Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Error Bounds in the Discretisation of the Input-constrained LQR Problem

José De Doná [*,**] **Claus Müller** [*] **Ryan McCloy** [*]

[*] *School of Electrical Engineering and Computer Science, The University of Newcastle, Callaghan NSW 2308, Australia*
[**] *Corresponding author; email:* Jose.Dedona@newcastle.edu.au

Abstract: This paper studies the discrete-time sampled-data approximation of the input-constrained finite-horizon linear quadratic regulator problem. Explicit estimates for the error in the performance index rendered by the solution of a discrete-time approximation are provided. This result can be used to evaluate the level of sub-optimality of the sampled-data solution corresponding to a given discretisation interval, or to compute a discretisation interval length that guarantees a given approximation error. In addition, the (in general) unknown structure of the solution of the continuous-time constrained linear quadratic regulator problem can be investigated via this discrete-time approximation result. Examples of such a study are provided.

Keywords: sampled-data optimal control, error bounds, discretisation, constrained LQR problem

1. INTRODUCTION

This paper is concerned with the following basic optimal control problem: Given a linear control system $\dot{x}(t) = Ax(t) + Bu(t)$ with initial condition $x(t_0) = x_0$ and input *constraint set* U (assumed to be a closed convex set with $0 \in U$ and $U = -U$), find the control input $u \in L_2([t_0, T], U)$ that minimises a performance index, over the fixed time interval $[t_0, T]$, given by $J_{x_0}(u) = x(T)^T P x(T) + \int_{t_0}^{T} \left(x(t)^T Q x(t) + u(t)^T R u(t) \right) dt$. This problem is known as the (continuous-time) input-constrained Linear Quadratic Regulator (LQR) problem.

Despite some elegant formalisms to deal with constrained optimal control problems (notably, the Pontryagin Maximum Principle and the Hamilton-Jacobi-Bellman equation) no exact solutions to the aforementioned basic constrained LQR problem are known in general, except in a few cases, as in the unconstrained problem (when $U = \mathbb{R}^m$). Thus, in practice, the problem is usually discretised in time and posed as a finite dimension optimisation problem. Typically, the time discretisation of the problem is performed by means of a *sampled-data* (or zero-order hold) control problem, whereby the control inputs are held constant during each discretisation interval. In recent times there has been a substantial amount of research related to fixed horizon discrete-time constrained optimal control problems. Part of this interest is due to the fact that these problems form the main building block of Model Predictive Control (MPC) strategies, one of the most extensively used control techniques in modern industrial applications [see, e.g., Mayne et al. (2000); Rawlings and Mayne (2009)]. This research has lead to a number of well-established methods for computing the solution to finite dimensional constrained optimal control problems, particularly for those that can be casted as a Quadratic Program, ranging from efficient numerical algorithms to explicit solutions [see, for example, Goodwin et al. (2005); Mayne et al. (2000); Seron et al. (2003) and references therein].

These discrete-time solutions are relevant to the original continuous-time problem as long as they provide good approximations to the (typically unknown) continuous-time solution. In Yuz et al. (2005) it has been shown that the optimal performance index achieved by solving the discrete-time sampled-data problem converges (as the length of the discretisation interval tends to zero) to the optimal performance index achievable in the continuous-time framework. However, no rate of convergence is provided in Yuz et al. (2005); nor bounds on the error for a given length of the discretisation interval are computed. So, while the results of Yuz et al. (2005) are reassuring and theoretically important in that they provide a sound foundation to constrained sampled-data optimal control, they do not provide a practical answer as to what level of discretisation one should choose so as to achieve a given degree of approximation to the problem. For other related work concerning discrete-time approximations for time-varying linear systems see for example Dontchev (1981), where an optimal control problem for a system with linear dynamics is discretised using Euler's integration scheme and it is shown that the discretisation error is bounded by a linear function of the length of the discretisation interval. However, no explicit expression for the proportionality constant is provided and hence it is not directly clear how to determine a discretisation interval length that guarantees a given error of approximation.

The main contribution of this paper is to provide an explicit bound on the error in the performance index rendered by the solution of a discrete-time approximation. The error estimation that is derived in this paper has the following form:

$$0 \leq J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t) \bar{u}_j^* \right) - J_{x_0}(u_{x_0}^*) \leq F(A, B, T - t_0, Q, R, P) \cdot \frac{\|x_0\|^2}{L}, \quad (1)$$

where $u_{x_0}^*$ is the (unknown) optimal solution of the constrained continuous-time control problem and $\sum_{j=0}^{L-1} f_j(t) \bar{u}_j^*$ is the ap-

proximated solution rendered by a discrete-time optimisation problem (details are provided in the remainder of the paper). That is, the bound on the error is directly proportional to the square of the size of the initial condition x_0 and inversely proportional to the number of discretisation intervals (degree of approximation) L. (In our notation, $(T-t_0)/L$ denotes the length of the discretisation interval.) The term F in (1) does not depend on the number of discretisation intervals L and only depends on the data of the underlying continuous-time problem: the system model matrices A and B, the length of the time interval $T-t_0$, and the performance index weighting matrices Q, R and P. An explicit expression to compute $F(A,B,T-t_0,Q,R,P)$ is easily obtained from the results presented in this paper and is summarised in Table 1. In addition, an estimation of the norm of the error on the optimal control input solution is also provided, which depends on $1/\sqrt{L}$.

The results in this paper differ from previous existing results in the following way. In contrast with the results in Yuz et al. (2005), we not only establish convergence of the sampled-data optimal solution to the optimal continuous-time solution as the length of the discretisation interval tends to zero, but also provide explicit bounds on the estimation error, of the form (1), that correspond to a given discretisation interval length $(T-t_0)/L$. The main difference with the results in Dontchev (1981) is that we do not use Euler's integration method to approximate the system with a discrete-time one but, since we focus attention on linear time-invariant systems, we compute the exact integral solutions that correspond to *piecewise-constant* [1] input functions. Similarly to the results in Dontchev (1981), we obtain error bounds given by a linear function of the length of the discretisation interval $(T-t_0)/L$. In addition, we provide explicit expressions to compute the term $F(A,B,T-t_0,Q,R,P)$ in (1). The practical relevance of expression (1) is that given a set of problem data, it is straightforward to select the number of sampling intervals L [equivalently the discretisation interval length $(T-t_0)/L$] so as to achieve a desired degree of approximation to the continuous-time solution. Moreover, the rate of convergence to the continuous-time optimal solution as L is increased can be directly estimated from expression (1). Expression (1) also allows to evaluate the level of sub-optimality of the sampled-data solution that corresponds to a given discretisation interval length.

