PERGAMON

Automatica 38 (2002) 217–225

automatica

www.elsevier.com/locate/automatica

Monte Carlo *TD*(λ)-methods for the optimal control of discrete-time Markovian jump linear systems[☆]

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

Departamento de Engenharia de Telecomunicações e Controle, Escola Politécnica da Universidade de São Paulo, CEP: 05508 900 São Paulo SP Brazil

Received 15 June 2000; revised 7 February 2001; received in final form 22 August 2001

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*(λ) algorithms for Markovian decision processes to develop a *TD*(λ) 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. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: *TD*(λ) 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 $(\Omega, P, \{\mathscr{F}_k\}, \mathscr{F})$:

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

$$x(0) = x_0 \quad \theta(0) = \theta_0. \tag{1}$$

Here, $\theta(k)$ is a Markov chain taking values in $\{1,\ldots,N\}$ with transition probability matrix $\mathscr{P} = (p_{ij})$ and $u = (u(0),\ldots)$ is the control input sequence assumed to belong to $\mathscr{U} = \{u = (u(0),\ldots); u(k)$ is $\mathscr{F}_k$-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, Abou-Kandil, Freiling, & Jank, 1995; Ait-Rami & Ghaoui, 1996; Blair & Sworder, 1975; Costa & Fragoso, 1993, 1995; Costa & Marques, 1999; Costa, Do Val, & Geromel, 1997; Gajic & Borno, 1995; Ji & Chizeck, 1988; Ji, Chizeck, Feng, & Loparo, 1991; Mariton, 1988, 1990; Sworder & Rogers, 1981; Do Val, Geromel, & Costa, 1998).

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

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

and it is desired to minimize (2) over $u \in \mathscr{U}$. It has been shown in the literature (cf. Costa & Fragoso, 1995; Ji & Chizeck, 1988; Ji et al., 1991) that the solution of this problem is associated to the existence of a solution $P = (P_1,\ldots,P_N)$, $P_i \geqslant 0$, $i = 1,\ldots,N$, to the following set of coupled algebraic Riccati equations (CARE), for $i = 1,\ldots,N$

$$X_i = A_i^*\mathscr{E}_i(X)A_i + C_i^*C_i - A_i^*\mathscr{E}_i(X)B_i(B_i^*\mathscr{E}_i(X)B_i + D_i^*D_i)^{-1}B_i^*\mathscr{E}_i(X)A_i, \tag{3}$$

[☆] This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor R. Srikant under the direction of Editor Tamer Basar. This work was supported 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.

[*] Corresponding author.

E-mail addresses: oswaldo@lac.usp.br (O.L.V. Costa), julio@lac.usp.br (J.C.C Aya).

PII: S0005-1098(01)00215-1

where $\mathscr{E}(X)=(\mathscr{E}_1(X),\ldots,\mathscr{E}_N(X))$ is defined, for $X=(X_1,\ldots,X_N)$ as $\mathscr{E}_i(X)=\sum_{j=1}^{N} p_{ij}X_j$. If such a solution P exists, it can be shown (see Costa & Fragoso, 1995) that the optimal control law for the problem posed by Eqs. (1)–(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

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

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 linear matrix inequalities (LMI) formulations have been proposed in the current literature (see Abou-Kandil et al., 1995; Ait-Rami & Ghaoui, 1996; Blair & Sworder, 1975; Costa & Fragoso, 1995; Costa & Marques, 1999; Costa et al., 1997; Gajic & Borno, 1995; Ji & Chizeck, 1988; Ji et al., 1991; Mariton, 1990; Do Val et al., 1998). All these methods assume that the transition probability matrix $\mathscr{P}$ 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 $\mathscr{P}$ is not known. We assume that either past observations of the Markov chain $\theta(k)$ are available or that, like in Bertesekas and Tsitsilklis (1996), Sutton and Barto (1998), 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*(λ) methods (see Bertesekas & Tsitsilklis (1996); Sutton and Barto (1998)) for Markovian decision processes (MDP).

