

$$\Sigma_t = \sigma\Big\{Y(0), \ldots, Y(t), \Theta_i(s), s = 1, \ldots, t-1, \\ i = 1, \ldots, N, \gamma(s), s = 1, \ldots, t\Big\}.$$

Therefore for each $i = 1, \ldots, N$,

$$\mathbb{E}\left(W_i(t)\middle|\Sigma_t\right) = \sum_{k=0}^{\infty} \lambda^k \bigg( E\Big(\mathcal{D}_i(t,k,Y(t))\Big|\theta(0) = i\Big) \\ - \Big(\mathcal{G}_i^k(Q) + \mathcal{G}_i^{k+1}(Y(t)) - \mathcal{G}_i^k(Y(t))\Big)\bigg) = 0,$$

satisfying equation (12). We have now the main result of this section, proved in the next section, establishing the convergence of $Y(t)$ to $S$. We assume that all matrices are real.

**Theorem 1** *If $\lambda^2 r_\sigma(\mathcal{H}) < 1$ then the sequence $\{Y(t); t = 1, 2, \ldots\}$ converges to $S$ with probability 1.*

**Remark 2** *Lemma 2 establishes an upperbound for $\lambda$ so that $\lambda^2 r_\sigma(\mathcal{H}) < 1$ will hold. Of course Lemma 2 is just a sufficient condition, and it is possible that the convergence of $Y(t)$ to $S$ takes place even for bigger values of $\lambda$.*

Summing up, the algorithm is as follows. Suppose we have sample paths $\Theta_i(t) = (\theta_i(t,0), \theta_i(t,1), \ldots)$, $\theta_i(t,0) = i$ of the Markov chain $\{\theta(k)\}$, $t = 1, 2, \ldots$, and that $\{\gamma(t); t = 1, 2, \ldots\}$ satisfies (11) with probability 1. Then for $\ell = 0, 1, \ldots$,

i) $F_i^\ell$ is calculated as (15) (except for $\ell = 0$).
ii) $S^\ell$ is the stationary value of equation (18).

According to Lemma 4 we have that $F^\ell$ tends to $\mathcal{F}(P^+)$, where $P^+$ is the maximal solution of the CARE (3).

## 4 Proofs

First we present the proofs of Lemmas 1 and 2.

**Proof.** Lemma 1: Since $x(k+1) = \Gamma_{\theta(k)}x(k)$, it is easy to check that $x_j(k+1)x_j(k+1)^* = 1_{\{\theta(k+1)=j\}} \sum_{i=1}^N \Gamma_i x_i(k) x_i(k)^* \Gamma_i^*$ where $x_i(k) := x(k)1_{\{\theta(k)=i\}}$. Let $z_i(k) = vec\{x_i(k)x_i(k)^*\}$, $i = 1, \ldots, N$. After some manipulation it follows that $z_j(k+1)z_j(k+1)^* = 1_{\{\theta(k+1)=j\}} \sum_{i=1}^N (\overline{\Gamma}_i \otimes \Gamma_i) z_i(k) z_i(k)^* (\Gamma_i' \otimes \Gamma_i^*)$ and writing $S_i(k) = E\left(z_i(k)z_i(k)^* 1_{\{\theta(k)=i\}}\right)$ it follows that $S_j(k+1) = \sum_{i=1}^N p_{ij} (\overline{\Gamma}_i \otimes \Gamma_i) S_i(k) (\Gamma_i' \otimes \Gamma_i^*)$. Finally notice that $tr\{z_i(k)z_i(k)^*\} = \|x_i(k)\|^2\, tr\{x_i(k)\, x_i(k)^*\} = \|x_i(k)\|^4$. ■

**Proof.** Lemma 2: For any $S = (S_1, \ldots, S_N) \in \mathbb{H}^{n^2}$, $\|\mathcal{H}(S)\|_1 = \sum_{j=1}^N \|\mathcal{H}_j(S)\|_2 \leq \left\|\overline{\Gamma} \otimes \Gamma\right\|_{max}^2 \|S\|_1$, and therefore $\|\mathcal{H}\|_1 = sup\{\frac{\|\mathcal{H}(S)\|_1}{\|S\|_1}; S \in \mathbb{H}^{n^2}\} \leq \left\|\overline{\Gamma} \otimes \Gamma\right\|_{max}^2$. Since $r_\sigma(\mathcal{H}) \leq \|\mathcal{H}\|_1$ the result follows. ■

