ON MODEL-FREE AND MODEL-BASED TECHNIQUES FOR THE QUADRATIC CONTROL OF MARKOV JUMP LINEAR SYSTEMS WITH UNKNOWN TRANSITION PROBABILITIES

RAFAEL L. BEIRIGO*, MARCOS G. TODOROV*, ANDRÉ M. S. BARRETO*

*National Laboratory for Scientific Computing, Rio de Janeiro, Brazil

Emails: rafaelb@lncc.br, todorov@lncc.br, amsb@lncc.br

Abstract— Markov jump linear systems encompass a theoretically sound and solid framework for pursuing the optimal control of systems with switching dynamics. Despite its broad applicability, demand for *a priori* perfect knowledge of the transition model may render the application of the solution techniques impractical. To circumvent this limitation, two techniques were recently proposed that prescind from the perfect prior knowledge of the transition model. In the face of one being *model-free*, and the other *model-based*, the promising results that were presented by each technique separately may admit a comparative analysis. Here, we provide an experimental evaluation of both techniques, applying them to the control in a simulator of a robotic arm whose joints are subject to failure. Additionally, we test two variations of the policy update strategy for the model-free technique. The experimental results suggest a comparable performance for both techniques.

Keywords— Markov jump linear systems, reinforcement learning, adaptive control, robotics.

Resumo— Sistemas lineares com saltos Markovianos compreendem um arcabouço de sólida fundamentação teórica para a busca do controle ótimo em sistemas cuja dinâmica está sujeita a chaveamento. Apesar de possuir larga aplicabilidade, a demanda por conhecimento perfeito do modelo de transição *a priori* pode tornar a sua aplicação impraticável. Para contornar essa limitação, recentemente foram propostas duas técnicas que prescindem do conhecimento prévio do modelo de transição. Considerando que uma das técnicas é *livre de modelo*, sendo a segunda *baseada em modelo*, os resultados promissores apresentados por ambas permitem uma análise comparativa. Neste trabalho apresentamos uma avaliação experimental das referidas técnicas, realizada através de aplicação ao controle em um simulador de braço robótico cujas articulações estão sujeitas a falha. Adicionalmente, apresentamos testes envolvendo duas variações na estratégia de atualização de política para a técnica livre de modelo. Os resultados da avaliação experimental sugerem desempenhos comparáveis para ambas as técnicas avaliadas.

Palavras-chave— Sistemas com saltos Markovianos, aprendizado por reforço, controle adaptativo, robótica.

1 Introduction

Considering the ubiquitous possibility of system malfunction that accompanies each control design process, taking such phenomenon into account may present considerable and valuable advantages. Systems in general may undergo *abrupt changes*, whose examples include not only likely catastrophic situations, such as parts malfunction, infrastructure breakdown, economic collapse, environmental disasters, but also dynamical changes in general, like the weather sensibility to the seasons, service demand variation in time, or even robotic dynamical changes due to warming of its corresponding parts, for instance. These matters are investigated in a vast amount of recent literature (see [1, 2, 3, 4], for a small sample), but open problems still abound.

In this work, we apply the formalization framework provided by the Markov Jump Linear Systems (MJLS) paradigm, where, essentially, the switching behavior in the system's dynamics is modeled by means of a Markov chain, with each state corresponding to one operational mode of the system. This class of systems exhibits broad applicability, encompassing flight systems, networked control, robotics, and economics, for instance. A more comprehensive discussion on the subject may be found in the books [5, 6, 7, 8, 9, 10], presenting MJLS' solid theoretical foundations, which provide fertile grounds for the investigation of this class of systems.

Notwithstanding the enormous flexibility given by the MJLS model, assuming perfect knowledge of the system's transition probabilities may present a barrier, that, if not unsurpassable, could be rather restrictive to its practical applicability. Indeed, substantial research endeavor was undertaken in several scenarios, with the investigation of methods that were able to *estimate* or *obtain bounds* for the transition probabilities. These include, for instance, *polytopic uncertainty*, in which a polytope with known vertices may contain the unknown parameters [11, 12, 13]; the *multi-simplex* setup [14]; the *partially known* case [15]; the *norm-bounded* setup [16]; a randomized Gaussian modeling [17]; *maximum-likelihood estimation* [18], possibly with *transfer* [19]; *temporal differences* [20, 21]. A more recent discussion on the subject may be found in the book [22], which presents also several other setups.

In this paper, we investigate the optimal quadratic control of MJLS when there is no prior knowledge of the Markov chain transition probabilities. We analyse two recently proposed techniques: *maximum likelihood estimation*, referred to as MLE$_{alg}$ [18], and a method that applies *online temporal differences with eligibility traces*, dubbed Online TD(λ) [21]. MLE$_{alg}$ is *model-based*, meaning it circumvents the problem

The authors are with the National Laboratory for Scientific Computing - LNCC/MCTIC, Av. Getúlio Vargas 333, Petrópolis, Rio de Janeiro, CEP 25651-070, Brazil. The third author is currently with Google DeepMind, London, UK. This work was partially supported by the Brazilian National Research Council – CNPq, Grants 421486/2016-3 and 461739/2014-3, and Coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, Grant 1528969/2015-5.

of not having the transition model by using transition samples to *approximate* one. On the other hand, Online TD(λ) is able to completely prescind from the transition model, thus being qualified as a *model-free* technique. Model-based techniques usually possess two main characteristics: (i) the model approximation process demands *memory* to keep it, and (ii) sampled data may not be used immediately after being obtained. From these characteristics, (i) may imply an increase in the *memory complexity* of the corresponding algorithm, while (ii) has the potential of delaying the technique's convergence process, by preventing it from immediately applying the sampled data.