As pointed out in Bertesekas and Tsitsilklis (1996), Sutton and Barto (1998), 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 Sutton & Barto, 1998, pp. 112–116). *TD*(λ) methods have been applied to solve problems related to MDP (see for instance Bertesekas & Tsitsilklis, 1996; Sutton & Barto, 1998) 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. Bertesekas & Tsitsilklis, 1996; Sutton & Barto, 1998) 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*(λ) 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 (Sutton & Barto, 1998, p. 115). Therefore our technique would be useful in situations similar to those presented (Bertesekas & Tsitsilklis, 1996; Sutton & Barto, 1998), 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 Bertesekas and Tsitsilklis (1996). 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 Bertesekas and Tsitsilklis (1996), based on the fact that either the MDP has a discount factor α 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*(λ) 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 $\mathscr{T}\in\mathbb{B}(\mathbb{X})$ will be denoted by $r_\sigma(\mathscr{T})$. If $\mathbb{X}$ is a Hilbert space then the inner product will be denoted by $\langle\cdot;\cdot\rangle$, and for $\mathscr{T}\in\mathbb{B}(\mathbb{X})$, $\mathscr{T}^*$ will denote the adjoint operator of $\mathscr{T}$. As usual, $\mathscr{T}\geqslant 0$ ($\mathscr{T}>0$ respectively) will denote that the operator $\mathscr{T}\in\mathbb{B}(\mathbb{X})$ 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 $\langle H;V\rangle=\operatorname{tr}\{H^*V\}$ for $H,V\in\mathbb{B}(\mathbb{C}^n)$ (and $\|H\|^2=\operatorname{tr}\{H^*H\}$). 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 $H=(H_1,\dots,H_N)$, $V=(V_1,\dots,V_N)\in\mathbb{H}^n$ we shall define $\langle\cdot;\cdot\rangle$ in $\mathbb{H}^n$ as follows: $\langle H;V\rangle := \sum_{i=1}^N \operatorname{tr}\{H_i^* V_i\}$, 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 $H=(H_1,\dots,H_N)\in\mathbb{H}^n$, set $\|H\|_{\max} := \max\{\|H_i\|_2;\ i=1,\dots,N\}$ and $\|H\|_1 := \sum_{i=1}^N \|H_i\|_2$. We set $\mathbb{H}^{n+} := \{V=(V_1,\dots,V_N)\in\mathbb{H}^n;\ V_i\geqslant 0,\ i=1,\dots,N\}$ and shall write, for $V=(V_1,\dots,V_N)\in\mathbb{H}^n$ and $S=(S_1,\dots,S_N)\in\mathbb{H}^n$, that $V>S$ if $V_i-S_i>0$ for $i=1,\dots,N$ (similarly for $\geqslant$). For $\Gamma=(\Gamma_1,\dots,\Gamma_N)\in\mathbb{H}^n$, we define the following operators $\mathscr{L}(\cdot)=(\mathscr{L}_1(\cdot),\dots,\mathscr{L}_N(\cdot)), \mathscr{T}(\cdot)=(\mathscr{T}_1(\cdot),\dots,\mathscr{T}_N(\cdot))\in\mathbb{B}(\mathbb{H}^n)$ for $V=(V_1,\dots,V_N)\in\mathbb{H}^n$

$$\mathscr{L}_i(V) := \Gamma_i^*\mathscr{E}_i(\cdot)\Gamma_i, \quad = \Gamma_i^*\sum_{j=1}^N p_{ij}V_j\Gamma_i \tag{5}$$

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

where the operator $\mathscr{E}(\cdot)=(\mathscr{E}_1(\cdot),\dots,\mathscr{E}_N(\cdot))\in\mathbb{B}(\mathbb{H}^n)$ is defined as in Section 1. It is easy to verify that the operators $\mathscr{E}$, $\mathscr{L}$, and $\mathscr{T}$ map $\mathbb{H}^{n+}$ into $\mathbb{H}^{n+}$, and it has been proved in Costa and Fragoso (1993) that $r_\sigma(\mathscr{L})=r_\sigma(\mathscr{T})$ (in fact, $\mathscr{L}=\mathscr{T}^*$ according to the inner product defined above). We also define the operator $\mathscr{G}\in\mathbb{B}(\mathbb{H}^n)$ as

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

where again $V=(V_1,\dots,V_N)\in\mathbb{H}^n$ and $\mathscr{G}(V)=(\mathscr{G}_1(V),\dots,\mathscr{G}_1(V))$. It has been shown in Costa and Fragoso (1993) that $r_\sigma(\mathscr{G})=r_\sigma(\mathscr{L})=r_\sigma(\mathscr{T})$.

We assume in (1) and (2) that $A=(A_1,\dots,A_N)\in\mathbb{H}^n$, $B=(B_1,\dots,B_N)\in\mathbb{H}^{m,n}$, $C=(C_1,\dots,C_N)\in\mathbb{H}^{n,p}$ and $D=(D_1,\dots,D_N)\in\mathbb{H}^{m,p}$. It has been shown in Costa and Fragoso (1993), Mariton (1988), that model (1) with $u(k)=F_{\theta(k)}x(k)$, and $V_i(k)=E(x(k)x(k)^* 1_{\{\theta(k)=i\}})$, $V(k)=(V_1(k),\dots,V_N(k))\in\mathbb{H}^{n+}$ leads to $V(k+1)=\mathscr{T}(V(k))$, $k=0,1,\dots$, where $\Gamma_i=A_i+B_iF_i$ in (6), and $E(\|x(k)\|^2)=\sum_{i=1}^N \operatorname{tr}\{V_i(k)\}$. In what follows, we shall also need the operator for the fourth 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 Brewer, 1978), let the operator $\mathscr{H}\in\mathbb{B}(\mathbb{H}^{n^2})$ be as follows: for $S=(S_1,\dots,S_N)\in\mathbb{H}^{n^2}$, $\mathscr{H}(S)=(\mathscr{H}_1(S),\dots,\mathscr{H}_N(S))$ is defined as

