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

Oswaldo L. V. Costa[1] Julio C.C. Aya[2]
Departamento de Engenharia de Telecomunicações e Controle
Escola Politécnica da Universidade de São Paulo
CEP: 05508 900 São Paulo SP Brazil
oswaldo@lac.usp.br, julio@lac.usp.br

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. It is well known that the optimal control of this problem is given in terms of the maximal solution of a set of coupled algebraic Riccati equations (CARE), which have been extensively studied over the last few years. 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 the CARE. Some numerical examples are also presented.

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

1 Introduction

In this paper we consider the following class of models in an appropriate probabilistic space $(\Omega, P, \{\mathcal{F}_k\}, \mathcal{F})$, known in the international literature as discrete-time Markovian jump linear systems (cf. [11]):

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

Here $\theta(k)$ is a Markov chain taking values in $\{1, \ldots, N\}$ with transition probability matrix $\mathcal{P} = (p_{ij})$. Let $\Sigma = \{u = (u(0), \ldots); u(k)$ is $\mathcal{F}_k$-measurable for each $k\}$. For $u \in \Sigma$, 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 $u \in \Sigma$. It has been shown in the literature (cf. [6], [8], [9]) 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 $i = 1, \ldots, N$

$$\begin{aligned} X_i &= A_i'\mathcal{E}_i(X)A_i + C_i'C_i \\ &\quad - A_i'\mathcal{E}_i(X)B_i\left(B_i'\mathcal{E}_i(X)B_i + D_i'D_i\right)B_i'\mathcal{E}_i(X)A_i \end{aligned} \tag{3}$$

where $\mathcal{E}(X) = (\mathcal{E}_1(X), \ldots, \mathcal{E}_N(X))$ is defined, for $X = (X_1, \ldots, X_N)$ as

$$\mathcal{E}_i(X) = \sum_{j=1}^{N} p_{ij}X_j. \tag{4}$$

If such a solution P exists, it can be shown (see [6]) that the optimal control law for the problem posed by equations (1), (2) and (3) is given by the feedback control law

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

where $F = (F_1, \ldots, F_N)$ is given by

$$F_i = -(B_i^*\mathcal{E}_i(P)B_i + D_i^*D_i)^{-1}B_i^*\mathcal{E}_i(P)A_i. \tag{6}$$

Tracing a parallel with the theory of Markovian decision processes (MDP for short), we could relate the iterations of the so-called quasi-linearization method for deriving the solution of (3) (see Lemma 5 below) with those in the policy iteration technique for MDP, involving a policy evaluation part, and a policy improvement part.

$TD(\lambda)$ methods have been applied to solve problems related to MDP (see for instance [1],[12]) for the case in which the transition probability matrix of the Markov chain is not known. It is assumed that 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.

The goal of this paper is to apply $TD(\lambda)$ methods for obtaining the optimal control (5),(6) associated to the solution P of the CARE (3), (4) for the case in which the transition probability matrix $\mathcal{P}$ is not known, but it is possible to simulate trajectories for the Markov chain $\theta(k)$. We propose a Monte Carlo policy evaluation like algorithm that incrementally updates the estimates for the optimal control F given by (5),(6). Proof of convergence is obtained by following arguments similar to those in [1].

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

[1]Partially supported by FAPESP (Research Council of the State of São Paulo), CNPq(Brazilian National Research Council) and PRONEX
[2]Supported by CNPq(Brazilian National Research Council)

main result is Theorem 1, in which the proof of convergence is established. Section 4 presents a numerical example and section 5 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 Euclidean norm in $\mathbb{C}^n$ and the induced norm in $\mathbb{B}(\mathbb{C}^n)$.

The following results will be useful in the next sections.