To prove Theorem 1 we shall need the following propositions:

**Proposition 1** $r_\sigma(\mathcal{J}) < 1$.

**Proof.** As mentioned above, $r_\sigma(\mathcal{G}) = r_\sigma(\mathcal{L}) < 1$. Suppose we had $r_\sigma(\mathcal{J}) \geq 1$. Then for some $\beta \in \mathbb{C}$, $|\beta| \geq 1$, and $V \in \mathbb{H}^n$, $\mathcal{J}(V) = \beta V$. In this case, $(1-\lambda)|\beta| \geq (1-\lambda)$ which implies that $\frac{|\beta|}{1+\lambda\left(|\beta|-1\right)} \geq 1$ and $\lambda\beta\mathcal{G}(V) = \mathcal{J}(V) - (1-\lambda)\mathcal{G}(V) = \beta V - (1-\lambda)\,\mathcal{G}(V)$ so that

$$\mathcal{G}(V) = \frac{\beta}{1+\lambda(\beta-1)} V$$

and $\frac{\beta}{1+\lambda(\beta-1)}$ is an eigenvalue of $\mathcal{G}$ with associated eigenvector $V$. But

$$\left|\frac{\beta}{1+\lambda(\beta-1)}\right| \geq \frac{|\beta|}{1+\lambda(|\beta|-1)} \geq 1$$

which is a contradiction with the fact that $r_\sigma(\mathcal{G}) < 1$. Thus $r_\sigma(\mathcal{J}) < 1$. ■

In the next Proposition consider $\mathcal{H}$ as in (8) with $\Gamma_i = A_i + B_i F_i$.

**Proposition 2** *If $\lambda^2 r_\sigma(\mathcal{H}) < 1$ then there exist constants $a > 0$, $b > 0$ such that*

$$\mathbb{E}\left(\|W_i(t)\|^2 \middle|\Sigma_t\right) \leq a\,\|Y(t)\|^2 + b.$$

**Proof.** Since $W_i(t)$ is hermitian, we can find $x \in \mathbb{C}^n$, $\|x\| = 1$, such that $\|W_i(t)\|_2 = |x^* W_i(t) x|$. Next notice that

$$x^*W_i(t)x = \sum_{k=0}^{\infty}\lambda^k\Big(\Big\{\left\|C_{\theta_i(t,k+1)}x(k+1)\right\|^2$$
$$+\left\|D_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)}x(k+1)\right\|^2$$
$$+(1-\lambda)x(k+2)^*Y_{\theta_i(t,k+1)}(t)x(k+2)\Big\}$$
$$-\Big\{E\Big(\left\|C_{\theta_i(t,k+1)}x(k+1)\right\|^2$$
$$+\left\|D_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)}x(k+1)\right\|^2$$
$$+(1-\lambda)x(k+2)^*Y_{\theta_i(t,k+1)}(t)x(k+2)\Big)\Big\}\Big)$$

where $x(k+1) = (A_{\theta_i(t,k)} + B_{\theta_i(t,k)}F_{\theta_i(t,k)})x(k), x(0) = x, \theta_i(t,0) = i$. As seen in Section 2, $E\left(\|x(k)\|^2\right) = \sum_{i=1}^N tr\{V_i(k)\}$, where $V_i(k) = E(x(k)x(k)^*1_{\theta_i(t,k)=i})$, and $V(k+1) = \mathcal{T}(V(k))$, where $\mathit{\Gamma}_i = A_i + B_iF_i$ in (6). Since $r_\sigma(\mathcal{T}) < 1$, we can find $c_0 > 0$, $c_1 \in (0,1)$ such that $\left\|\mathcal{T}^k\right\| \le c_0c_1^k$. Thus writing $c_3 = \|C\|_{max} + \|D\|_{max}\|F\|_{max}$ and noticing that $\sum_{i=1}^N tr\{\mathcal{T}_i^k(V(0))\} \le nN\left\|\mathcal{T}_i^k(V(0))\right\|$ we have that

$$E\Big(\left\|C_{\theta_i(t,k+1)}x(k+1)\right\|^2$$
$$+\left\|D_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)}x(k+1)\right\|^2\Big)$$
$$\le c_3E\Big(\|x(k+1)\|^2\Big)$$
$$= c_3\sum_{i=1}^N tr\{\mathcal{T}_i^{k+1}(V(0))\} \le c_3c_0Nnc_1^{k+1}\|V(0)\|$$
$$= c_3c_0Nc_1^{k+1}\|x\|^2 = c_3c_0Nc_1^{k+1}.$$