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

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

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

Lemma 2. *If $\lambda^2 < 1/\|\bar{\Gamma}\otimes\Gamma\|_{\max}$, then $\lambda^2 r_\sigma(\mathscr{H})<1$.*

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

Definition 1. We say that $F=(F_1,\dots,F_N)\in\mathbb{H}^{n,m}$ 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 $F=(F_1,\dots,F_N)\in\mathbb{H}^{n,m}$, we have that F stabilizes (A,B) in the mean square sense.

The following result, proved in Costa and Fragoso (1993), shows that $F=(F_1,\dots,F_N)$ 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. *$F=(F_1,\dots,F_N)\in\mathbb{H}^{n,m}$ stabilizes (A,B) in the mean square sense if and only if $r_\sigma(\mathscr{T})<1$, where $\mathscr{T}$ is as in (6) with $\Gamma_i=A_i+B_iF_i$.*

We make the following definition:

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

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

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

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

$$\begin{aligned}
&(A_i+B_iF_i)^*\mathscr{E}_i(S)(A_i+B_iF_i)+F_i^*D_i^*D_iF_i\\
&\quad=(A_i+B_i\mathscr{F}_i(S))^*\mathscr{E}_i(S)(A_i+B_i\mathscr{F}_i(S))\\
&\qquad+\mathscr{F}_i(S)^*D_i^*D_i\mathscr{F}_i(S)+(F_i-\mathscr{F}_i(S))^*\\
&\qquad\times(B_i^*\mathscr{E}_i(S)B_i+D_i^*D_i)(F_i-\mathscr{F}_i(S)).
\end{aligned} \tag{10}$$

The next lemma, proved in Costa and Marques (1999), 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) $P^0\geqslant P^1\geqslant\cdots\geqslant P^l\geqslant X$, *for arbitrary* $X\in\mathbb{H}^{n+}$ *such that* $X\geqslant\mathscr{R}(X)$.
(b) $r_\sigma(\mathscr{L}^l)<1$, *where* $\mathscr{L}^l(\cdot)=(\mathscr{L}_1^l(\cdot),\ldots,\mathscr{L}_N^l(\cdot))$ *and for* $i=1,\ldots,N$,

$$\mathscr{L}_i^l(\cdot):=A_i^{l^*}\mathscr{E}_i(\cdot)A_i^l,\quad A_i^l:=A_i+B_iF_i^l,$$

$$F_i^l:=\mathscr{F}_i(P^{l-1})\quad for\ l=1,2,\ldots\ .$$

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

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

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

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

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 Σ_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, p. 161 in Bertesekas and Tsitsilklis (1996). □

3. *TD*(*λ*) algorithm

If the transition probability matrix $\mathscr{P}$ 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 $\mathscr{P}$ 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 $F^{l+1}=\mathscr{F}(P^l)$ since, as can be seen from (9), it depends on $\mathscr{P}$ through the operator $\mathscr{E}(\cdot)$. An alternative way would be to calculate directly $S^l:=\mathscr{E}(P^l)$, which would lead us to the following equation: $S^l=\mathscr{G}^l(S^l)+\mathscr{E}(\mathscr{V}(F^l))$. Let us write for simplicity $Q^l=\mathscr{E}(\mathscr{V}(F^l))$. Note that as seen in Section 2, $r_\sigma(\mathscr{G}^l)=r_\sigma(\mathscr{L}^l)<1$, and Theorem 5.17, p. 102 in Weidman (1980) can be applied to say that the equation in Y

$$Y=\mathscr{G}^l(Y)+Q^l \qquad (14)$$

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

$$F_i^{l+1}=-(B_i^*S_i^lB_i+D_i^*D_i)^{-1}B_i^*S_i^lA_i. \qquad (15)$$

The remaining of this section is now devoted to calculating S^l through Monte Carlo simulations, tracing a parallel with *TD*(λ) methods (see Bertesekas & Tsitsilklis, 1996). For simplicity, we shall suppress the superscript l. For any $\lambda\in[0,1)$, define the operator $\mathscr{J}\in\mathbb{B}(\mathbb{H}^n)$ and $Z\in\mathbb{H}^n$ in the following way:

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

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

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

We have from (16) that

$$S=\sum_{k=0}^{\infty}(1-\lambda)\lambda^k S$$
$$=\sum_{t=0}^{\infty}(\lambda\mathscr{G})^t(Q)+(1-\lambda)\sum_{k=0}^{\infty}\lambda^k\mathscr{G}^{k+1}(S)$$
$$=Z+\mathscr{J}(S)$$
$$=\sum_{k=0}^{\infty}\lambda^k[\mathscr{G}^k(Q)+\mathscr{G}^{k+1}(S)-\mathscr{G}^k(S)]+S. \tag{17}$$

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

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

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

$$\Upsilon_i(t,k):=(A_{\theta_i(t,k)}+B_{\theta_i(t,k)}F_{\theta_i(t,k)})\Upsilon_i(t,k-1),$$
$$\Upsilon_i(t,0):=I,$$
$$\mathscr{D}_i(t,k,V):=\mathscr{B}_i(t,k)+\mathscr{C}_i(t,k,V),$$
$$\mathscr{B}_i(t,k):=\Upsilon_i(t,k)^*[C^*_{\theta_i(t,k+1)}C_{\theta_i(t,k+1)}$$
$$+F^*_{\theta_i(t,k+1)}D^*_{\theta_i(t,k+1)}D_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)}]\Upsilon_i(t,k),$$
$$\mathscr{C}_i(t,k,V):=\Upsilon_i(t,k)^*[(A_{\theta_i(t,k+1)}$$
$$+B_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)})^*V_{\theta_i(t,k+1)}(A_{\theta_i(t,k+1)}$$
$$+B_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)})-V_{\theta_i(t,k)}]\Upsilon_i(t,k),$$
$$D(t,k,V):=(\mathscr{D}_1(t,k,V),\dots,\mathscr{D}_N(t,k,V)).$$

Notice that

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

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

$$W_i(t):=\sum_{k=0}^{\infty}\lambda^k\Bigg(\mathscr{D}_i(t,k,Y(t))$$
$$-(\mathscr{G}_i^k(Q)+\mathscr{G}_i^{k+1}(Y(t))-\mathscr{G}_i^k(Y(t)))\Bigg).$$

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

$$Y(t+1)=(1-\gamma(t))Y(t)$$
$$+\gamma(t)(Z+\mathscr{J}(Y(t))+W(t)). \tag{19}$$

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

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

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

$$\mathbb{E}(W_i(t)|\Sigma_t)=\sum_{k=0}^{\infty}\lambda^k\Bigg(E(\mathscr{D}_i(t,k,Y(t))|\theta(0)=i)$$
$$-(\mathscr{G}_i^k(Q)+\mathscr{G}_i^{k+1}(Y(t))-\mathscr{G}_i^k(Y(t)))\Bigg)=0,$$

satisfying Eq. (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 $\lambda^2 r_\sigma(\mathscr{H})<1$, then the sequence $\{Y(t);\ t=1,2,\dots\}$ converges to S with probability* 1.

Remark 2. Lemma 2 establishes an upperbound for λ so that $\lambda^2 r_\sigma(\mathscr{H})<1$ 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 λ.

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

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

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

4. Proofs

First we present the proofs of Lemmas 1 and 2.

Proof of 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 $z_i(k)=vec\{x_i(k)x_i(k)^*\}$, $i=1,\dots,N$. After some manipulation, it follows that $z_j(k+1)z_j(k+1)^*=1_{\{\theta(k+1)=j\}}\sum_{i=1}^N(\bar{\Gamma}_i\otimes\Gamma_i)z_i(k)z_i(k)^*\,(\Gamma_i'\otimes\Gamma_i^*)$ and writing $S_i(k)=E(z_i(k)z_i(k)^*\,1_{\{\theta(k)=i\}})$, it follows that $S_j(k+1)=\sum_{i=1}^N p_{ij}(\bar{\Gamma}_i\otimes\Gamma_i)S_i(k)(\Gamma_i'\otimes\Gamma_i^*)$. Finally, notice that $\mathrm{tr}\{z_i(k)z_i(k)^*\}=\|x_i(k)\|^2\,\mathrm{tr}\{x_i(k)\,x_i(k)^*\}=\|x_i(k)\|^4$. □

Proof of Lemma 2. For any $S=(S_1,\dots,S_N)\in\mathbb{H}^{n^2}$, $\|\mathscr{H}(S)\|_1=\sum_{j=1}^N\|\mathscr{H}_j(S)\|_2\leqslant\|\bar{\Gamma}\otimes\Gamma\|^2_{\max}\|S\|_1$, and therefore, $\|\mathscr{H}\|_1 = \sup\{\|\mathscr{H}(S)\|_1/\|S\|_1;\ S \in \mathbb{H}^{n^2}\} \leqslant \|\bar{\Gamma} \otimes \Gamma\|_{\max}^2$. Since $r_\sigma(\mathscr{H}) \leqslant \|\mathscr{H}\|_1$ the result follows. □