The layout of the remainder of the paper is as follows. In Section 2 we formulate the problem and present some preliminary definitions. In Section 3 we show existence and uniqueness of the solution of the continuous-time problem and prove the continuity of the performance index J_{x_0}. In Section 4 we show that the optimal control and corresponding adjoint trajectory are Lipschitz continuous and compute bounds for the relevant Lipschitz constants. The main result of the paper is presented in Section 5, which allows to compute explicitly error estimates for the optimal performance index and the optimal control trajectories that result from optimising a discrete-time problem. In Section 6 we present two examples illustrating the results. In the first example, we use the error bounds obtained in the paper to verify a conjecture that for first-order systems ($n=1$), when $U=[-1,1]$, the optimal control that minimises the performance index considered (with terminal weight given by the infinite-horizon unconstrained optimal value function) is equal to $\mathrm{sat}(Kx)$ [where $\mathrm{sat}(\cdot)$ is the usual saturation function between ± 1 and K is the optimal state feedback gain that solves the 'unconstrained' LQR problem]. To the best of the authors knowledge, this is a result that has not been previously reported in the literature. Here, we verify it by comparing the resulting performance index with the one obtained with the discretised problem, and by using the error bound (converging to zero for L increasing) afforded by the main result (1) of this paper. In the second example we illustrate two situations for a second-order system, one where $\mathrm{sat}(Kx)$ is still the (nontrivial) optimal control solution and one where $\mathrm{sat}(Kx)$ is no longer the optimal solution. Conclusions are provided in Section 7. To avoid interrupting the flow in reading the conceptual ideas contained in the main body, some of the lengthy mathematical proofs have been included in appendices at the end of the paper.

Notation: For a vector $x\in\mathbb{R}^n$ we denote the Euclidean norm $\|x\|:=\sqrt{x^Tx}$. For a matrix $M\in\mathbb{R}^{n\times m}$ we denote the induced norm $\|M\|:=\sup\limits_{0\neq x\in\mathbb{R}^m}\dfrac{\|Mx\|}{\|x\|}$. The nonnegative integers are denoted $\mathbb{N}_0:=\mathbb{N}\cup\{0\}$.

2. PROBLEM FORMULATION AND ADJOINT SYSTEM

In this section we define the optimal control problem and introduce some preliminary definitions. Let $A\in\mathbb{R}^{n\times n}$ and $B\in\mathbb{R}^{n\times m}$ be matrices, let $P,Q\in\mathbb{R}^{n\times n}$ be (symmetric and) positive semidefinite, and let $R\in\mathbb{R}^{m\times m}$ be (symmetric and) positive definite. Let $t_0<T$ be real numbers, and $x_0\in\mathbb{R}^n$ be some initial state. Let $U\subset\mathbb{R}^m$ be a nonempty, closed and convex set. We consider a linear time-invariant system given by

$$\dot{x}(t)=Ax(t)+Bu(t),\quad x(t_0)=x_0. \tag{2}$$

If $u\in L_2([t_0,T],U)$ is some control the corresponding trajectory is given by the Integral Equation $IE(x_0,u)$:

$$x(t)=x_0+\int_{t_0}^{t}\big(Ax(\gamma)+Bu(\gamma)\big)\,d\gamma\quad\text{if } t\in[t_0,T],$$

i.e. $x(t)=e^{(t-t_0)A}x_0+\int_{t_0}^{t}e^{(t-\gamma)A}Bu(\gamma)\,d\gamma$.

The optimisation problem considered here consists in minimising the performance index

$$\begin{aligned}J_{x_0}(u)=&\,x(T)^TPx(T)\\&+\int_{t_0}^{T}\big(x(t)^TQx(t)+u(t)^TRu(t)\big)\,dt,\end{aligned} \tag{3}$$

i.e. to find $u^*\in L_2([t_0,T],U)$ that minimises J_{x_0}.

If $u\in L_2([t_0,T],U)$ is some control and x given by $IE(x_0,u)$ is the corresponding trajectory, the Cauchy problem $AS(x_0,u)$ given by

$$\begin{cases}\dot{\psi}(t)=2Qx(t)-A^T\psi(t)\ \text{ if }\ t\in[t_0,T]\\ \psi(T)=-2Px(T)\end{cases}$$

is called the Adjoint System. The solution of $AS(x_0,u)$ is

$$\psi(t)=-2e^{(T-t)A^T}Px(T)-2\int_{t}^{T}e^{(\gamma-t)A^T}Qx(\gamma)\,d\gamma. \tag{4}$$

Moreover it is a straightforward calculation that if we define the negative semidefinite matrix

[1] Considering piecewise-constant input functions is in agreement with usual practice in sampled-data systems and is standard in technological implementations (e.g., using a 'zero-order hold' device).

$$h(t) := -\int_t^T e^{\gamma A^T} Q e^{\gamma A}\, d\gamma - e^{TA^T} P e^{TA} \in \mathbb{R}^{n\times n}$$

and the mapping

$$k(t,\beta) := \begin{cases} e^{-tA^T} h(t) e^{-\beta A} & \text{if } \beta \le t \\ e^{-tA^T} h(\beta) e^{-\beta A} & \text{if } t \le \beta \end{cases} \tag{5}$$

then,

$$\frac{\psi(t)}{2} = k(t,t_0)x_0 + \int_{t_0}^T k(t,\beta) B u(\beta)\, d\beta.$$

3. EXISTENCE AND UNIQUENESS OF THE OPTIMAL SOLUTION

The following theorem provides a preliminary result for the error estimation that will be derived later. It is presented here since it also shows that J_{x_0} is continuous.