Lemma 1 *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 $\mathcal{W} \in \mathbb{B}(\mathbb{X}^n)$ 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 [10], pages 31-32. ∎

Lemma 2 *Let $\mathbb{X}$ be a Hilbert space and $\mathcal{T} \in \mathbb{B}(\mathbb{X})$. If $r_\sigma(\mathcal{T}) < 1$ then for each $Q \in \mathbb{X}$ there exists a unique solution $S(Q) \in \mathbb{X}$ for the system in X*

$$X - \mathcal{T}(X) = Q. \tag{7}$$

Moreover

$$S(Q) = \sum_{k=0}^{\infty} \mathcal{T}^k(Q). \tag{8}$$

Proof. See Theorem 5.17, page 102 in [13]. ∎

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:

$$\langle H;V\rangle := \sum_{i=1}^{N} tr\{H_i^*V_i\} \tag{9}$$

and $\|H\|^2 := \langle H;H\rangle$ (so that $\|H\|^2 = \sum_{i=1}^N \|H_i\|^2$). Therefore with the above inner product given by (9), $\mathbb{H}^n$ is a Hilbert space. We shall also set the following $\|H\|_{max}$ norm in $\mathbb{H}^n$. For $H = (H_1,\ldots,H_N) \in \mathbb{H}^n$, set

$$\|H\|_{max} := max\{\|H_i\|_2\,; i = 1,\ldots,N\}.$$

We set

$$\mathbb{H}^{n+} := \left\{V = (V_1,\ldots,V_N) \in \mathbb{H}^n; V_i \geq 0,\ i = 1,\ldots,N\right\}$$

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 $V \geq S$ if

$$V - S = (V_1 - S_1,\ldots,V_N - S_N) \in \mathbb{H}^{n+},$$

and that $V > S$ if $V_i - S_i > 0$ for $i = 1,\ldots,N$. For $\Gamma = (\Gamma_1,\ldots,\Gamma_N) \in \mathbb{H}^n$ we define the following operator $\mathcal{L}(\cdot) = (\mathcal{L}_1(\cdot),\ldots,\mathcal{L}_N(\cdot)) \in \mathbb{B}(\mathbb{H}^n)$ for $V = (V_1,\ldots,V_N) \in \mathbb{H}^n$ and $i,j = 1,\ldots,N$,

$$\mathcal{L}_i(V) := \Gamma_i^*\mathcal{E}_i(\cdot)\Gamma_i \tag{10}$$

$$\mathcal{T}_j(V) := \sum_{i=1}^{N} p_{ij}\Gamma_i V_i \Gamma_i^* \tag{11}$$

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 (4). 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 [5] that $r_\sigma(\mathcal{L}) = r_\sigma(\mathcal{T})$ (in fact $\mathcal{L} = \mathcal{T}^*$ according to (9)). We also define the operator $\mathcal{G} \in \mathbb{B}(\mathbb{H}^n)$ as

$$\mathcal{G}_i(V) := \sum_{j=1}^{N} p_{ij}\Gamma_j^* V_j \Gamma_j \tag{12}$$

where again $V = (V_1,...,V_N) \in \mathbb{H}^n$ and $\mathcal{G}(V) = (\mathcal{G}_1(V),\ldots,\mathcal{G}_N(V))$. It has been shown in [5] 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 [5], that model (1) with $u(k) = F_{\theta(k)}x(k)$, and $V_i(k) = E\left(x(k)x(k)^*1_{\{\theta(k)=i\}}\right)$, $V(k) = (V_1(k),\ldots,V_N(k)) \in \mathbb{H}^{n+}$ leads to

$$V(k+1) = \mathcal{T}(V(k)), \qquad k = 0,1,\ldots$$

where $\Gamma_i = A_i + B_iF_i$ in (11), 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 [3]), let the operator $\mathcal{H} \in \mathbb{B}\left(\mathbb{H}^{n^2}\right)$ 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

$$\mathcal{H}_j(S) := \sum_{i=1}^{N} p_{ij}(\overline{\Gamma}_i \otimes \Gamma_i)S_i(\Gamma_i' \otimes \Gamma_i^*). \tag{13}$$

Let $S_i(k) = E\left(vec\{x(k)x(k)^*\}vec\{x(k)x(k)^*\}^*1_{\{\theta(k)=i\}}\right)$, $S(k) = (S_1(k),\ldots,S_N(k)) \in \mathbb{H}^{n^2+}$. We have the following result.

Lemma 3 *$S(k+1) = \mathcal{H}(S(k))$ and $E\left(\|x(k)\|^4\right) = \sum_{i=1}^N tr\{S_i(k)\}$.*

Proof. 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 $tr\{z_i(k)z_i(k)^*\} = \|x_i(k)\|^2 tr\{x_i(k)x_i(k)\} = \|x_i(k)\|^4$. ■

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 [5], 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 (11) (or (10), (12)) in closed loop is less than one.