We experimentally analyse both techniques in the optimal control problem of a robotic arm whose actuators are subject to failure. Results of applying the algorithms as originally proposed are presented, along with two variations of Online TD(λ), concerning the time step interval related to the policy update. The experimental results show that the analysed techniques, along with the proposed variations, are able to adequately solve the problem, with qualitatively comparable performance.

This paper is organized as follows. Section 2 presents the notation, along with the definitions and relevant results from the literature. The MLE$_{alg}$ and Online TD(λ) algorithms are presented in Section 3. Experimental results are put together in Section 4. We present some concluding remarks together with directions for future research in Section 5.

2 Preliminaries

Consider a homogeneous Markov chain $\theta = \{\theta^k;\ k = 0,1,2,\ldots\}$ in a complete stochastic basis $(\Omega,\mathfrak{F},\{\mathfrak{F}_k\},\mathbb{P})$, along with the state space $\mathcal{S} \triangleq \{1,\ldots,N\}$, with $\theta^k \in \mathcal{S}, \forall k$, where the transition matrix $P = [p_{ij}]$, i.e., $P \in \Pi \triangleq \{Q = [q_{ij}] \in \mathbb{R}^{N\times N};\ q_{ij} \geq 0, \sum_{j=1}^{N} q_{ij} = 1\ \forall i,j \in \mathcal{S}\}$ is an N-by-N row-stochastic matrix. The initial distribution of θ will be denoted $\nu_i = \mathbb{P}(\theta^0 = i)$ for all such i. The mathematical expectation with respect to $\mathbb{P}$ will be represented as $\mathbb{E}$, and 1_A will stand for the indicator of the event A (a random variable that can be equal to 1 or 0, depending on whether the event occurs or not). For later use, consider also the following operation:

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

for any N-sequence of matrices of the form $X = (X_1,\ldots,X_N)$, with $\mathcal{E}(X) \triangleq (\mathcal{E}_1(X),\ldots,\mathcal{E}_N(X))$, with the slight variation $\mathcal{E}_i^t(X) \triangleq \sum_{j=1}^{N} p_{ij}^t X_j$, $i \in \mathcal{S}$. The set of all N-sequences of positive semidefinite n-by-n matrices, i.e., objects of the form $X = (X_1,\ldots,X_N)$ such that $0 \leq X_i \in \mathbb{R}^{n\times n}$ for all $i \in \mathcal{S}$, will be denoted $\mathbb{H}_N^{n+}$. Throughout the paper, $\|\cdot\|$ will stand for the euclidean norm of vectors.

The following discrete-time control system comprises the subject of our study:

$$\begin{cases} x^{k+1} = A_{\theta^k} x^k + B_{\theta^k} u^k \\ z^k = C_{\theta^k} x^k + D_{\theta^k} u^k \\ x^0 = \tilde{x},\ \theta^0 = \tilde{\theta}, \end{cases} \tag{2}$$

with $x = \{x^k \in \mathbb{R}^{n_x};\ k = 0,1,2,\ldots\}$ corresponding to the state, $u = \{u^k \in \mathbb{R}^{n_u};\ k = 0,1,2,\ldots\}$ to the control, and $z = \{z^k \in \mathbb{R}^{n_z};\ k = 0,1,2,\ldots\}$ representing the controlled output. The shorthand notation $(\cdot)_{\theta^k} \equiv (\cdot)_i$ shall be used when $\theta^k = i$ (e.g., if $\theta^k = i$, then $A_{\theta^k} \equiv A_i$).

2.1 Jump linear quadratic optimal control

We define

$$\begin{aligned} \mathbf{A}_i &\triangleq A_i + B_i F_i, \\ \mathbf{C}_i &\triangleq C_i + D_i F_i, \end{aligned} \quad i \in \mathcal{S}, \tag{3}$$

assuming[1]

$$\begin{aligned} C_i' D_i &\equiv 0, \\ D_i' D_i &> 0, \end{aligned} \quad i \in \mathcal{S}, \tag{4}$$

and focus our attention on the so-called *state-feedback* controllers, which have the form

$$u = \{u^k = F_{\theta^k} x^k;\ k = 0,1,2,\ldots\}, \tag{5}$$

thus arriving at the *closed-loop* variant of system (2), which is given by

$$\begin{cases} x^{k+1} = \mathbf{A}_{\theta^k} x^k \\ z^k = \mathbf{C}_{\theta^k} x^k \\ x^0 = \tilde{x},\ \theta^0 = \tilde{\theta}. \end{cases} \tag{6}$$

Our interest resides in optimizing the controlled output z^k of (6), when the control performance is measured by the *infinite horizon quadratic cost*, defined by

$$\mathfrak{J}(\tilde{\theta},\tilde{x},u) \triangleq \sum_{k=0}^{\infty} \mathbb{E}\big(\|z^k\|^2\big). \tag{7}$$

The notion of *stability* for the system (6) used in this work is formalized in Definition 1.