Theorem 1. Let $u_1, u_2 \in L_2([t_0,T],\mathbb{R}^m)$. Define:

$$\lambda := \frac{1}{2}\lambda_{\max}(A + A^T),$$

the half of the biggest eigenvalue of $A + A^T$, and

$$\eta(\lambda) := \int_{t_0}^T e^{2\lambda(T-\beta)}\, d\beta = \int_{t_0}^T e^{2\lambda(\beta - t_0)}\, d\beta$$
$$= \begin{cases} T - t_0 & \text{if } \lambda = 0 \\ \dfrac{e^{2\lambda(T-t_0)} - 1}{2\lambda} & \text{if } \lambda \neq 0 \end{cases}.$$

Moreover define

$$\eta_1(\lambda) := 2\|B\| \left(\|Q\| \sqrt{T - t_0}\, \eta(\lambda) + \|P\| e^{\lambda(T-t_0)} \sqrt{\eta(\lambda)} \right),$$

and

$$\eta_2(\lambda) := \|R\| + \|Q\| \cdot \|B\|^2 (T - t_0)\eta(\lambda) + \|P\| \cdot \|B\|^2 \eta(\lambda).$$

Then

$$|J_{x_0}(u_1) - J_{x_0}(u_2)| \le \|x_0\| \cdot \|u_1 - u_2\|_{L_2} \eta_1(\lambda) + \|u_1 - u_2\|_{L_2} \|u_1 + u_2\|_{L_2} \eta_2(\lambda).$$

Proof. The proof is included in Appendix A. □

It is a known fact that J_{x_0} is strictly convex and thus we omit the proof. Hence, it follows from the convexity of U that there is at most one optimal control in $L_2([t_0,T],U)$. Existence of a solution follows for example from Theorem 4.2.1 and Lemma 4.2.2 in Jost and Li-Jost (1998).

The following result provides an expression for the optimisation problem derived from the Maximum Principle.

Theorem 2. The optimal control solution $u^*_{x_0} := \arg\min_{u \in L_2([t_0,T],U)} J_{x_0}(u)$ is the unique solution of the equation:

$$u^*_{x_0}(t) = \arg\min_{v \in U} \left(v^T R v - v^T B^T \psi^*_{x_0}(t) \right)$$
$$= R^{-1/2} pr_{R^{1/2}U} \left(\frac{1}{2} R^{-1/2} B^T \psi^*_{x_0}(t) \right), \tag{6}$$

where $\psi^*_{x_0}$ solves $AS(x_0, u^*_{x_0})$, and $pr_{R^{1/2}U}$ is the Euclidean projection on the set $R^{1/2}U$.

Proof. The proof is included in Appendix B. □

4. REGULARITY OF THE OPTIMAL CONTROL

From the expression for the optimiser (6) and the convexity of U it follows that the optimal solution $u^*_{x_0}$ of the constrained optimal control problem is continuous. In this section we prove that it is in fact Lipschitz continuous and give an estimation of the Lipschitz constant. First, we present the following result concerning the projection operator that appears in expression (6).

Lemma 3. We have for all $z_1, z_2 \in \mathbb{R}^m$

$$\left\| R^{-1/2} pr_{R^{1/2}U} \left(\frac{1}{2} R^{-1/2} z_1 \right) - R^{-1/2} pr_{R^{1/2}U} \left(\frac{1}{2} R^{-1/2} z_2 \right) \right\| \le \frac{\|z_1 - z_2\|}{\lambda_{\min}(R)}.$$

Proof. The proof is included in Appendix C. □

The following result provides bounds for the optimal control $u^*_{x_0}$, the optimal state $x^*_{x_0}$ and the optimal adjoint system $\psi^*_{x_0}$ trajectories.

Lemma 4. Let U be as before with $0 \in U$, and define,

$$\delta(\lambda) := \|P\| e^{2\lambda(T-t_0)} + \|Q\| \eta(\lambda),$$
$$\varepsilon(\lambda) := \max\left\{1, e^{\lambda(T-t_0)}\right\} + \|B\| \left(\frac{\delta(\lambda)\eta(\lambda)}{\lambda_{\min}(R)} \right)^{1/2},$$

and

$$\varphi(\lambda) := \varepsilon(\lambda) \left(\|P\| \max\left\{1, e^{\lambda(T-t_0)}\right\} + \|Q\| \sqrt{(T-t_0)\eta(\lambda)} \right).$$

Then we have:

1) $\|u^*_{x_0}\|^2_{L_2} \le \frac{\|x_0\|^2}{\lambda_{\min}(R)} \delta(\lambda)$.
2) $\|x^*_{x_0}\|_\infty \le \varepsilon(\lambda)\|x_0\|$.
3) $\|\psi^*_{x_0}\|_\infty \le 2\varphi(\lambda)\|x_0\|$.

Proof. The proof is included in Appendix D. □

The following corollary provides explicit estimates of the Lipschitz constants. We recall that, for a vector valued function $f : [t_0,T] \to \mathbb{R}^n$ the Lipschitz constant is defined as $\|f\|_{Lip} := \inf\{c > 0 : \|f(t) - f(s)\| \le c|t - s|,\ \forall\, t,s \in [t_0,T]\}$.

Corollary 5. We have

1) $\|\psi^*_{x_0}\|_{Lip} \le 2\|x_0\|(\|Q\|\varepsilon(\lambda) + \|A\|\varphi(\lambda))$.
2) $\|u^*_{x_0}\|_{Lip} \le \frac{\|B\|\|\psi^*_{x_0}\|_{Lip}}{\lambda_{\min}(R)}$.

Proof.

1) follows from the equation $\dot{\psi^*_{x_0}}(t) = 2Qx^*_{x_0}(t) - A^T\psi^*_{x_0}(t)$, and Lemma 4.