Lemma 4 *$F = (F_1,\ldots,F_N) \in \mathbb{H}^{n,m}$ stabilizes (A,B) in the mean square sense if and only if $r_\sigma(\mathcal{T}) < 1$, where $\mathcal{T}$ is as in (11) with $\Gamma_i = A_i + B_iF_i$.*

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*

$$\begin{aligned}
\mathcal{F}_i(X) &:= -\left(B_i'\mathcal{E}_i(X)B_i + D_i'D_i\right)^{-1}B_i'\mathcal{E}_i(X)A_i\\
\mathcal{V}_i(F) &:= C_i'C_i + F_i'D_i'D_iF_i\\
\mathcal{R}_i(X) &:= A_i'\mathcal{E}_i(X)A_i + C_i'C_i \qquad (14)\\
&\quad - A_i'\mathcal{E}_i(X)B_i\left(B_i'\mathcal{E}_i(X)B_i + D_i'D_i\right)^{-1}B_i'\mathcal{E}_i(X)A_i
\end{aligned}$$

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

$$\begin{aligned}
&(A_i + B_iF_i)^*\mathcal{E}_i(S)(A_i + B_iF_i) + F_i^*D_i^*D_iF_i = \qquad (15)\\
&(A_i + B_i\mathcal{F}_i(S))^*\mathcal{E}_i(S)(A_i + B_i\mathcal{F}_i(S)) + \mathcal{F}_i(S)^*D_i^*D_i\mathcal{F}_i(S) +\\
&(F_i - \mathcal{F}_i(S))^*(B_i^*\mathcal{E}_i(S)B_i + D_i^*D_i)(F_i - \mathcal{F}_i(S)).
\end{aligned}$$

The next Lemma, proved in [7], 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 5 *Suppose that (A,B) is mean square stabilizable and considers $F^0 = \left(F_1^0,\ldots,F_N^0\right) \in \mathbb{H}^{n,m}$ such that stabilizes (A,B) in the mean square sense. Then for $l = 0,1,2,\ldots$, there exists $P^l = \left(P_1^l,\ldots,P_N^l\right)$ 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 $\mathcal{L}^l(\cdot) = \left(\mathcal{L}_1^l(\cdot),\ldots,\mathcal{L}_N^l(\cdot)\right)$ and for $i = 1,\ldots,N$,

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

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

Moreover there exists $P^+ = \left(P_1^+,\ldots,P_N^+\right) \in \mathbb{H}^{n+}$ such that $P^+ = \mathcal{R}\left(P^+\right)$, $P^+ \geq X$ 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 $\mathcal{L}^+(\cdot) = \left(\mathcal{L}_1^+(\cdot),\ldots,\mathcal{L}_N^+(\cdot)\right)$ is defined as $\mathcal{L}_i^+(\cdot) = A_i^{+*}\mathcal{E}_i(\cdot)A_i^+$, for $i = 1,\ldots,N$, and

$$\begin{aligned}
F_i^+ &= \mathcal{F}_i\left(P^+\right)\\
A_i^+ &= A_i + B_iF_i^+.
\end{aligned}$$

Remark 1 *Step c) in Lemma 5, which corresponds the calculation of the solution of the linear system*

$$X^l = \mathcal{L}^l\left(X^l\right) + \mathcal{V}\left(F^l\right),$$

can be seen as the policy evaluation step in the policy iteration technique for MDP, while from identity (15),

$$\begin{aligned}
&(A_i + B_iF_i)'\mathcal{E}_i(X^l)(A_i + B_iF_i) + C_i'C_i + F_i'D_i'D_iF_i =\\
&X_i^l - (F_i^l - \mathcal{F}_i(X^l))'(B_i'\mathcal{E}_i(X^l)B_i + D_i'D_i)(F_i^l - \mathcal{F}_i(X^l))\\
&+ (F_i - \mathcal{F}_i(X^l))'(B_i'\mathcal{E}_i(X^l)B_i + D_i'D_i)(F_i - \mathcal{F}_i(X^l))
\end{aligned}$$

and the right hand side is minimized in F_i by choosing $F_i = F_i^{l+1} = \mathcal{F}_i\left(P^l\right)$, which can be seen as the policy improvement step.