Definition 1 (mean square stability) *A control $u = \{u^k;\ k = 0,1,2,\ldots\}$ is said to* stabilize system (2) in the mean square sense *if, regardless of $x^0 \in \mathbb{R}^n$ and $\theta^0 \in \mathcal{S}$, the application of u in* (6)*, yields*

$$\lim_{k\to\infty} \mathbb{E}\big(\|x^k\|^2\big) = 0. \tag{8}$$

In such case, we say that (6) *is* mean square stable *(MSS).* ▽

Here, we are interested in the solution of the following problem.

[1] It is shown in [5, Chapter 4], that the orthogonality between C and D is without loss of generality, besides presenting the advantage of simplifying the derivations made subsequently. Singular controls are ruled out by the other assumption, which guarantees non trivial penalization to every control action.

Problem 1 (Jump linear quadratic (JLQ)) *Find a control $\bar{u}$ that satisfies*

$$\mathfrak{J}(\theta^0, x^0, \hat{u}) \leq \mathfrak{J}(\theta^0, x^0, u), \forall u. \tag{9}$$

for system (6)*, and has the form* (5)*.* ▽

For mean square stabilizing controls of the form (5), the next lemma shows that an algebraic characterization of the closed-loop cost is possible.

Lemma 1 *If system* (6) *is stabilized in the mean square sense by a given controller u of the form* (5)*, then the corresponding cost in* (7) *is obtained by*

$$\mathfrak{J}(\theta^0, x^0, u) = x^{0\prime} X_{\theta^0} x^0, \tag{10}$$

with $\mathbb{H}_N^{n+} \ni X = (X_1, \ldots, X_N)$ being the unique solution of the following Lyapunov-like equation

$$X_i = \mathbf{A}_i' \mathcal{E}_i(X) \mathbf{A}_i + \mathbf{G}_i, \quad i \in \mathcal{S}, \tag{11}$$

where

$$\mathbf{G}_i \triangleq \boldsymbol{C}_i' \boldsymbol{C}_i, \quad i \in \mathcal{S}. \tag{12}$$

Proof: See [5, Chapter 4]. □

By Lemma 1, it is clear that, in order to solve Problem 1, we must be able to find *control gains*

$$F \triangleq (F_1, \ldots, F_N), \tag{13}$$

which give us a control with the form (5), that minimize (10), under the constraint (11), for each $i \in \mathcal{S}$. Nonetheless, two severe complications clearly arise by a quick examination of this setup, and will be addressed in the sequel:

(i) The minimization of (10) subject to (11), that underlies the optimization problem carries a great deal of implicit dependence upon the controller gains;

(ii) The direct solution of (11) demands the knowledge of the transition probabilities in (1).

2.2 Maximum-likelihood estimation of transition parameters

In this section we present the key concepts applied by the MLE$_{alg}$ algorithm [18] to approximate the transition model by means of maximum-likelihood estimation using the sampled transition data. First notice that the probability of a Markov chain visiting a given sequence of Markov states $\{\theta^0 = i_0, \theta^1 = i_1, \ldots, \theta^k = i_k\}$, corresponds to

$$\mathbb{P}\big(\theta^0 = i_0, \ldots, \theta^k = i_k\big) = \nu_{i_0} \prod_{t=0}^{k-1} p_{i_t i_{t+1}}. \tag{14}$$

For the scenario where the transition parameters are unknown, a sensible approximation scheme would involve applying the observed outcome to find the transition parameters that maximize a likelihood function of the form $Q \mapsto f(Q; i_0, \ldots, i_k) \triangleq \nu_{i_0} \prod_{t=0}^{k-1} q_{i_t i_{t+1}}$, which is subject to the constraints $q_{ij} \geq 0$, $\sum_{j=1}^{N} q_{ij} = 1$. Bearing in mind that, by taking the logarithm of the above expression, the essence of the underlying optimization problem suffers no change, besides converting the product into a sum, and by also discarding a constant term that is not related to the transition parameters, we end up with the log-likelihood objective function:

$$L(Q; i_1, \ldots, i_k) \triangleq \sum_{t=0}^{k-1} \log\big(q_{i_t i_{t+1}}\big). \tag{15}$$

Hence, it is not difficult to see that, if the process just described should not sample transitions from some state $\hat{\imath}$, (i.e., $\hat{\imath} \neq i_t$ for all $t = 0, \ldots, k-1$), then the resulting log-likelihood function is independent of the transition parameters of the $\hat{\imath}$th row of the transition model (i.e., (15) does not depend upon $q_{\hat{\imath} j}$ for either $j \in \mathcal{S}$). For this specific case, the arbitrary "flat prior" choice $q_{\hat{\imath} j}^{(k)} \equiv 1/N$ shall be made for the corresponding lines of the maximum-likelihood estimate $Q^{(k)} \in \arg\max\{L(Q; i_1, \ldots, i_k);\, Q \in \Pi\}$.