and similarly

$$E\left(x(k+2)^*Y_{\theta_i(t,k+1)}x(k+2)\right) \le c_0Nnc_1^{k+2}\|Y(t)\|.$$

Therefore for some constant $c_4 > 0$,

$$\mathbb{E}\Big(\|W_i(t)\|^2|\Sigma_t\Big) \le c_4(1+\|Y(t)\|^2)$$
$$+c_4\left(E\left(\left(\sum_{k=0}^{\infty}\lambda^k\|x(k+1)\|^2\right)^2\right)\right.$$
$$\left.+E\left(\left(\sum_{k=0}^{\infty}\lambda^k\|x(k+2)\|^2\right)^2\right)\|Y(t)\|^2\right).$$

Finally notice from Lemma 1 that

$$E\left(\left(\sum_{k=0}^{\infty}\lambda^k\|x(k+1)\|^2\right)^2\right) \le$$
$$\left(\sum_{k=0}^{\infty}\lambda^k\sqrt{E(\|x(k+1)\|^4)}\right)^2$$
$$\le \left(\frac{1}{\lambda}Nn^2\sum_{k=0}^{\infty}\sqrt{\lambda^{2(k+1)}\left\|\mathcal{H}^{k+1}\right\|}\right)^2.$$

If $\lambda^2r_\sigma(\mathcal{H}) < 1$ then we can find $c_5 > 0$, $c_6 \in (0,1)$ such that $\lambda^{2k}\left\|\mathcal{H}^k\right\| \le c_5c_6^k$ and

$$\sum_{k=0}^{\infty}\sqrt{\lambda^{2(k+1)}\left\|\mathcal{H}^{k+1}\right\|} \le c_5^{\frac{1}{2}}\sum_{k=0}^{\infty}\left(c_6^{\frac{1}{2}}\right)^{k+1} = c_7.$$

Thus for some $c_8 > 0$,

$$c_4\left(E\left(\left(\sum_{k=0}^{\infty}\lambda^k\|x(k+1)\|^2\right)^2\right)\right.$$
$$\left.+E\left(\left(\sum_{k=0}^{\infty}\lambda^k\|x(k+2)\|^2\right)^2\right)\|Y(t)\|^2\right)$$
$$\le c_8\left(1+\|Y(t)\|^2\right)$$

and the result follows. ■

Thus (12) is satisfied for each $W_i(t)$, $i = 1,\ldots,N$. The next result shows that through a linear transformation in the algorithm (19) we can assume that $\|\mathcal{J}\| < 1$. First we need the following lemma.

**Lemma 6** *Let $\mathbb{X}$ be a Hilbert space and $\mathcal{T} \in \mathbb{B}(\mathbb{X})$. The following assertions are equivalent:*

*a) $r_\sigma(\mathcal{T}) < 1$*
*b) there exist $\mathcal{W} \in \mathbb{B}(\mathbb{X})$ invertible and $\mathcal{Q} \in \mathbb{B}(\mathbb{X})$ such that $\|\mathcal{Q}\| < 1$ and $\mathcal{T} = \mathcal{W}\mathcal{Q}\mathcal{W}^{-1}$.*

**Proof.** Corollary 1.14 of [13], pages 31-32. ■

**Proposition 3** *There is no loss of generality in assuming that $\|\mathcal{J}\| < 1$.*

**Proof.** Suppose this is not the case. Since $r_\sigma(\mathcal{J}) < 1$ (Proposition 1), we know from Lemma 6 that there exist $\mathcal{W} \in \mathbb{B}(\mathbb{X})$ invertible and real and $\widetilde{\mathcal{J}} \in \mathbb{B}(\mathbb{X})$ real such that $\left\|\widetilde{\mathcal{J}}\right\| < 1$ and $\mathcal{J} = \mathcal{W}\widetilde{\mathcal{J}}\mathcal{W}^{-1}$. Consider the following transformation in the algorithm (18)

$$\widetilde{Y}(t) = \mathcal{W}Y(t),\ \widetilde{\mathcal{D}}_i(t,k,\cdot) = \widetilde{\mathcal{B}}_i(t,k) + \widetilde{\mathcal{C}}_i(t,k,\cdot),$$
$$\widetilde{\mathcal{B}}_i(t,k) = \mathcal{W}\mathcal{B}_i(t,k),\ \widetilde{\mathcal{C}}_i(t,k,\cdot) = \mathcal{W}\mathcal{C}_i(t,k,\cdot)\mathcal{W}^{-1}.$$