2) follows from Theorem 2 and Lemma 3. □

5. APPROXIMATION AND ERROR ESTIMATION

This section presents the main result of the paper; namely, the estimation of the error on the optimal performance index and on the optimal control trajectory that result from optimising a zero-order hold (zoh) discrete-time approximation of the problem. (This problem is also known as optimal sampled-data problem.) Let $\mathbf{1}_M$ denote the characteristic function on the set M. In order to approximate the continuous-time system by a zoh discrete-time system with a given number of *sampling intervals* $L \in \mathbb{N}$ we define $T_i = t_0 + \frac{i}{L}(T - t_0)$ for $0 \leq i \leq L$ and $f_j(t) := \sqrt{\frac{L}{T-t_0}} \cdot \mathbf{1}_{[T_j, T_{j+1})} \in L_2([t_0, T], \mathbb{R})$ for $0 \leq j \leq L-1$. The control inputs considered in the discrete-time approximation consist of piecewise-constant functions

$$u(t) = \sum_{j=0}^{L-1} f_j(t)\bar{u}_j, \tag{7}$$

with $\bar{u}_j \in \mathbb{R}^m$. An easy calculation shows that the zoh discrete-time performance index (that is, the performance index (3) evaluated on a piecewise constant function (7)) is given by the following quadratic expression: $J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j\right) = -{x_0}^T k(t_0,t_0)x_0 - 2\bar{u}^T \mathcal{A}x_0 + \bar{u}^T \mathcal{B}\bar{u}$, where $\bar{u} = (\bar{u}_0, \ldots, \bar{u}_{L-1})^T$,

$$\mathcal{A} = \begin{pmatrix} \mathcal{A}_0 \\ \ldots \\ \mathcal{A}_{L-1} \end{pmatrix} \in \mathbb{R}^{mL \times n},$$

$$\mathcal{A}_j = B^T \int_{t_0}^{T} f_j(t) k(t, t_0)\, dt \in \mathbb{R}^{m \times n},$$

$$\mathcal{B} = \begin{pmatrix} \mathcal{B}_{0,0} & \ldots & \mathcal{B}_{0,L-1} \\ \ldots & \ldots & \ldots \\ \mathcal{B}_{L-1,0} & \ldots & \mathcal{B}_{L-1,L-1} \end{pmatrix} \in \mathbb{R}^{mL \times mL},$$

$$\mathcal{B}_{j,k} = \int_{t_0}^{T} f_j(t) f_k(t)\, dtR$$

$$-B^T \int_{t_0}^{T}\int_{t_0}^{T} f_j(\delta) f_k(\gamma) k(\delta,\gamma)\, d\delta\, d\gamma B \in \mathbb{R}^{m \times m},$$

and where the matrices $k(\delta, \gamma)$ are defined in (5). The matrix $\mathcal{B}$ is (symmetric and) positive definite.

We also define the constraint set $\overline{U} = \sqrt{\frac{T-t_0}{L}} \cdot U$ and consider the following optimisation problem:

$$\min_{\bar{u} \in \overline{U}^L} -{x_0}^T k(t_0,t_0)x_0 - 2\bar{u}^T \mathcal{A}x_0 + \bar{u}^T \mathcal{B}\bar{u}. \tag{8}$$

Notice that when the set U (and hence $\overline{U}$) is a polytope (i.e., a set defined by linear inequalities) then problem (8) is a Quadratic Program that can be solved with a number of efficient numerical algorithms.

The following theorem provides an estimation of the error between the optimal (unknown) continuous-time solution and the (approximated) zoh discrete-time solution rendered by solving problem (8).

Theorem 6. Let $U \subset \mathbb{R}^m$ be closed and convex with $0 \in U = -U$. Let $\overline{U} = \sqrt{\frac{T-t_0}{L}} \cdot U$ and let $\bar{u}^* = (\bar{u}_0^*, \ldots, \bar{u}_{L-1}^*)^T \in \overline{U}^L$ be the solution of the discrete-time optimal control problem consisting in the Quadratic Program (8).

Then we have $\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^* \in L_2([t_0, T], U)$, and if $u_{x_0}^* \in L_2([t_0, T], U)$ is the optimal solution of the constrained (continuous-time) optimal control problem $\min_{u \in L_2([t_0,T],U)} J_{x_0}(u)$ we have the error estimation:

$$0 \leq J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right) - J_{x_0}(u_{x_0}^*) \leq \frac{(T-t_0)^{3/2}}{2\sqrt{3}L} \cdot \|u_{x_0}^*\|_{Lip}\Big[\|x_0\|\eta_1(\lambda) + 2\|u_{x_0}^*\|_{L_2}\eta_2(\lambda)\Big].$$

Proof. The proof is included in Appendix E. □

Remark 7. Note from the expression of the error estimation in Theorem 6 that, using the bound for $\|u_{x_0}^*\|_{Lip}$ computed in Corollary 5, the bound for $\|u_{x_0}^*\|_{L_2}$ computed in Lemma 4, and the expressions of $\eta_1(\lambda)$ and $\eta_2(\lambda)$ given in Theorem 1, the expression for the term $F(A, B, T - t_0, Q, R, P)$ in expression (1) can be readily obtained as a function of the data of the underlying continuous-time problem: the system model matrices A and B, the length of the time interval $T - t_0$, and the performance index weighting matrices Q, R and P. The required computations are summarised in Table 1.

Finally, we present an estimation of the error between the optimal continuous-time control input $u_{x_0}^*$ and an arbitrary input $u \in L_2([t_0, T], U)$ provided the performance index of u is within ϵ of the optimal performance index. This result is particularly useful in estimating the error rendered by the zoh discrete-time approximation $u(t) = \sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*$, obtained from the solution to (8), since we have already obtained in Theorem 6 an estimation for the performance index error.

Theorem 8. Assume that $0 \leq J_{x_0}(u) - J_{x_0}(u_{x_0}^*) \leq \varepsilon$.

Then $\|u - u_{x_0}^*\|_{L_2}^2 \leq \varepsilon / \lambda_{\min}(R)$.

Proof. The proof is included in Appendix F. □

6. EXAMPLES

The results obtained thus far are for any symmetric positive semidefinite matrix P in the 'terminal cost' term of the performance index (3). In the examples presented in this section, we make the following additional assumption.