Focusing now on the states $i \in \mathcal{S}$ that were visited during the sample trajectory $\{\theta^0 = i_0, \theta^1 = i_1, \ldots, \theta^k = i_k\}$, i.e., those which

$$\nu_i^{(k)} \triangleq \sum_{t=0}^{k-1} 1_{\{\theta(t)=i\}} > 0, \tag{16a}$$

we have the well-known result (see e.g. [23, Section 1.10]) that establishes the maximum-likelihood estimate given by an empirical mean with the form

$$q_{ij}^{(k)} = \frac{r_{ij}^{(k)}}{\nu_i^{(k)}}, \qquad r_{ij}^{(k)} \triangleq \sum_{t=0}^{k-1} 1_{\{\theta(t)=i, \theta(t+1)=j\}}, \tag{16b}$$

corresponding to the ratio between the number of transitions $i \to j$ and the total count of visits to i. Therefore, in face of the sampled data $\{\theta^0 = i_0, \theta^1 = i_1, \ldots, \theta^k = i_k\}$ the most likely transition model, according to the log-likelihood criterion (15), is given by the N-by-N row-stochastic matrix $Q^{(k)} = [q_{ij}^{(k)}]$:

$$q_{ij}^{(k)} = \begin{cases} N^{-1}, & \text{if } i \notin \{i_1, \ldots, i_k\}, \\ r_{ij}^{(k)} / \nu_i^{(k)} \text{ as in (16)}, & \text{otherwise}. \end{cases} \tag{17}$$

Concerning the consistency of this maximum-likelihood estimation setup, we have the following well-known result (proved in [23, Section 1.10]): if the Markov state $i \in \mathcal{S}$ is visited infinitely often (i.e., $\nu_i^{(k)} \to \infty$), then the empirical means in (16b) will almost surely converge to the (true) probability distribution $(p_{i1}, \ldots, p_{iN})$.

Lemma 2 *The random variables in* (16) *satisfy, with probability one,*

$$\nu_i^{(k)} \to \infty \quad \Rightarrow \quad q_{ij}^{(k)} \to p_{ij} \quad \forall\, j \in \mathcal{S}. \tag{18}$$

3 Algorithms

In this section we describe the algorithms that are analysed in this work. We start by presenting MLE$_{alg}$ [18], in Section 3.1, a *model-based* technique that applies maximum likelihood estimation to approximate the transition model using sampled transition data. In Section 3.2, we present Online TD(λ), a *model-free* technique that, besides prescinding from the transition model altogether, may immediately apply the sampled transition data to improve the policy.

Before introducing the solution techniques, we first formalize the problem being solved. We consider a variant of the JLQ control problem, defined in Section 2, where the transition probabilities $P = [p_{ij}] \in \Pi$ are unknown, hence, the optimal control $\bar{u}$ cannot be computed, in principle. We apply MLE$_{alg}$ and Online TD(λ) to calculate the optimal control $\bar{u}$, under the following assumption.

Assumption 1 *Regarding system* (6)*, we assume that:*

(A1) The transition probability matrix P is unknown, but all the other system parameters are perfectly known.

(A2) The conditions of Lemma 1 are satisfied, thus there exists a stabilizing solution to (11)*.*

(A3) For each $k = 0, 1, \ldots$, a sample path obeying the transition matrix $P = [p_{ij}]$ of Markov states $\{\theta^0 = i_0, \theta^1 = i_1, \ldots, \theta^k = i_k\}$ is available.

(A4) P is irreducible. ▽

Remark 1 *Condition (A1) guarantees that the control problem is, except from the fact that P cannot be directly used for design purposes, exactly the one treated in Lemma 1. The existence of a solution to the control problem we are interested in, on the other hand, is ensured by condition (A2), so it makes sense to seek the corresponding approximation. Availability, at any time instant, of all the past trajectory of the Markov chain, is guaranteed by (A3) (even if represented by access to an accurate "simulator" of it, e.g., in an experimental setup). Finally, (A4) ensures with probability one that, regardless of the initial distribution ν of the Markov chain, for all $i \in \mathbb{S}$, the number of visits to i tends to infinity as $k \to \infty$.* ▽

3.1 MLE$_{alg}$

We first introduce the algorithm MLE$_{alg}$ [18], a model-based algorithm that applies sampled transition data to build the model, which is then used to calculate improving versions of the control policy. One of the features presented by MLE$_{alg}$ is being able to deal with two distinct scenarios involving the prior knowledge of P: it might be entirely unknown, or there may be available some prior estimate $\bar{P}^0$ of P, in which case it may be included in the computation, as formalized in (19).

$$\bar{P}^t = (1-\alpha_t)\bar{P}^{t-1} + \alpha_t Q^t. \tag{19}$$

Algorithm 1: MLE$_{alg}$ (adapted from [18])