Then from (18) it follows that

$$\widetilde{Y}(t+1) = \widetilde{Y}(t) + \gamma(t)\sum_{k=0}^{\infty}\lambda^k\widetilde{\mathcal{D}}\left(t,k,\widetilde{Y}(t)\right) \qquad (20)$$

which can be rewritten as

$$\widetilde{Y}(t+1) = (1-\gamma(t))\,\widetilde{Y}(t) \\ + \gamma(t)\left(\widetilde{Z} + \widetilde{J}\left(\widetilde{Y}(t)\right) + \widetilde{W}(t)\right) \qquad (21)$$

where

$$\widetilde{Z} = \mathcal{W}Z, \quad \widetilde{W}(t) = \mathcal{W}W(t).$$

It is easy to check that $\mathbb{E}\left(\widetilde{W}_i(t)|\widetilde{\Sigma}_t\right) = 0$ and $\mathbb{E}\left(\left\|\widetilde{W}_i(t)\right\|^2 |\widetilde{\Sigma}_t\right) \leq \widetilde{a}\left\|\widetilde{Y}(t)\right\|^2 + \widetilde{b}$ for some $\widetilde{a} > 0$ ,$\widetilde{b} > 0$ where $\widetilde{\Sigma}_t$ is defined appropriately with $\widetilde{Y}(s)$ replacing $Y(s)$. Thus if $\|\mathcal{J}\| \geq 1$ we could apply the linear transformation above and deal with the representation of the algorithm given by (20), (21) instead of (18), (19). ■

Thus from now on we shall assume without loss of generality that $\|\mathcal{J}\| < 1$ in (19). The next results follow the arguments presented in [3], pages 162-167.

**Proposition 4** *If $\lambda^2 r_\sigma(\mathcal{H}) < 1$ then the sequence $\{Y(t); t = 1, 2, \ldots\}$ is bounded with probability 1.*

**Proof.** This proof follows the same steps as the proof of Proposition 4.7 in [3], pages 162-166. Since $\|\mathcal{J}\| < 1$ we can find $G$ such that $G \geq 1$ and $G > \frac{\|Z\|}{1-\|\mathcal{J}\|}$, and define $\eta$ and $\epsilon$ as $\|\mathcal{J}\| < \eta := \frac{\|Z\|}{G} + \|\mathcal{J}\| \in (0,1)$, $\epsilon := \frac{1}{\eta} - 1 > 0$. As in [3], page 163, the non-decreasing $\Sigma_t$-adapted sequence $\{G(t); t = 1, 2, \ldots\}$ is defined as follows: $G(0) = max\{\|Y(0)\|, G\}$ and

$$G(t+1) = \begin{cases} G(t), & \text{if } \|Y(t+1)\| \leq (1+\epsilon)G(t) \\ G(0)(1+\epsilon)^\tau, & \text{otherwise} \end{cases}$$

where $\tau = min\{s \geq 0; \|Y(t+1)\| \leq (1+\epsilon)^s G(0)\}$. From this definition it follows that for all $t = 1, 2, \ldots$, $\|Y(t)\| \leq (1+\epsilon)G(t)$ and $\|Y(t)\| \leq G(t)$ if $G(t-1) < G(t)$. Next we define $\widetilde{W}(t) = \frac{1}{G(t)}W(t)$, so that

$$\mathbb{E}\left(\widetilde{W}_i(t)|\Sigma_t\right) = \frac{1}{G(t)}\mathbb{E}\left(W_i(t)|\Sigma_t\right) = 0 \qquad (22)$$

and for $c = a(1+\epsilon)^2 + b$ (see Proposition 2)

$$\mathbb{E}\left(\left\|\widetilde{W}_i(t)\right\|^2 |\Sigma_t\right) \leq \frac{1}{G(t)^2}\left(a\left\|Y(t)\right\|^2 + b\right) \\ \leq \frac{1}{G(t)^2}\left(a(1+\epsilon)^2 G(t)^2 + b\right) \leq c. \qquad (23)$$