We close this section with the following result that 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

$$\sum_{t=0}^\infty \gamma(t) = \infty \qquad (16)$$

$$\sum_{t=0}^\infty \gamma(t)^2 < \infty. \qquad (17)$$

Assume also that the noise matrix terms satisfy

$$\mathbb{E}(W(t)\big|\Sigma_t) = 0 \tag{18}$$

$$\mathbb{E}(\|W(t)\|^2\,\big|\Sigma_t) \leq A(t) \tag{19}$$

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, given by the sequence

$$R(t+1) = (1 - \gamma(t))R(t) + \gamma(t)W(t) \tag{20}$$

Lemma 6 *Suppose that* (16)-(19) *are satisfied and the sequence of $n \times n$ matrices $\{R(t); t = 0, 1, \ldots\}$ are given by* (20). *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 [1]. ■

3 $TD(\lambda)$ Algorithm

If the transition probability matrix $\mathcal{P}$ were known in advance we could use Lemma 5 to obtain an iterative algorithm for the maximal solution P of the CARE (3),(4) and then from (6) 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 5. But this would not be enough to obtain $F^{l+1} = \mathcal{F}\left(P^l\right)$ since, as can be seen from (14), 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

$$S_i^l = \sum_{j=1}^N p_{ij}\left(A_j + B_j F_j^l\right)^* S_j^l \left(A_j + B_j F_j^l\right) + \sum_{j=1}^N p_{ij}\mathcal{V}_j\left(F^l\right), \qquad i = 1, \ldots, N$$

or

$$S^l = \mathcal{G}^l(S^l) + \mathcal{E}(\mathcal{V}(F^l)).$$

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 Lemma 2 can be applied to say that the equation in Y

$$Y = \mathcal{G}^l(Y) + Q^l \tag{21}$$

has a unique solution S^l given by

$$S^l = \sum_{k=0}^\infty \left(\mathcal{G}^l\right)^k (Q^l). \tag{22}$$

Once S^l is calculated, F^{l+1} can be obtained from (14) as

$$F_i^{l+1} = -(B_i^* S_i^l B_i + D_i^* D_i)^{-1} B_i^* S_i^l A_i, \qquad i = 1, \ldots, N. \tag{23}$$

The remaining of this section is now devoted to calculating S^l through Monte Carlo simulations, tracing a parallel with $TD(\lambda)$ methods (see [1]). 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:

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

$$Z := \sum_{k=0}^\infty (\lambda\mathcal{G})^k(Q).$$

We have the following result.

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

Proof. See [4]. ■

By iterating (21) with $Y = S$ we have

$$S = \mathcal{G}^{k+1}(S) + \sum_{t=0}^k \mathcal{G}^{k-t}(Q), \qquad k = 0, 1, \ldots. \tag{24}$$

We have from (24) that

$$\begin{aligned} S &= \sum_{k=0}^\infty (1-\lambda)\lambda^k S \\ &= (1-\lambda)\left[\sum_{k=0}^\infty \lambda^k \left(\mathcal{G}^{k+1}(S) + \sum_{t=0}^k \mathcal{G}^{k-t}(Q)\right)\right] \qquad (25) \\ &= \sum_{t=0}^\infty (\lambda\mathcal{G})^t(Q) + (1-\lambda)\sum_{k=0}^\infty \lambda^k \mathcal{G}^{k+1}(S) = Z + \mathcal{J}(S) \\ &= \sum_{k=0}^\infty \lambda^k \left[\mathcal{G}^k(Q) + \mathcal{G}^{k+1}(S) - \mathcal{G}^k(S)\right] + S. \end{aligned}$$

Equation (25) 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 (16), (17) 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:

$$Y_i(t+1) = Y_i(t) + \gamma(t)\sum_{k=0}^\infty \lambda^k \mathcal{D}_i(t,k) \qquad i = 1, \ldots, N,$$

or in a more compact form,

$$Y(t+1) = Y(t) + \gamma(t)\sum_{k=0}^\infty \lambda^k \mathcal{D}(t,k,Y(t)), \tag{26}$$