Assumption 9. The matrix P in the 'terminal cost' term in (3) is given by the solution of the algebraic Riccati equation:

$$PA + A^T P - PBR^{-1}B^T P + Q = 0.$$

Remark 10. Assumption 9 is a reasonable choice for a number of good reasons; including:

(i) Performance: In the case when the 'terminal state' at the end of the time-horizon, $x(T)$, is in the region where the constraints are no longer active, the terminal cost term $x(T)^T P x(T)$ in (3) is the infinite-horizon optimal performance index—from initial state $x(T)$—and, hence, we recover infinite-horizon optimal performance in the finite-horizon optimal control problem.

(ii) Stability: The choice of the algebraic Riccati solution for the terminal weight matrix can be effectively used to prove stability of receding horizon implementations [see, e.g., Goodwin et al. (2005); Mayne et al. (2000)].

$\lambda = \frac{1}{2}\lambda_{\max}(A + A^T), \qquad \eta(\lambda) = \begin{cases} T - t_0 & \text{if } \lambda = 0, \\ \dfrac{e^{2\lambda(T-t_0)} - 1}{2\lambda} & \text{if } \lambda \neq 0, \end{cases}$ ($\lambda_{\max}(\cdot)$ denotes maximum eigenvalue)
$\eta_1(\lambda) = 2\|B\| \left(\|Q\|\sqrt{T - t_0}\eta(\lambda) + \|P\|e^{\lambda(T-t_0)}\sqrt{\eta(\lambda)} \right),$
$\eta_2(\lambda) = \|R\| + \|Q\| \cdot \|B\|^2 (T - t_0)\eta(\lambda) + \|P\| \cdot \|B\|^2 \eta(\lambda),$
$\varepsilon(\lambda) = \max\left\{1, e^{\lambda(T-t_0)}\right\} + \|B\| \sqrt{\frac{\|P\|e^{2\lambda(T-t_0)}\eta(\lambda) + \|Q\|\eta(\lambda)^2}{\lambda_{\min}(R)}},$ ($\lambda_{\min}(\cdot)$ denotes minimum eigenvalue)
$\varphi(\lambda) = \varepsilon(\lambda) \left(\|P\| \max\left\{1, e^{\lambda(T-t_0)}\right\} + \|Q\| \sqrt{(T - t_0)\eta(\lambda)} \right),$
$F(A, B, T - t_0, Q, R, P) = \frac{(T-t_0)^{3/2}\|B\|(\|Q\|\varepsilon(\lambda) + \|A\|\varphi(\lambda))}{\lambda_{\min}(R)\sqrt{3}} \left(\eta_1(\lambda) + 2\sqrt{\frac{\|P\|e^{2\lambda(T-t_0)} + \|Q\|\eta(\lambda)}{\lambda_{\min}(R)}}\eta_2(\lambda) \right).$

Table 1. Summary of the computation of the constant term $F(A, B, T - t_0, Q, R, P)$ in expression (1), which is a function of only the data of the underlying continuous-time problem: system model matrices A and B, length of time interval $T - t_0$, and performance weighting matrices Q, R and P. (Note that this term does not depend on the discrete-time approximation—in particular, it does not depend on the number of discretisation intervals L—nor does it depend on the magnitude of the initial condition x_0.)

6.1 First-order system

In this first example, we use the error bound (1) to verify a conjecture [2] that, for first-order systems with constraint set $U = [-1, 1]$, the optimal control that minimises *globally* (i.e., for any initial condition $x_0 \in \mathbb{R}$) the quadratic performance index (3) with terminal weight matrix P chosen as in Assumption 9 is equal to $\text{sat}(Kx)$ [where $\text{sat}(\cdot)$ is the usual saturation function between -1 and 1, and K is the optimal state feedback gain that solves the 'unconstrained' LQR problem corresponding to the terminal weight matrix P; i.e., $K = -R^{-1}B^T P$]. For system (2) with $A = -0.5$, $B = 1$, initial condition $x_0 = 5$, and performance index (3) with $Q = 1$, $R = 1$, $P = 0.618$ (solution of the algebraic Riccati equation), $K = -0.618$, and $t_0 = 0$, $T = 5$, we computed the upper bound in (1), which resulted in $F = 552$. (Notice that, with the initial condition considered, $x_0 = 5$, the unconstrained control law has initial value $Kx_0 = -3.09$ and hence it exceeds the allowed range $[-1, 1]$; in other words, the case considered is not a trivial one where the unconstrained solution solves the problem.) In Figure 1 we have plotted the error upper bound $F\|x_0\|^2/L$ (right hand side of (1)) and the difference between $J_1 := J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right)$ and $J_2 := J_{x_0}(\text{sat}(Kx))$. It is verified that $J_1 - J_2$ is always under the error upper bound $F\|x_0\|^2/L$. From (1) we have $\left|J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right) - J_{x_0}(u_{x_0}^*)\right| \leq F\|x_0\|^2/L$, and from the simulation (Figure 1) we have $\left|J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right) - J_{x_0}(\text{sat}(Kx))\right| \leq F\|x_0\|^2/L$. We then have, from the triangular inequality, that $\left|J_{x_0}(\text{sat}(Kx)) - J_{x_0}(u_{x_0}^*)\right| \leq 2F\|x_0\|^2/L$, for all $L \in \mathbb{N}$. Since $2F\|x_0\|^2/L$ converges to zero as $L \to \infty$, we then have that $J_{x_0}(\text{sat}(Kx)) = J_{x_0}(u_{x_0}^*)$. We thus conclude, from the uniqueness of the optimal control (and, also, from the result of Theorem 8), that $u_{x_0}^* = \text{sat}(Kx)$, and the conjecture is verified. In Figure 2, the performance indices $J_1 := J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right)$ and $J_2 := J_{x_0}(\text{sat}(Kx))$ are plotted, where the convergence of the discretised solution to the conjectured optimal continuous-time solution can also be verified.

[2] To the best of the authors' knowledge, this result has hitherto not been reported in the literature.