To prove Theorem 1, we shall need the following propositions:

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

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

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

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

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

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

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

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

$$\mathbb{E}(\|W_i(t)\|^2 | \Sigma_t) \leqslant a\|Y(t)\|^2 + b.$$

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

$$\begin{aligned} x^* W_i(t) x = \sum_{k=0}^{\infty} \lambda^k \Bigg(& \{\|C_{\theta_i(t,k+1)} x(k+1)\|^2 \\ & + \|D_{\theta_i(t,k+1)} F_{\theta_i(t,k+1)} x(k+1)\|^2 \\ & + (1-\lambda) x(k+2)^* Y_{\theta_i(t,k+1)}(t) x(k+2)\} \\ & - \{E(\|C_{\theta_i(t,k+1)} x(k+1)\|^2 \\ & + \|D_{\theta_i(t,k+1)} F_{\theta_i(t,k+1)} x(k+1)\|^2 \\ & + (1-\lambda) x(k+2)^* Y_{\theta_i(t,k+1)}(t) x(k+2))\} \Bigg), \end{aligned}$$

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, $E(\|x(k)\|^2) = \sum_{i=1}^N \operatorname{tr}\{V_i(k)\}$, where $V_i(k) = E(x(k)x(k)^* 1_{\theta_i(t,k)=i})$, and $V(k+1) = \mathscr{T}(V(k))$, where $\Gamma_i = A_i + B_i F_i$ in (6). Since $r_\sigma(\mathscr{T}) < 1$, we can find $c_0 > 0$, $c_1 \in (0,1)$ such that $\|\mathscr{T}^k\| \leqslant c_0 c_1^k$. Thus writing $c_3 = \|C\|_{\max} + \|D\|_{\max}\|F\|_{\max}$ and noticing that $\sum_{i=1}^N \operatorname{tr}\{\mathscr{T}_i^k(V(0))\} \leqslant nN\|\mathscr{T}_i^k(V(0))\|$, we have that

$$\begin{aligned} & E(\|C_{\theta_i(t,k+1)} x(k+1)\|^2 + \|D_{\theta_i(t,k+1)} F_{\theta_i(t,k+1)} x(k+1)\|^2) \\ & \quad \leqslant c_3 E(\|x(k+1)\|^2) \\ & \quad = c_3 \sum_{i=1}^N \operatorname{tr}\{\mathscr{T}_i^{k+1}(V(0))\} \leqslant c_3 c_0 N n c_1^{k+1} \|V(0)\| \\ & \quad = c_3 c_0 N c_1^{k+1} \|x\|^2 = c_3 c_0 N c_1^{k+1}. \end{aligned}$$

and similarly,

$$E(x(k+2)^* Y_{\theta_i(t,k+1)} x(k+2)) \leqslant c_0 N n c_1^{k+2} \|Y(t)\|.$$

Therefore, for some constant $c_4 > 0$,

$$\begin{aligned} & \mathbb{E}w(\|W_i(t)\|^2 | \Sigma_t) \leqslant c_4(1 + \|Y(t)\|^2) \\ & \quad + c_4 \left(E\left(\left(\sum_{k=0}^{\infty} \lambda^k \|x(k+1)\|^2 \right)^2 \right) \right. \\ & \quad \left. + E\left(\left(\sum_{k=0}^{\infty} \lambda^k \|x(k+2)\|^2 \right)^2 \right) \|Y(t)\|^2 \right). \end{aligned}$$

Finally, notice from Lemma 1 that

$$\begin{aligned} & E\left(\left(\sum_{k=0}^{\infty} \lambda^k \|x(k+1)\|^2 \right)^2 \right) \\ & \quad \leqslant \left(\sum_{k=0}^{\infty} \lambda^k \sqrt{E(\|x(k+1)\|^4)} \right)^2 \\ & \quad \leqslant \left(\frac{1}{\lambda} N n^2 \sum_{k=0}^{\infty} \sqrt{\lambda^{2(k+1)} \|\mathscr{H}^{k+1}\|} \right)^2. \end{aligned}$$

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

$$\sum_{k=0}^{\infty} \sqrt{\lambda^{2(k+1)} \|\mathscr{H}^{k+1}\|} \leqslant c_5^{1/2} \sum_{k=0}^{\infty} (c_6^{1/2})^{k+1} = c_7.$$