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

$$\Upsilon_i(t,k) := \left(A_{\theta_i(t,k)} + B_{\theta_i(t,k)}F_{\theta_i(t,k)}\right)\Upsilon_i(t,k-1)$$

$$\Upsilon_i(t,0) := I$$

$$\mathcal{D}_i(t,k,V) := \mathcal{B}_i(t,k) + \mathcal{C}_i(t,k,V)$$

$$\mathcal{B}_i(t,k) := \Upsilon_i(t,k)^*\Big[C^*_{\theta_i(t,k+1)}C_{\theta_i(t,k+1)} + F^*_{\theta_i(t,k+1)}D^*_{\theta_i(t,k+1)}D_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)}\Big]\Upsilon_i(t,k)$$

$$\mathcal{C}_i(t,k,V) := \Upsilon_i(t,k)^*\Big[\left(A_{\theta_i(t,k+1)} + B_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)}\right)^* V_{\theta_i(t,k+1)}\left(A_{\theta_i(t,k+1)} + B_{\theta_i(t,k+1)}F_{\theta_i(t,k+1)}\right) - V_{\theta_i(t,k)}\Big]\Upsilon_i(t,k).$$

Notice that

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

and we define

$$W_i(t):=\sum_{k=0}^{\infty}\lambda^k\Big(\mathcal{D}_i(t,k,Y(t))-\Big(\mathcal{G}_i^k(Q)+\mathcal{G}_i^{k+1}(Y(t))-\mathcal{G}_i^k(Y(t))\Big)\Big),\quad i=1,\ldots,N.$$

Notice that from (25) we can rewrite (26) as follows:

$$Y(t+1)=(1-\gamma(t))Y(t)+\gamma(t)(Z+\mathcal{J}(Y(t))+W(t)). \tag{27}$$

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

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

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

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

In the next Proposition consider $\mathcal{H}$ as in (13) with $\Gamma_i=A_i+B_iF_i$.

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

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

Proof. See [4]. ∎

Thus (18) and (19) are satisfied for each $W_i(t)$, $i=1,\ldots,N$. The next result shows that through a linear transformation in the algorithm (27) we can assume that $\|\mathcal{J}\|<1$.

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

Proof. See [4]. ∎

Thus from now on we shall assume without loss of generality that $\|\mathcal{J}\|<1$ in (27). The next results follows the arguments presented in [1], 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. See [4]. ∎

We have now the main result of this section, proving the convergence of $Y(t)$ to S.

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.*

Proof. This proof follows the same steps as the proof of Proposition 4.5 in [1], pages 166-167. First notice that from (25),

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

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 $i=1,\ldots,N$, $t\ge t_0\ge 1$, $R_i(t_0,t_0)=0$ and

$$R_i(t+1,t_0)=(1-\gamma(t))R_i(t,t_0)+\gamma(t)W_i(t)$$

we have from Lemma 6, 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 $i=1,\ldots,N$, $t_0\ge 1$. 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 $t\ge 1$

$$\|X(t)(\omega)\|\le\|Y(t)(\omega)\|+\|S\|\le d(\omega).$$

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 $s\ge u(\omega)$) and prove that there is always $t_{k+1}(\omega)\ge t_k(\omega)$ such that

$$\|X(t)(\omega)\|\le d_k(\omega),\qquad t\ge t_k(\omega). \tag{28}$$

For simplicity we shall suppress ω from now on. For $t=0$ the result is immediate. Suppose (28) holds for k. Consider the sequence

$$y(t+1)=(1-\gamma(t))y(t)+\gamma(t)\,\|\mathcal{J}\|\,d_k, t\ge t_k,\ y(t_k)=d_k. \tag{29}$$