Require: F^t is stabilizing, $t = 0, \ldots, \tau$

$k \leftarrow 0$
for $t = 0, 1, 2, \ldots, \tau$ **do**
  if $k \in \Xi$ **then**
    $\theta^k \leftarrow \theta^0$, $x^k \leftarrow x^0$
  end if
  repeat
    Perform $u^k \equiv F^t_{\theta^k} x^k$; observe $\theta^k \to \theta^{k+1}$
    Compute Q^{k+1} as in (16)
    $k \leftarrow k+1$
  until $k = k_t$
  $\alpha_t = \min\left\{\frac{k}{k_s}, 1\right\}$
  $Q^t \leftarrow Q^{k_t}$
  Compute $\bar{P}^t$ as in (19)
  if (11) can be solved applying $\bar{P}^t$ **then**
    $F_i^{t+1} \leftarrow -\left(B_i' \mathcal{E}_i^t(X^t) B_i + D_i' D_i\right)^{-1} B_i' \mathcal{E}_i^t(X^t) A_i$
  else
    $F_i^{t+1} \leftarrow F_i^t$
  end if
end for

Ensure: $F^\tau \approx F$

In this case, the current transition model $\bar{P}^t$ is a composition involving the maximum-likelihood estimate (16) and the preceding transition model. The non-negative parameter $\alpha_t = \min\{k/k_t, 1\}$ may be viewed as determining how much "credit" should be attributed to each of $\bar{P}$ and Q. For small values of t (corresponding to situations where yet few samples of θ are available), the parameter α_t diminishes the importance of the maximum likelihood estimation $\bar{P}$, favoring $\bar{P}^0$, which was provided as a reasonable estimate of P. This may make sense, as one ponders that, for yet a few samples (i.e., small values of t), $\bar{P}^t$ is likely not to be an accurate estimate of the true transition matrix P. With the increase in t, the model's belief will progressively nudge towards the maximum-likelihood estimates $\{Q^t;\ t = 1, 2, \ldots\}$. At the time instant $k_t = k_s$, it is expected that the MLE$_{alg}$ will represent a "fairly good" model by then, as α_t will be equal to one, ceasing this process from that moment on. In this work, k_s is referred to as a *bootstrapping horizon*, because it determines the time steps needed for the maximum likelihood estimation to be fully trustable (also considering the possibility of imprecision). Considering the case where P is entirely unknown and there is no prior, k_s is set to 1, thus the transition models coincides with the maximum likelihood estimation for all steps. As noted by the authors in [18], this structure may be adequate in situations where the transition parameters are accurate at the start of operation, becoming increasingly inaccurate as time progresses. MLE$_{alg}$ is presented in pseudocode in Algorithm 1.

Notice that the hypothesis of the Markov chain being irreducible ensures that every state is visited in-

finitely often, regardless of its initial distribution ν. For absorbing chains, however, the counting procedure (16) should yield, to all states, a finite number of visits, with the exception of those present in the recurrent class that eventually absorbs the process in that particular run. Therefore, the algorithm must be able to deal with the possibility of "getting stuck" in some absorbing class, thus ensuring that, in any particular run, all states were visited infinitely often. One way this may be done is by imposing, at prespecified time instants, *resets* to both the Markov and the continuous states. Algorithm 1 depicts this *episodic* version of MLE$_{alg}$, which, for a set of time instants $\Xi \subseteq \{k_0, k_1, \ldots, k_\tau\}$, for each $\xi \in \Xi$, the Markov and the continuous states may be reset (i.e., $x(\xi) = x^0$, $\theta(\xi) = \theta^0$), which is exemplified in the case where an experimental task is repeated multiple times.

3.2 Online TD(λ)

We now present Online TD(λ) [21], an algorithm that is able to solve the JLQ problem in a *model-free* basis, thus prescinding completely from the transition model P. Online TD(λ) applies *policy iteration* to refine an approximation of $\mathcal{E}(X)$, with $\mathbb{H}_N^{n+} \ni X = (X_1, \ldots, X_N)$ corresponding to the unique solution of (11), then using this approximation to calculate a better policy. The approximation sequence $\{\bar{Y}^t;\ t = 1, 2, \ldots\}$ is composed by elements $\bar{Y}^t$ calculated by

$$\bar{Y}_i^{t+1} = \bar{Y}_i^t + \gamma_t \sum_{k=0}^{\infty} e_i^{t,k} \mathcal{D}_i^{t,k}(\bar{Y}^{t,k}), \tag{20}$$

whose incremental form is given by

$$\bar{Y}_i^{t,k+1} = \bar{Y}_i^{t,k} + \gamma_t e_i^{t,k} \mathcal{D}_i^{t,k}(\bar{Y}^{t,k}), \tag{21a}$$

$$\text{with} \quad \bar{Y}_i^{t,0} = \bar{Y}_i^t, \quad \bar{Y}_i^{t+1} = \bar{Y}_i^{t,N_t}, \tag{21b}$$

where N_t is the episode length[2], the *stepsize* γ_t is assumed to satisfy the usual conditions

$$\sum_{t=0}^{\infty} \gamma_t = \infty \ \text{ and } \ \sum_{t=0}^{\infty} {\gamma_t}^2 < \infty, \tag{22}$$

whereas $\mathcal{D}_i^{t,k}(\cdot)$, the *temporal difference*, is given by

$$\mathcal{D}_i^{t,k}(\cdot) = \Upsilon_i'^{t,k}(\mathbf{G}_{\theta_t^{k+1}} + \mathbf{A}'_{\theta_t^{k+1}}(\cdot)_{\theta_t^{k+1}} \mathbf{A}_{\theta_t^{k+1}} - (\cdot)_{\theta_t^k}) \Upsilon_i^{t,k},$$