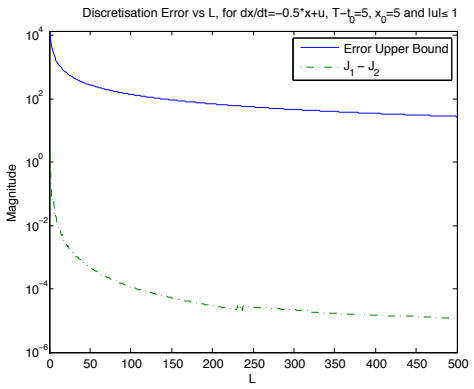

Fig. 1. Error upper bound $F\|x_0\|^2/L$ and difference between $J_1 := J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right)$ and $J_2 := J_{x_0}(\text{sat}(Kx))$ versus L, for first-order system example.

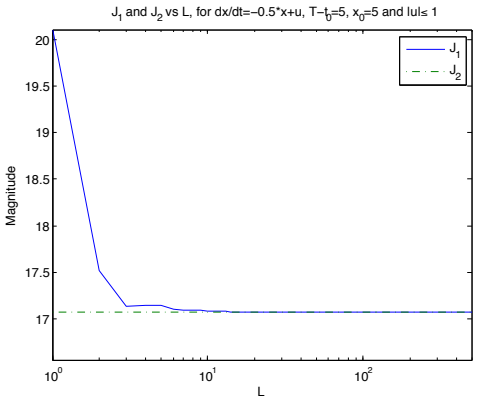

Fig. 2. $J_1 := J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right)$ and $J_2 := J_{x_0}(\text{sat}(Kx))$ versus L, for first-order system example.

6.2 Second-order system

We explore here a second-order system consisting in a double integrator, with $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $R = 1$, $P = \begin{bmatrix} 1.7321 & 1 \\ 1 & 1.7321 \end{bmatrix}$ (solution of the algebraic Riccati equation), $K = [-1 \quad -1.7321]$, $U = [-1, 1]$, $t_0 = 0$ and $T = 10$. In Figure 3, the performance indices $J_1 := J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right)$ and $J_2 := J_{x_0}(\mathrm{sat}(Kx))$ are plotted for

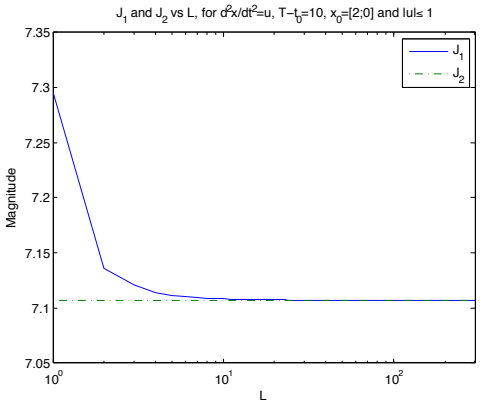

Fig. 3. $J_1 := J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right)$ and $J_2 := J_{x_0}(\mathrm{sat}(Kx))$ versus L, for second-order system example with initial condition: $x_0 = [2 \quad 0]^T$.

the case of an initial condition $x_0 = [2 \quad 0]^T$, where it can be seen (using the convergence of the discretised solution to the optimal continuous-time solution) that, as in the case of the first-order system, $u_{x_0}^* = \mathrm{sat}(Kx)$. Notice again that, with $x_0 = [2 \quad 0]^T$, the initial value of the unconstrained control law is $Kx_0 = -2$, thus exceeding the allowed range $[-1, 1]$ and making this a nontrivial case. Contrary to first-order systems, this situation cannot be expected to hold 'globally' (i.e., for any initial condition $x_0 \in \mathbb{R}^2$), and in Figure 4 both performance indices J_1 and J_2 are plotted for an initial condition $x_0 =$

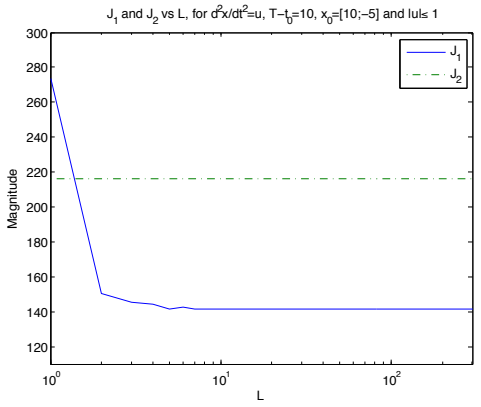

Fig. 4. $J_1 := J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right)$ and $J_2 := J_{x_0}(\mathrm{sat}(Kx))$ versus L, for second-order system example with initial condition: $x_0 = [10 \quad -5]^T$.

$[10 \quad -5]^T$, where it can be seen that the discretised solution converges to the (unknown) continuous-time optimal solution which is lower than the one corresponding to $\mathrm{sat}(Kx)$ (i.e., the latter control is no longer optimal in this case).

7. CONCLUSIONS

This paper has explored the discrete-time sampled-data approximation of the input-constrained finite-horizon linear quadratic regulator problem. The main result consists in providing explicit expressions for estimates of the error rendered by the solution of the discrete-time approximation. This result can be used, for example, to evaluate the level of sub-optimality of the sampled-data solution that corresponds to a given length of the discretisation interval, or to compute a discretisation interval length that guarantees a given error of approximation. Moreover, it also allows to estimate the rate of convergence to the continuous-time optimal solution as the length of the discretisation interval is decreased to zero. Two examples illustrating the results of the paper have been presented, a first-order system example where a conjecture that the optimal control solution is $\mathrm{sat}(Kx)$ was verified, and a second-order system with an initial condition for which $\mathrm{sat}(Kx)$ is still the (nontrivial) optimal control solution and with a different initial condition for which $\mathrm{sat}(Kx)$ is no longer the optimal control solution.

REFERENCES

Afanasev, V., Kolmanovskii, V., and Nosov, V. (1996). *Mathematical Theory of Control Systems Design*. Kluwer.