Thus, for some $c_8 > 0$,

$$\begin{aligned} & c_4 \left(E\left(\left(\sum_{k=0}^{\infty} \lambda^k \|x(k+1)\|^2 \right)^2 \right) \right. \\ & \quad \left. + E\left(\left(\sum_{k=0}^{\infty} \lambda^k \|x(k+2)\|^2 \right)^2 \right) \|Y(t)\|^2 \right) \\ & \quad \leqslant c_8(1 + \|Y(t)\|^2) \end{aligned}$$

and the result follows. □

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

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

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

Proof. Corollary 1.14 of Kubrusly (1997), pp. 31–32. □

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

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

$$\tilde{Y}(t)=\mathscr{W}Y(t),\quad \tilde{\mathscr{D}}_i(t,k,\cdot)=\tilde{\mathscr{B}}_i(t,k)+\tilde{\mathscr{C}}_i(t,k,\cdot),$$

$$\tilde{\mathscr{B}}_i(t,k)=\mathscr{W}\mathscr{B}_i(t,k),\quad \tilde{\mathscr{C}}_i(t,k,\cdot)=\mathscr{W}\mathscr{C}_i(t,k,\cdot)\mathscr{W}^{-1}.$$

Then, from (18), it follows that

$$\tilde{Y}(t+1)=\tilde{Y}(t)+\gamma(t)\sum_{k=0}^{\infty}\lambda^k\tilde{\mathscr{D}}(t,k,\tilde{Y}(t)) \tag{20}$$

which can be rewritten as

$$\tilde{Y}(t+1)=(1-\gamma(t))\tilde{Y}(t)+\gamma(t)(\tilde{Z}+\tilde{J}(\tilde{Y}(t))+\tilde{W}(t)), \tag{21}$$

where

$$\tilde{Z}=\mathscr{W}Z,\quad \tilde{W}(t)=\mathscr{W}W(t).$$

It is easy to check that $\mathbb{E}(\tilde{W}_i(t)|\tilde{\Sigma}_t)=0$ and $\mathbb{E}(\|\tilde{W}_i(t)\|^2|\tilde{\Sigma}_t)\leqslant\tilde{a}\|\tilde{Y}(t)\|^2+\tilde{b}$ for some $\tilde{a}>0$, $\tilde{b}>0$ where $\tilde{\Sigma}_t$ is defined appropriately with $\tilde{Y}(s)$ replacing $Y(s)$. Thus, if $\|\mathscr{J}\|\geqslant 1$ 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 $\|\mathscr{J}\|<1$ in (19). The next results follow the arguments presented in Bertesekas and Tsitsilklis (1996, pp. 162–167).

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

Proof. This proof follows the same steps as the proof of Proposition 4.7 in Bertesekas and Tsitsilklis (1996, pp. 162–166). Since $\|\mathscr{J}\|<1$, we can find G such that $G\geqslant 1$ and $G>\|Z\|/1-\|\mathscr{J}\|$, and define η and ε as $\|\mathscr{J}\|<\eta:=\|Z\|/G+\|\mathscr{J}\|\in(0,1)$, $\varepsilon:=1/\eta-1>0$. As in Bertesekas and Tsitsilklis (1996, p. 163), the non-decreasing Σ_t-adapted sequence $\{G(t);\ t=1,2,\dots\}$ is defined as follows: $G(0)=\max\{\|Y(0)\|,G\}$ and

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

where $\tau=\min\{s\geqslant 0;\|Y(t+1)\|\leqslant(1+\varepsilon)^sG(0)\}$. From this definition, it follows that for all $t=1,2,\dots$, $\|Y(t)\|\leqslant(1+\varepsilon)G(t)$ and $\|Y(t)\|\leqslant G(t)$ if $G(t-1)<G(t)$. Next we define $\tilde{W}(t)=(1/G(t))W(t)$, so that

$$\mathbb{E}(\tilde{W}_i(t)|\Sigma_t)=\frac{1}{G(t)}\mathbb{E}(W_i(t)|\Sigma_t)=0 \tag{22}$$

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

$$\begin{aligned}\mathbb{E}(\|\tilde{W}_i(t)\|^2|\Sigma_t)&\leqslant\frac{1}{G(t)^2}(a\|Y(t)\|^2+b)\\&\leqslant\frac{1}{G(t)^2}(a(1+\varepsilon)^2G(t)^2+b)\leqslant c.\end{aligned} \tag{23}$$