with

$$\Upsilon_i^{t,k} = \begin{cases} I, & \text{if } k = 0, \\ \mathbf{A}_{\theta_t^k} \Upsilon_i^{t,k-1}, & \text{if } k > 0, \end{cases} \tag{23}$$

the *cost* is given by

$$\mathbf{G}_i \triangleq \mathbf{C}_i' \mathbf{C}_i, \quad i \in \mathcal{S}, \tag{24}$$

and the *eligibility coefficients* by

$$e_i^{t,k} = \begin{cases} 0, & k < k_i^t, \\ \lambda^{k-k_i^t}, & k \geq k_i^t, \end{cases} \tag{25}$$

[2]For instance, N_t could correspond to the length of the trajectory θ^t, or a limit in the number of iterations for the algorithm.

Algorithm 2: Online TD(λ) (adapted from [21])

Require: F is stabilizing

1: **for** $\ell = 1, \ldots, L$ **do**
2: **for** $t = 1, \ldots, T$ **do**
3: Initialize θ^0
4: $e \leftarrow 0e$
5: **for** $k = 1, \ldots, K$ **do**
6: Perform $u^k \equiv F_{\theta^k} x^k$; observe $\theta^k \to \theta^{k+1}$
7: **if** $e_{\theta^k} = 0$ **then**
8: $e_{\theta^k} \leftarrow 1$
9: **end if**
10: $\bar{Y}_i \leftarrow \bar{Y}_i + \gamma_t e_i \mathcal{D}_i^{t,k}(\bar{Y}), \forall i \in \mathcal{S}$
11: $e \leftarrow \lambda e$
12: $(F \leftarrow \mathcal{F}(\bar{Y}))$ ▷ K-TD(λ) variation
13: **end for**
14: $(F \leftarrow \mathcal{F}(\bar{Y}))$ ▷ T-TD(λ) variation
15: **end for**
16: $(F \leftarrow \mathcal{F}(\bar{Y}))$ ▷ L-TD(λ) variation
17: **end for**

Ensure: $\bar{Y} \approx \mathcal{E}(X)$

where k_i^t is the first time that state i is visited in trajectory t, i.e.,

$$k_i^t = \inf_k \{\theta_t^k = i\}, \quad i \in \mathcal{S}. \tag{26}$$

A new policy F is calculated at the *policy improvement* step, when the current approximation $\bar{Y}$ of $\mathcal{E}(X)$ is applied in

$$F_i = \mathcal{F}(\bar{Y}) \triangleq -\left(B_i' \bar{Y}_i B_i + D_i' D_i\right)^{-1} B_i' \bar{Y}_i A_i, \quad i \in \mathcal{S}.$$

The original algorithm [21] calculates the new policy at the end of each "ℓ-cycle", represented by Line 16 in Algorithm 2, and is referred to by the "L-TD(λ)" variation. A natural question that could arise about which cycle the policy is updated at would be "what impact on the algorithm's performance would changing the cycle where the policy is updated have?" To answer this question, we analyse two variants of Online TD(λ), that update the policy either at the end of the "k-cycle" (Line 12) (denominated "K-TD(λ)"), or at the end of the "t-cycle" (Line 14) (denominated "T-TD(λ)"), instead of updating it at the end of the "ℓ-cycle" (Line 16).

One of the main distinctions between the algorithms presented in this section is that while MLE$_{alg}$, by being model-based, has to construct (and, thus, preserve in memory) an approximation of the model, Online TD(λ) does not approximate the model, thus being free from the corresponding memory complexity burden. This characteristic may present itself advantageous in scenarios where memory limitation is a concern. In the next section, we present experimental results from the application of both techniques in a simulator to control an underactuated robotic arm subject to actuator failure.

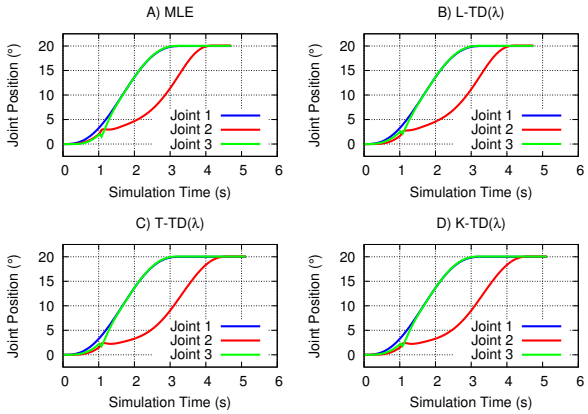

Figure 1: Evolution of the joint positions of the robotic manipulator's arm throughout time for the MLE$_{alg}$ algorithm (A), along with the three variants of TD(λ) (B, C, and D). The collective analysis of these results suggest a comparable performance presented by both the model-based (MLE$_{alg}$), and model-free (TD(λ) variants) techniques.