Let us show by induction that

$$\|X(t)-R(t,t_k)\|\le y(t),\qquad t\ge t_k. \tag{30}$$

For $t=t_k$ we have $y(t_k)=d_k$, $R(t_k,t_k)=0$ and the result follows from (28). Suppose (30) holds for t. Then

$$X(t+1)-R(t+1,t_k)=(1-\gamma(t))(X(t)-R(t,t_k))+\gamma(t)\mathcal{J}(X(t))$$

and since $\|\mathcal{J}(X(t))\|\le\|\mathcal{J}\|\|X(t)\|$, we have from (28), (29) and (30) that

$$\begin{aligned}\|X(t+1)-R(t+1,t_k)\|&\le(1-\gamma(t))y(t)+\gamma(t)\|\mathcal{J}\|d_k\\&=y(t+1)\end{aligned}$$

showing (30) for $t+1$. Thus (30) holds for all $t\ge t_k$. 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 $t_{k+1}\ge t_k$ such that $y(t)\le(\|\mathcal{J}\|+\frac{\nu}{2})d_k$ and $\|R(t,t_k)\|\le\frac{\nu}{2}d_k$ for all $t\ge t_{k+1}$. Thus from (30) we have for all $t\ge t_{k+1}$

$$\begin{aligned}\|X(t)\|&\le\|X(t)-R(t,t_k)\|+\|R(t,t_k)\|\\&\le y(t)+\frac{\nu}{2}d_k\le(\|\mathcal{J}\|+\nu)d_k=d_{k+1}\end{aligned}$$

proving (28) for $k+1$. Since d_k goes to zero as k goes to infinity, the result follows. ∎

Remark 2 *In practice one should try λ such that the convergence of $Y(t)$ to S takes place, since $\mathcal{P}$ is not known, and thus it is not possible to check if $\lambda^2 r_\sigma(\mathcal{H})<1$*

4 Numerical Example

To illustrate the use of the results developed in the previous sections, we have chosen a simple economic system based on Samuelson's multiplier-accelerator model [2] which appears in state equation form:

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

We considered the following three-operating mode discrete time jump linear: The discrete time state transition probability matrix is assumed to be as

Table 1: Datas for the multiplier-accelerator model

i	1	2	3
$A_i =$	$\begin{bmatrix} 0 & 1 \\ -2.5 & 3.2 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ -4.3 & 4.5 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 5.3 & -5.2 \end{bmatrix}$
$B_i =$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
$C_i'C_i =$	$\begin{bmatrix} 3.6 & -3.8 \\ -3.8 & 4.87 \end{bmatrix}$	$\begin{bmatrix} 10 & -3 \\ -3 & 8 \end{bmatrix}$	$\begin{bmatrix} 5 & -4.5 \\ -4.5 & 4.5 \end{bmatrix}$
$D_i'D_i =$	2.6	1.165	1.111

$$P = \begin{bmatrix} 0.67 & 0.17 & 0.16 \\ 0.30 & 0.47 & 0.23 \\ 0.26 & 0.10 & 0.64 \end{bmatrix}.$$

We define the next algorithm for $\ell = 0, 1, \ldots$, in the following way:

i) F_i^ℓ is calculated as (23) (except for $\ell = 0$).

ii) S^ℓ is the stationary value of equation (26).

We have performed this algorithm with different values of λ (see table 2). Final convergence to the optimal gain controller took place after $\ell = 20$ iterations. On the 3rd column of the table it is shown the error calculated in accordance with the following equation:

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

where F_i^{real} is the optimal feedback gain controller associated to mode i.

Table 2: $TD(\lambda)$-simulation method applied to the multiplier-accelerator model

λ	ℓ	Δ_i		
0.1	20	$\Delta_1 =$	0.0409	0.1465
		$\Delta_2 =$	0.0134	0.3296
		$\Delta_3 =$	0.1029	0.2365
0.5	20	$\Delta_1 =$	0.3123	0.2707
		$\Delta_2 =$	0.0377	0.5293
		$\Delta_3 =$	0.0070	0.5351
0.9	20	$\Delta_1 =$	0.6091	0.1545
		$\Delta_2 =$	0.1658	0.0704
		$\Delta_3 =$	0.1747	0.1477

5 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 it is possible to simulate trajectories for the Markov chain $\theta(k)$.

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. Some numerical examples to illustrate the results are presented.

References

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

[2] 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.

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

[4] O.L.V. Costa and J.C.C Aya. Monte carlo TD(λ)-methods for the optimal control of discrete-time markovian jump linear systems. to appear, 2000.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

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

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

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