Defining for $i=1,\dots,N$, $t\geqslant t_0\geqslant 1$, $R_i(t_0,t_0)=0$ and $R_i(t+1,t_0)=(1-\gamma(t))R_i(t,t_0)+\gamma(t)\tilde{W}_i(t)$ 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 $t_0=1,2,\dots$, $i=1,\dots,N$. It is easy to check that Γ 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 $\|Y(t)(\omega)\|\leqslant G(t)(\omega)$ for infinitely often t. Let $u(\omega)\geqslant 1$ be such that $\gamma(s)(\omega)<1$ for $s\geqslant u(\omega)$. Consider $t_0(\omega)\geqslant u(\omega)$ such that $\|Y(t_0)(\omega)\|\leqslant G(t_0)(\omega)$ and $\|R_i(t,0)(\omega)\|\leqslant\varepsilon/2\sqrt{N}$ for $i=1,\dots,N$, $t\geqslant t_0(\omega)\geqslant u(\omega)$. Set $R(t,t_0)(\omega)=(R_1(t,t_0)(\omega),\dots,R_N(t,t_0)(\omega))\in\mathbb{H}^n$. As in Bertesekas and Tsitsilklis (1996) it can be shown by induction that for $t\geqslant t_0(\omega)$

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

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 Bertesekas and Tsitsilklis (1996, pp. 166–167).

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

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

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

$$X(t+1)=(1-\gamma(t))X(t)+\gamma(t)(\mathscr{I}(X(t))+W(t)).$$

Defining for $i=1,\dots,N$, $t\geqslant t_0\geqslant 1$, $R_i(t_0,t_0)=0$ 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 $\{Y(t);\ t=1,2,\dots\}$ is bounded and $R_i(t,t_0)$ goes to zero as t goes to infinity for every $i=1,\dots,N$, $t_0\geqslant 1$. Since this set is the countable intersection of sets with probability one, we have that $\mathbb{P}(\Lambda)=1$. Let us write $R(t,t_0)=(R_1(t,t_0),\dots,R_N(t,t_0))\in\mathbb{H}^n$. For each $\omega\in\Lambda$ there exists $d(\omega)$ such that for all $t\geqslant 1$, $\|X(t)(\omega)\|\leqslant\|Y(t)(\omega)\|+\|S\|\leqslant d(\omega)$. Set $v>0$ such that $\|\mathscr{I}\|+v<1$ and $d_{k+1}=(\|\mathscr{I}\|+v)d_k$, $d_0=d$. Let us take $t_0(\omega)=u(\omega)$ (where $u(\omega)$ is such that $\gamma(s)(\omega)<1$ for $s\geqslant u(\omega)$) and prove that there is always $t_{k+1}(\omega)\geqslant t_k(\omega)$ such that

$$\|X(t)(\omega)\|\leqslant d_k(\omega),\quad t\geqslant t_k(\omega). \tag{24}$$

For simplicity, we shall suppress ω from now on. For $t=0$ the result is immediate. Suppose (24) holds for k. Consider the sequence $y(t+1)=(1-\gamma(t))y(t)+\gamma(t)\|\mathscr{I}\|d_k$, $t\geqslant t_k$, $y(t_k)=d_k$. As in Bertesekas and Tsitsilklis (1996), it can be shown by induction that

$$\|X(t)-R(t,t_k)\|\leqslant y(t),\quad t\geqslant t_k. \tag{25}$$

Since $y(t)$ converges to $\|\mathscr{I}\|d_k$ and $R(t,t_k)$ goes to 0 as t goes to infinity, we can find $t_{k+1}\geqslant t_k$ such that $y(t)\leqslant(\|\mathscr{I}\|+v/2)d_k$ and $\|R(t,t_k)\|\leqslant(v/2)d_k$ for all $t\geqslant t_{k+1}$. Thus, from (25) we have for all $t\geqslant t_{k+1}$, $\|X(t)\|\leqslant\|X(t)-R(t,t_k)\|+\|R(t,t_k)\|\leqslant y(t)+(v/2)d_k\leqslant(\|\mathscr{I}\|+v)d_k=d_{k+1}$ 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 (Blair & Sworder, 1975) 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:

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

$$A_3=\begin{bmatrix}0 & 1\\ 5.3 & -5.2\end{bmatrix}$$

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

$$Q_3=\begin{bmatrix}5 & -4.5\\ -4.5 & 4.5\end{bmatrix}$$

and $D_1^*D_1=2.6$, $D_2^*D_2=1.165$, $D_3^*\,D_3=1.111$, $B_1^*=B_2^*=B_3^*=[0\ 1]$. 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 $\mathscr{P}$ 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 do not 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 three different values of λ: $\lambda=0.1$, $\lambda=0.5$, and $\lambda=0.9$. For all three cases, final convergence to the optimal gain controller took place after $\ell=20$ iterations. We define the error as follows:

$$\Delta_i=\left\|\frac{F_i^{\mathrm{real}}-F_i}{F_i^{\mathrm{real}}}\right\|\times 100,$$

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 λ, 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 $\mathscr{P}$ would not be known. Also in a real situation we might have to try several different values of λ to have the convergence since we need to have $\lambda^2 r_\sigma(\mathscr{H})<1$. As pointed out in Remark 2, Lemma 2 establishes a sufficient condition for λ so that $\lambda^2 r_\sigma(\mathscr{H})<1$ 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) 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 the optimal control of discrete-time Markovian jump linear systems. It is assumed that the transition probability matrix $\mathscr{P}$ 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, Bertesekas & Tsitsilklis, 1996; Sutton & Barto, 1998). 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-linearization 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 λ is chosen small enough, convergence of the $TD(\lambda)$ algorithm in the cost evaluation occurs with probability 1.