Defining for $i = 1, \ldots, N$, $t \geq t_0 \geq 1$, $R_i(t_0, t_0) = 0$ and $R_i(t+1, t_0) = (1-\gamma(t))R_i(t, t_0) + \gamma(t)\widetilde{W}_i(t)$ we have from Lemma 5, (22) and (23), that $R_i(t, t_0)$ goes to zero as $t$ goes to infinity with probability 1. Consider the set $\Gamma \subset \Sigma$ such that (11) holds and $R_i(t, t_0)$ goes to zero as $t$ goes to infinity for every $t_0 = 1, 2, \ldots$, $i = 1, \ldots, N$. It is easy to check that $\Gamma$ has probability 1 (it is the countable intersection of sets with probability 1). Suppose by contradiction that $Y(t)(\omega)$, $\omega \in \Gamma$, is unbounded, so that $G(t)(\omega) \to \infty$ as $t \to \infty$ and $\|Y(t)(\omega)\| \leq G(t)(\omega)$ for infinitely often $t$. Let $u(\omega) \geq 1$ be such that $\gamma(s)(\omega) < 1$ for $s \geq u(\omega)$. Consider $t_0(\omega) \geq u(\omega)$ such that $\|Y(t_0)(\omega)\| \leq G(t_0)(\omega)$ and $\|R_i(t, 0)(\omega)\| \leq \frac{\epsilon}{2\sqrt{N}}$ for $i = 1, \ldots, N$, $t \geq t_0(\omega) \geq u(\omega)$. Set $R(t, t_0)(\omega) = (R_1(t, t_0)(\omega), \ldots, R_N(t, t_0)(\omega)) \in \mathbb{H}^n$. As in [3] it can be shown by induction that for $t \geq t_0(\omega)$

$$\|Y(t)(\omega) - R(t, t_0)(\omega)G(t_0)(\omega)\| \leq G(t_0)(\omega)$$

and $G(t)(\omega) = G(t_0)(\omega)$. But this is a contradiction with the fact that $G(t)(\omega)$ is going to infinity, proving the result. ■

We can now prove Theorem 1. This proof follows the same steps as the proof of Proposition 4.5 in [3], pages 166-167.

**Proof.** Theorem 1: First notice that from (17),

$$S = Z + \mathcal{J}(S) = (1-\gamma(t))S + \gamma(t)(Z + \mathcal{J}(S))$$

and therefore defining $X(t) = Y(t) - S$, we have

$$X(t+1) = (1-\gamma(t))X(t) + \gamma(t)(\mathcal{J}(X(t)) + W(t)).$$

Defining for $i = 1, \ldots, N$, $t \geq t_0 \geq 1$, $R_i(t_0, t_0) = 0$ and $R_i(t+1, t_0) = (1-\gamma(t))R_i(t, t_0) + \gamma(t)W_i(t)$ we have from Lemma 5, Propositions 2 and 4 that $R_i(t, t_0)$ goes to zero as $t$ goes to infinity with probability 1. Let $\Lambda \subset \Sigma$ be the set such that the sequence $\{Y(t); t = 1, 2, \ldots\}$ is bounded and $R_i(t, t_0)$ goes to zero as $t$ goes to infinity for every $i = 1, \ldots, N$, $t_0 \geq 1$. Since this set is the countable intersection of sets with probability one, we have that $\mathbb{P}(\Lambda) = 1$. Let us write $R(t, t_0) = (R_1(t, t_0), \ldots, R_N(t, t_0)) \in \mathbb{H}^n$. For each $\omega \in \Lambda$ there exists $d(\omega)$ such that for all $t \geq 1$, $\|X(t)(\omega)\| \leq \|Y(t)(\omega)\| + \|S\| \leq d(\omega)$. Set $\nu > 0$ such that $\|\mathcal{J}\| + \nu < 1$ and $d_{k+1} = (\|\mathcal{J}\| + \nu)\, d_k$, $d_0 = d$. Let us take $t_0(\omega) = u(\omega)$ (where $u(\omega)$ is such that $\gamma(s)(\omega) < 1$ for $s \geq u(\omega)$) and prove that there is always $t_{k+1}(\omega) \geq t_k(\omega)$ such that

$$\|X(t)(\omega)\| \leq d_k(\omega), \qquad t \geq t_k(\omega). \tag{24}$$

For simplicity we shall suppress $\omega$ from now on. For $t = 0$ the result is immediate. Suppose (24) holds for $k$. Consider the sequence $y(t+1) = (1-\gamma(t))y(t) + \gamma(t)\,\|\mathcal{J}\|\, d_k$, $t \geq t_k, y(t_k) = d_k$. As in [3] it can be shown by induction that