Dontchev, A. (1981). Error estimates for a discrete approximation to constrained control problems. *SIAM J. of Numerical Analysis*, 18(3), 500–514.

Goodwin, G., Seron, M., and De Doná, J. (2005). *Constrained Control and Estimation. An Optimisation Approach*. Springer-Verlag, London.

Jost, J. and Li-Jost, X. (1998). *Calculus of variations*. Cambridge University Press.

Kantorovich, L. and Akilov, G. (1964). *Functional Analysis in Normed Spaces*. Pergamon Press.

Mayne, D., Rawlings, J., Rao, C., and Scokaert, P. (2000). Constrained model predictive control: Stability and optimality. *Automatica*, 36, 789–814.

Rawlings, J. and Mayne, D. (2009). *Model Predictive Control: Theory and Design*. Nob Hill Publishing.

Seron, M., Goodwin, G., and De Doná, J. (2003). Characterisation of receding horizon control for constrained linear systems. *Asian Journal of Control*, 5(2), 271–286.

Yuz, J., Goodwin, G., Feuer, A., and De Doná, J. (2005). Control of constrained linear systems using fast sampling rates. *Systems & Control Letters*, 54, 981–990.

Appendix A. PROOF OF THEOREM 1

Proof. If x_i is given by $IE(x_0, u_i)$, $i = 1, 2$, then we have:

$$\begin{aligned}\|x_1(t) + x_2(t)\| &\le 2\|x_0\|e^{\lambda(t-t_0)} \\ &\quad + \int_{t_0}^{t} \|e^{(t-\gamma)A}B\| \cdot \|u_1(\gamma) + u_2(\gamma)\|\, d\gamma \\ &\le 2\|x_0\|e^{\lambda(t-t_0)} \\ &\quad + \|B\| \cdot \|u_1 + u_2\|_{L_2} \left(\int_{t_0}^{t} e^{2\lambda(t-\gamma)}\, d\gamma\right)^{1/2},\end{aligned}$$

where we used Hölder's inequality and the fact that if $\mu \geq 0$ then $\|e^{\mu A}\| = e^{\mu\lambda}$. Moreover,

$$\|x_1(t) - x_2(t)\| \leq \int_{t_0}^{t} \|B\| \cdot \|e^{(t-\gamma)A}\| \cdot \|u_1(\gamma) - u_2(\gamma)\| \, d\gamma$$
$$\leq \|u_1 - u_2\|_{L_2} \|B\| \left(\int_{t_0}^{t} e^{2\lambda(t-\gamma)} \, d\gamma \right)^{1/2}.$$

Thus,

$$\left| \int_{t_0}^{T} \left(x_1(t)^T Q x_1(t) - x_2(t)^T Q x_2(t) \right) dt \right|$$
$$= \left| \int_{t_0}^{T} (x_1(t) - x_2(t))^T Q (x_1(t) + x_2(t)) \, dt \right|$$
$$\leq \|Q\| \cdot \|u_1 - u_2\|_{L_2} \|B\| \int_{t_0}^{T} \left(\int_{t_0}^{t} e^{2\lambda(t-\gamma)} \, d\gamma \right)^{1/2}$$
$$\cdot \Bigg(2\|x_0\| e^{\lambda(t-t_0)} + \|B\| \cdot \|u_1 + u_2\|_{L_2}$$
$$\left(\int_{t_0}^{t} e^{2\lambda(t-\gamma)} \, d\gamma \right)^{1/2} \Bigg) \, dt$$
$$\leq \|Q\| \cdot \|u_1 - u_2\|_{L_2} \|B\| \cdot 2 \cdot \|x_0\|$$
$$\left(\int_{t_0}^{T} \int_{t_0}^{t} e^{2\lambda(t-\gamma)} \, d\gamma \, dt \right)^{1/2} \left(\int_{t_0}^{T} e^{2\lambda(t-t_0)} \, dt \right)^{1/2} + \|Q\|$$
$$\cdot \|u_1 - u_2\|_{L_2} \|u_1 + u_2\|_{L_2} \|B\|^2 \int_{t_0}^{T} \int_{t_0}^{t} e^{2\lambda(t-\gamma)} \, d\gamma \, dt.$$

Since, for $t \in [t_0, T]$, $\int_{t_0}^{t} e^{2\lambda(t-\gamma)} \, d\gamma \leq \eta(\lambda)$ we obtain:

$$\left| \int_{t_0}^{T} \left(x_1(t)^T Q x_1(t) - x_2(t)^T Q x_2(t) \right) dt \right|$$
$$\leq 2\|x_0\| \cdot \|Q\| \cdot \|B\| \cdot \|u_1 - u_2\|_{L_2} \sqrt{T - t_0} \eta(\lambda)$$
$$+ \|Q\| \cdot \|B\|^2 \cdot \|u_1 - u_2\|_{L_2} \|u_1 + u_2\|_{L_2} (T - t_0) \eta(\lambda).$$

Moreover,

$$\left| x_1(T)^T P x_1(T) - x_2(T) P x_2(T) \right|$$
$$= \left| (x_1(T) - x_2(T))^T P (x_1(T) + x_2(T)) \right|$$
$$\leq \|P\| \cdot \|u_1 - u_2\|_{L_2} \|B\| \sqrt{\eta(\lambda)} \Big(2\|x_0\| e^{\lambda(T-t_0)}$$
$$+ \|B\| \cdot \|u_1 + u_2\|_{L_2} \sqrt{\eta(\lambda)} \Big),$$

and finally,

$$\left| \int_{t_0}^{T} \left(u_1(t)^T R u_1(t) - u_2(t)^T R u_2(t) \right) dt \right|$$
$$\leq \int_{t_0}^{T} \|R\| \cdot \|u_1(t) - u_2(t)\| \cdot \|u_1(t) + u_2(t)\| \, dt$$
$$\leq \|R\| \cdot \|u_1 - u_2\|_{L_2} \|u_1 + u_2\|_{L_2}.$$

The assertion follows from these estimations. □

Appendix B. PROOF OF THEOREM 2

Proof. The Maximum Principle [see for example Afanasev et al. (1996)] shows that

$$\int_{t_0}^{T} \frac{\partial H}{\partial u}(t, x^*_{x_0}(t), u^*_{x_0}(t), \psi^*_{x_0}(t)) \cdot (v(t) - u^*_{x_0}(t)) \, dt \leq 0, \tag{B.1}$$

for all $v \in L_2([t_0, T], U)$, where $x^*_{x_0}$ is the optimal trajectory, and the Hamiltonian is given by $H(t, x, u, \lambda) = \lambda^T (Ax + Bu) - x^T Q x - u^T R u$. We note that we cannot use the more familiar form of the Maximum Principle (see Afanasev et al. (1996)), since this requires the pre-knowledge that the optimal control is continuous from the right.