Acknowledgements

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

References

Abou-Kandil, H., Freiling, G., & Jank, G. (1995). On the solution of discrete-time Markovian jump linear quadratic control problems. *Automatica*, *31*(5), 765–768.

Ait-Rami, M., & Ghaoui, L. E. (1996). LMI optimization for nonstandard Riccati equations arising in stochastic control. *IEEE Transactions on Automatic Control*, *41*(11), 1666–1671.

Bertesekas, D. P., & Tsitsilklis, J. N. (1996). Neuro-dynamic programming. Belmont, MA: Athena Scientific.

Blair Jr., W. P., & Sworder, D. D. (1975). Feedback control of a class of linear discrete system with jump parameters and quadratic cost criteria. *International Journal of Control*, *21*, 833–841.

Brewer, W. (1978). Kronecker product and matrix calculus in system theory. *IEEE Transactions on Circuits and Systems*, *25*, 772–781.

Costa, O. L. V., & Fragoso, M. D. (1993). Stability results for discrete-time linear systems with Markovian jumping parameters. *Journal of Mathematical Analysis and Applications*, *179*, 154–178.

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

Costa, O. L. V., & Marques, R. P. (1999). Maximal and stabilizing Hermitian solutions for discrete-time coupled algebraic Riccati equations. *Mathematics of Control, Signals and Systems*, *12*(2), 167–195.

Costa, O. L. V., Do Val, J. B. R., & Geromel, J. C. (1997). A convex programming approach to $\mathscr{H}_2$-control of discrete-time Markovian jump linear systems. *International Journal of Control*, *66*, 557–579.

Gajic, Z., & Borno, I. (1995). Lyapunov iterations for optimal control of jump linear systems at steady state. *IEEE Transactions on Automatic Control*, *40*(11), 481–498.

Ji, Y., & Chizeck, H. J. (1988). Controllability, observability and discrete-time Markovian jump linear quadratic control. *International Journal of Control*, *48*, 481–498.

Ji, Y., Chizeck, H., Feng, X., & Loparo, K. A. (1991). Stability and control of discrete-time jump linear systems. *Control Theory and Advanced Technology*, *7*, 247–270.

Kubrusly, C. S. (1997). *An introduction to models and decompositions in operator theory*. New York: Springer.

Mariton, M. (1988). Almost sure and moments stability of jump linear systems. *Systems and Control Letters*, *11*, 393–397.

Mariton, M. (1990). *Jump linear systems in automatic control*. New York: Marcel Dekker.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning—an introduction*. Cambridge, MA: MIT Press.

Sworder, D. D., & Rogers, R. O. (~~1981~~ 1983). An ~~LQG~~ LQ solution to a control problem whit solar thermal receiver. *IEEE Transactions on Automatic Control*, *28*, 971–978.

Do Val, J. B. R., Geromel, J. C., & Costa, O. L. V. (1998). Uncoupled Riccati iterations for the linear quadratic control problem of discrete-time markov jump linear systems. *IEEE Transactions on Automatic Control*, *43*, 1727–1733.

Weidman, J. (1980). *Linear operators in Hilbert spaces*. New York: Springer.

Oswaldo Luiz do Valle Costa was born in 1959 in Rio de Janeiro, RJ, Brazil. He obtained his B.Sc. and M.Sc. degrees both in Electrical Engineering from the Catholic University of Rio de Janeiro, Brazil, in 1981 and 1983, respectively, and the Ph.D. degree in Electrical Engineering from Imperial College of Science and Technology in London in 1987. He also held a post-doctoral research assistantship position in the Department of Electrical Engineering at Imperial College from 1987 until 1988. He is presently Professor in the Control Group of the Department of Telecommunications and Control of the Polytechnic School of the University of São Paulo, Brazil. His research interests include stochastic control, optimal control, and jump systems.

Julio Cesar Ceballos Aya was born in Cali, Colombia, in December 1970. He received the B.Sc. degree in Electrical Engineering from "Universidad Autonoma de Occcidente", CaliColombia, in 1995 and the M.Sc. and Ph.D. degrees from "Universidade de São Paulo", São Paulo, Brazil, in 1996, and 2001, respectively. His main research interests are in temporal difference, reinforcement learning, optimal control, with applications to dynamics systems subject to abrupt random variations.