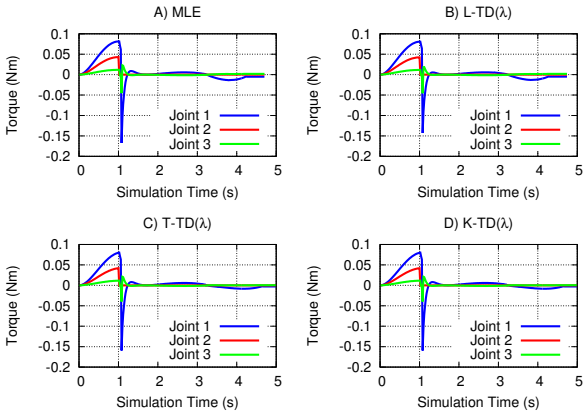

Figure 2: Applied torque *versus* time for the MLE$_{alg}$ algorithm (A), along with the three variants of TD(λ) (B, C, and D), corresponding to the angle variations presented in Figure 1. The collective analysis of these results corroborates the hypothesis of comparable performance presented by the model-based (MLE$_{alg}$), along with the model-free (TD(λ) variants) algorithms.

4 Experiments

We analysed MLE$_{alg}$ [18] and Online TD(λ) [21] in a robotic manipulator simulator of an arm with 3 joints, each containing an actuator and a brake, the Robust and Fault Tolerant Control Environment for Robots (CERob) [24]. In the simulator, the status of each actuator may be *active* or *passive*, the former corresponding to the situation where the motor responds infallibly to control. However, when the actuator is in the passive status, it does not respond to control, remaining on or off, with this situation being considered of *fault*. The arm's task consists of moving the arm from an initial set of joint angles to a final one, where this is accomplished by using the actuators to apply torque to the joints, considering the possibility of failure.

It was shown in [25] that computed torque plus $\mathcal{H}_\infty$ control is not sufficient to guarantee stability. In [25] this problem was modeled by a MJLS, with each state corresponding to the angle and velocity of each joint, together with the status of its actuator. The physical model, along with the linearization strategy, and a detailed description of their methodology can be found in [25, 26]. We applied for this example a setup for the simulation where only one of the three actuators may present a faulty behavior. Due to space restrictions, we refer the reader to [25], where a detailed description of the simulator-related parameters can be found.

To solve this problem, we applied Online TD(λ) with the following parameter values: $L = 200$, $T = 100$, $K = 10$, $\lambda = 0.1$, and $\gamma_k = 0.1/k$. We employed the following parameters for MLE$_{alg}$: $\tau = 200$, $k_0 = 0, k_1 = 1000, k_2 = 2000, \ldots, k_\tau = 200,000$, along with $k_s = 1$, corresponding to the more challenging scenario, where no prior model is provided, as mentioned in Section 3.1. For all variants of Online TD(λ), $\bar{Y}^{t=0}$ was initialized with zeros and for both MLE$_{alg}$ and Online TD(λ) variants, F^0 and F, respectively, were initialized with the optimal control gains for the problem.

Figure 1 shows the evolution of the angles for each joint corresponding to one task, comprising moving from the angles $[0°, 0°, 0°]$ to $[20°, 20°, 20°]$. The graphs A, B, C and D show the performance for the algorithms MLE$_{alg}$, along with the variants L-TD(λ), T-TD(λ) and K-TD(λ) of Online TD(λ). Analysing the graphs, the corresponding results suggest that MLE$_{alg}$, along with the TD(λ) variants were not only able to satisfactorily calculate adequate control gains, but they also presented a qualitatively comparable performance. While these results depict the information about the joint's trajectory throughout time, Figure 2 presents the evolution of the torque that was applied at each joint during the experiment. By analysing the corresponding graphs, one may see that the results appear to corroborate the hypothesis of performance comparability. Together, these results suggest that MLE$_{alg}$, along with the variants L-TD(λ), T-TD(λ), and K-TD(λ), were able to satisfactorily calculate the control policy for the robotic arm, considering the possibility of actuator failure, and without prior knowledge of the transition probabilities.

5 Concluding remarks

In this work, we presented a comparative analysis of a *model-based* and a *model-free* techniques to solve the *jump linear quadratic optimal control* problem (JLQ) for Markov jump linear systems (MJLS). The model-based technique was represented by MLE$_{alg}$ [18], a recently proposed algorithm that applies sampled transition data to approximate the

corresponding model, which is then used to obtain increasingly improved control policies. The model-free algorithm corresponded to the also recently proposed Online TD(λ) [21], which prescinds from the transition model altogether, thus being free from the associated memory complexity. Advantages presented by model-free techniques include not only situations where the transition model is difficult to establish in advance, but also those where the corresponding approximation cost is considerable, or may be subject to change throughout time (e.g., non-stationary domains). Both techniques were experimentally applied to the control of a robotic manipulator arm subject to actuator failure, and the results suggest comparable qualitative performance. Considering these results, along with the additional memory complexity needed by model-based techniques, one could tend to apply the model-free variant in situations where memory restriction is a concern.

References

[1] A. S. Morse, ed., *Control Using Logic Based Switching*. London: Springer-Verlag, 1997.

[2] C. G. Cassandras and J. Lygeros, *Stochastic Hybrid Systems*. Boca Raton, FL: Taylor & Francis, 2007.