$$\|X(t) - R(t,t_k)\| \leq y(t), \qquad t \geq t_k. \tag{25}$$

Since $y(t)$ converges to $\|\mathcal{J}\|\, d_k$ and $R(t,t_k)$ goes to 0 as $t$ goes to infinity, we can find $t_{k+1} \geq t_k$ such that $y(t) \leq \left(\|\mathcal{J}\| + \frac{\nu}{2}\right) d_k$ and $\|R(t,t_k)\| \leq \frac{\nu}{2} d_k$ for all $t \geq t_{k+1}$. Thus from (25) we have for all $t \geq t_{k+1}$, $\|X(t)\| \leq \|X(t) - R(t,t_k)\| + \|R(t,t_k)\| \leq y(t) + \frac{\nu}{2} d_k \leq (\|\mathcal{J}\| + \nu)\, d_k = d_{k+1}$ proving (24) for $k+1$. Since $d_k$ goes to zero as $k$ goes to infinity, the result follows. ∎

## 5 Numerical Example

To illustrate the results developed in the previous sections, we consider an economic system based on Samuelson's multiplier-accelerator model [4] which appears as in (1). We considered the following three-operating modes, corresponding to the states 1-Normal, 2-Boom or 3-Slump, for the economy:

$$A_1 = \begin{bmatrix} 0 & 1 \\ -2.5 & 3.2 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 1 \\ -4.3 & 4.5 \end{bmatrix}, A_3 = \begin{bmatrix} 0 & 1 \\ 5.3 & -5.2 \end{bmatrix}$$

$$Q_1 = \begin{bmatrix} 3.6 & -3.8 \\ -3.8 & 4.87 \end{bmatrix}, Q_2 = \begin{bmatrix} 10 & -3 \\ -3 & 8 \end{bmatrix}, Q_3 = \begin{bmatrix} 5 & -4.5 \\ -4.5 & 4.5 \end{bmatrix}$$

and $D_1^* D_1 = 2.6$, $D_2^* D_2 = 1.165$, $D_3^* D_3 = 1.111$, $B_1^* = B_2^* = B_3^* = \begin{bmatrix} 0 & 1 \end{bmatrix}$. We assume that the state of the economy (normal, boom or slump) has been observed over a past period of time, and we use these observations in our iterative algorithm. Since this is just a theoretical example, we generated the past observations considering a discrete time state transition probability matrix $\mathcal{P}$ as follows: $p_{11} = 0.67$, $p_{12} = 0.17$, $p_{13} = 0.16$, $p_{21} = 0.30$, $p_{22} = 0.47$, $p_{23} = 0.23$, $p_{31} = 0.26$, $p_{32} = 0.10$, $p_{33} = 0.64$. It is important to stress however that we would be considering the case in which we don't simulate the sample path of the Markov chain but instead we use the past observations of the Markov chain to obtain the optimal control law of the problem. We have performed the algorithm with 3 different values of $\lambda$: $\lambda = 0.1$, $\lambda = 0.5$, and $\lambda = 0.9$. For all 3 cases final convergence to the optimal gain controller took place after $\ell = 20$ iterations. We define the error as follows:

$$\Delta_i = \left\| \frac{F_i^{real} - F_i}{F_i^{real}} \right\| * 100.$$

where $F_i^{real}$ is the optimal feedback gain controller associated to mode $i$. For $\lambda = 0.5$ we obtained the following values: $\Delta_1 = [0.3123\ 0.2707]$, $\Delta_2 = [0.0377\ 0.5293]$, $\Delta_3 = [0.0070\ 0.5351]$. For the other values of $\lambda$ we obtained similar results. Notice that we could evaluate the error due to the fact that we knew the transition probability matrix, and thus $F_i^{real}$. In a real situation it would not be possible to make this estimation since $\mathcal{P}$ would not be known. Also in a real situation we might have to try several different values of $\lambda$ to have the convergence since we need to have $\lambda^2 r_\sigma(\mathcal{H}) < 1$. As pointed out in remark 2, Lemma 2 establishes a sufficient condition for $\lambda$ so that $\lambda^2 r_\sigma(\mathcal{H}) < 1$ will hold.

## 6 Conclusion

In this paper we have traced a parallel between the Monte Carlo $TD(\lambda)$-simulation method for Markovian decision processes (MDP for short) with a Monte Carlo $TD(\lambda)$-simulation like algorithm for obtaining the optimal control associated to the set of coupled algebraic Riccati equations (CARE for short) for the optimal control of discrete-time Markovian jump linear systems. It is assumed that the transition probability matrix $\mathcal{P}$ is not known, but either there is a sample of past observations of the Markov chain that can be used for the iterative algorithm, or it is possible to simulate trajectories for the Markov chain $\theta(k)$. We related the so-called quasi-linearization method presented in Lemma 4 with the policy iteration technique for MDP, involving a policy evaluation part, and a policy improvement part (see Remark 1).

Monte Carlo $TD(\lambda)$-simulation methods have been applied to solve problems related to MDP (see for instance [3],[16]). In the policy iteration technique, the policy evaluation is carried out by Monte Carlo simulations, using temporal difference methods. By applying these ideas to the quasi-liearization method presented in Lemma 4, we obtained an iterative method that traces a close parallel with the $TD(\lambda)$ simulation methods in MDP. It has been shown in Theorem 1 that if $\lambda$ is chosen small enough, convergence of the $TD(\lambda)$ algorithm in the cost evaluation occurs with probability 1.

**Acknowledgments:** The authors would like to thank the helpful comments made by the associated editor and the reviewers.

## References

[1] H. Abou-Kandil, G. Freiling, and G. Jank. On the solution of discrete-time markovian jump linear quadratic control problems. *Automatica*, 31(5):765–768, 1995.

[2] M. Ait-Rami and L. E. Ghaoui. LMI optimization for nonstandard Riccati equations arising in stochastic control. *IEEE Trans. Automatic Control*, 41(11):1666–1671, 1996.

[3] D. P. Bertesekas and J.N. Tsitsilklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[4] W.P. Jr Blair and D.D. Sworder. Feedback control of a class of linear discrete system with jump parameters and quadratic cost criteria. *Int. J. Control*, 21:833–841, 1975.

[5] W. Brewer. Kronecker product and matrix calculus in system theory. *IEEE Trans. Circuits and Systems*, 25:772–781, 1978.

[6] O.L.V. Costa and M.D. Fragoso. Stability results for discrete-time linear systems with markovian jumping parameters. *J. Math. Analysis and Applic*, 179:154–178, 1993.

[7] O.L.V. Costa and M.D. Fragoso. Discrete-time LQ-optimal control problems for infinite markov jump parameter systems. *IEEE Trans. Automat. Control*, 40:2076–2088, 1995.

[8] 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*, 12(2):167–195, 1999.

[9] O.L.V. Costa, J.B.R. Do Val, and J.C. Geromel. A convex programming approach to $\mathcal{H}_2$-control of discrete-time markovian jump linear systems. *Int. J. Control*, 66:557–579, 1997.

[10] Z. Gajic and I. Borno. Lyapunov iterations for optimal control of jump linear systems at steady state. *IEEE Trans. Automat. Control*, 40(11):481–498, 1995.

[11] Y. Ji and H.J. Chizeck. Controllability, observability and discrete-time markovian jump linear quadratic control. *Int. J. Control*, 48:481–498, 1988.

[12] Y. Ji, H.J. Chizeck, X. Feng, and K.A. Loparo. Stability and control of discrete-time jump linear systems. *Control Th. and Adv. Tech*, 7:247–270, 1991.

[13] C.S Kubrusly. *An Introduction to Models and Decompositions in Operator Theory*. Springer Verlag, New York, 1997.

[14] M. Mariton. Almost sure and moments stability of jump linear systems. *Systems and Control Letters*, 11:393–397, 1988.

[15] M. Mariton. *Jump Linear Systems in Automatic Control*. Marcel Dekker, New York, 1990.

[16] R. S. Sutton and A. G. Barto. *Reinforcement Learning - An Introduction*. MIT Press, 1998.

[17] D.D. Sworder and R.O. Rogers. An LQG solution to a control problem whit solar thermal receiver. *IEEE Transactions on Automatic Control*, 28:971–978, 1981.

[18] J.B.R. Do Val, J.C. Geromel, and O.L.V. Costa. Uncoupled Riccati iterations for the linear quadratic control problem of discrete-time markov jump linear systems. *IEEE Trans. Automat. Control*, 43:1727–1733, 1998.

[19] J. Weidman. *Linear Operators in Hilbert Spaces*. Springer Verlag, New York, 1980.