[3] D. Liberzon, *Switching in Systems and Control*. Boston: Birkhäuser, 2003.

[4] R. Goebel, R. G. Sanfelice, and A. R. Teel, *Hybrid Dynamical Systems: Modeling, Stability, and Robustness*. Princeton, NJ: Princeton University Press, 2012.

[5] O. L. V. Costa, M. D. Fragoso, and R. P. Marques, *Discrete-Time Markov Jump Linear Systems*. Probability and Its Applications, New York: Springer-Verlag, 2005.

[6] O. L. V. Costa, M. D. Fragoso, and M. G. Todorov, *Continuous-Time Markov Jump Linear Systems*. Probability and Its Applications, Heidelberg: Springer-Verlag, 2013.

[7] V. Dragan, T. Morozan, and A. Stoica, *Mathematical Methods in Robust Control of Linear Stochastic Systems*, vol. 50 of *Mathematical concepts and methods in science and engineering*. New York: Springer, 2006.

[8] M. Mariton, *Jump Linear Systems In Automatic Control*. New York: Marcel Dekker, 1990.

[9] E.-K. Boukas, *Stochastic Switching Systems: Analysis and Design*. Boston: Birkhäuser, 2005.

[10] V. Dragan, T. Morozan, and A. Stoica, *Mathematical Methods in Robust Control of Discrete-Time Linear Stochastic Systems*. Springer, 2010.

[11] L. El Ghaoui and M. A. Rami, "Robust state-feedback stabilization of jump linear systems via LMIs," *Internat. J. Robust Nonlinear Control*, vol. 6, pp. 1015–1022, Nov. 1996.

[12] C. E. de Souza, "Robust stability and stabilization of uncertain discrete-time markovian jump linear systems," *IEEE Trans. Automat. Control*, vol. 51, no. 5, pp. 836–841, 2006.

[13] M. G. Todorov and M. D. Fragoso, "New methods for mode-independent robust control of Markov jump linear systems," *Systems Control Lett.*, vol. 90, pp. 38–44, 2016.

[14] C. F. Morais, M. F. Braga, R. C. L. F. Oliveira, and P. L. D. Peres, "$\mathcal{H}_2$ control of discrete-time Markov jump linear systems with uncertain transition probability matrix: improved linear matrix inequality relaxations and multi-simplex modeling," *IET Control Theory and Applications*, vol. 7, pp. 1665–1674, 2013.

[15] L. Zhang, E.-K. Boukas, and J. Lam, "Analysis and synthesis of markov jump linear systems with time-varying delays and partially known transition probabilities," *IEEE Trans. Automat. Control*, vol. 53, no. 10, pp. 2458–2464, 2008.

[16] M. Karan, P. Shi, and C. Y. Kaya, "Transition probability bounds for the stochastic stability robustness of continuous- and discrete-time Markovian jump linear systems," *Automatica*, vol. 42, no. 12, pp. 2159–2168, 2006.

[17] X. Luan, S. Zhao, and F. Liu, "$\mathcal{H}_\infty$ control for discrete-time Markov jump systems with uncertain transition probabilities," *IEEE Trans. Automat. Control*, vol. 58, no. 16, pp. 1566–1572, 2013.

[18] R. L. Beirigo, M. G. Todorov, and A. M. S. Barreto, "Count-based quadratic control of markov jump linear systems with unknown transition probabilities," in *Proc. of the 56th IEEE Conference on Decision & Control*, (Melbourne, Australia), 2017.

[19] R. L. Beirigo, M. G. Todorov, and A. M. S. Barreto, "Transfer on count-based quadratic control of markov jump linear systems with unknown transition probabilities," in *Proc. of the Brazilian Conference on Dynamics, Control & Applications*, (São José do Rio Preto, Brazil), 2017.

[20] 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," *Automatica*, vol. 38, pp. 217–225, 2002.

[21] R. L. Beirigo, M. G. Todorov, and A. M. S. Barreto, "Online TD(λ) for discrete-time Markov jump linear systems," in *Proc. of the 57th IEEE Conference on Decision & Control*, (Miami, USA), submitted, 2018.

[22] L. Zhang, T. Yang, P. Shi, and Y. Zhu, *Analysis and Design of Markov Jump Systems with Complex Transition Probabilities*. Switzerland: Springer, 2016.

[23] J. R. Norris, *Markov Chains*. Cambridge: Cambridge University Press, 1997.

[24] A. A. G. Siqueira, M. H. Terra, and M. Bergerman, *Robust Control of Robots*. Heidelberg: Springer-Verlag, 2011.

[25] A. A. G. Siqueira and M. H. Terra, "Nonlinear and Markovian $\mathcal{H}_\infty$ controls of underactuated manipulators," *IEEE Trans. Control Syst. Technol.*, vol. 12, pp. 811–826, Nov. 2004.

[26] A. A. G. Siqueira and M. H. Terra, "A fault-tolerant manipulator robot based on $\mathcal{H}_2$, $\mathcal{H}_\infty$ and mixed $\mathcal{H}_2/\mathcal{H}_\infty$ Markovian controls," *IEEE/ASME Trans. Mechatronics*, vol. 14, pp. 257–263, Apr. 2009.