From (B.1) we have, for all $v \in L_2([t_0, T], U)$,

$$\int_{t_0}^{T} \left(\psi^*_{x_0}(t)^T B - 2u^*_{x_0}(t)^T R \right) v(t) \, dt$$
$$\leq \int_{t_0}^{T} \left(\psi^*_{x_0}(t)^T B - 2u^*_{x_0}(t)^T R \right) u^*_{x_0}(t) \, dt.$$

Therefore

$$\int_{t_0}^{T} \left(\psi^*_{x_0}(t)^T B u^*_{x_0}(t) - u^*_{x_0}(t)^T R u^*_{x_0}(t) \right) dt \geq$$
$$\int_{t_0}^{T} \left(\psi^*_{x_0}(t)^T B v(t) - 2u^*_{x_0}(t)^T R v(t) + u^*_{x_0}(t)^T R u^*_{x_0}(t) \right) dt$$
$$= \int_{t_0}^{T} \left(\psi^*_{x_0}(t)^T B v(t) - v(t)^T R v(t) \right.$$
$$\left. + (u^*_{x_0}(t) - v(t))^T R (u^*_{x_0}(t) - v(t)) \right) dt$$
$$\geq \int_{t_0}^{T} \left(\psi^*_{x_0}(t)^T B v(t) - v(t)^T R v(t) \right) dt.$$

Thus,

$$\int_{t_0}^{T} \left(\psi^*_{x_0}(t)^T B \overline{v}(t) - \overline{v}(t)^T R \overline{v}(t) \right) dt$$
$$\leq \int_{t_0}^{T} \left(\psi^*_{x_0}(t)^T B u^*_{x_0}(t) - u^*_{x_0}(t)^T R u^*_{x_0}(t) \right) dt,$$

where $\overline{v}(t) := arg\max_{v \in U} \psi^*_{x_0}(t)^T B v - v^T R v$. Since for almost all $t \in [t_0, T]$, $\psi^*_{x_0}(t)^T B u^*_{x_0}(t) -$

$u^*_{x_0}(t)^T R u^*_{x_0}(t) \le \psi^*_{x_0}(t)^T B\overline{v}(t) - \overline{v}(t)^T R\overline{v}(t)$, we conclude that $\psi^*_{x_0}(t)^T B u^*_{x_0}(t) - u^*_{x_0}(t)^T R u^*_{x_0}(t) = \psi^*_{x_0}(t)^T B\overline{v}(t) - \overline{v}(t)^T R\overline{v}(t)$, thus $u^*_{x_0}(t) = \overline{v}(t)$ almost everywhere. □

Appendix C. PROOF OF LEMMA 3

Proof. If $z \in \mathbb{R}^m$ define $f_z : \mathbb{R}^m \to \mathbb{R}$ by $f_z(w) = w^T R w - w^T z$. Then $w_z := R^{-1/2} pr_{R^{1/2}U}\left(\frac{1}{2}R^{-1/2}z\right) = arg\min_{w\in U} f_z(w)$.

If $z_1, z_2 \in \mathbb{R}^m$ we have

$$\begin{aligned}0 &\le f_{z_1}(w_{z_2}) - f_{z_1}(w_{z_1})\\ &\le f_{z_1}(w_{z_2}) - f_{z_1}(w_{z_1}) + f_{z_2}(w_{z_1}) - f_{z_2}(w_{z_2})\\ &= w_{z_2}^T(z_2 - z_1) + w_{z_1}^T(z_1 - z_2)\\ &= (w_{z_1} - w_{z_2})^T(z_1 - z_2).\end{aligned}$$

Moreover, $f_{z_1}(w_{z_2}) = f_{z_1}(w_{z_1}) + \nabla f_{z_1}(w_{z_1})\cdot(w_{z_2} - w_{z_1}) + (w_{z_2} - w_{z_1})^T R(w_{z_2} - w_{z_1})$, thus

$$\begin{aligned}0 &\le \nabla f_{z_1}(w_{z_1})\cdot(w_{z_2} - w_{z_1}) + (w_{z_2} - w_{z_1})^T R(w_{z_2} - w_{z_1})\\ &\le (w_{z_1} - w_{z_2})^T(z_1 - z_2).\end{aligned}$$

Since w_{z_1} is the global minimiser of f_{z_1} on U we have $\nabla f_{z_1}(w_{z_1})\cdot(w_{z_2} - w_{z_1}) \ge 0$, thus $(w_{z_1} - w_{z_2})^T(z_1 - z_2) \ge (w_{z_2} - w_{z_1})^T R(w_{z_2} - w_{z_1})$. Hence we have $\|w_{z_1} - w_{z_2}\|\|z_1 - z_2\| \ge (w_{z_1} - w_{z_2})^T(z_1 - z_2) \ge (w_{z_1} - w_{z_2})^T R(w_{z_1} - w_{z_2}) \ge \|w_{z_1} - w_{z_2}\|^2\lambda_{\min}(R)$, and we conclude that $\|w_{z_1} - w_{z_2}\| \le \|z_1 - z_2\|/\lambda_{\min}(R)$. □

Appendix D. PROOF OF LEMMA 4

Proof.

1) Since $x^T R x \ge \|x\|^2\lambda_{\min}(R)$ for all $x \in \mathbb{R}^m$, we have

$$\begin{aligned}\lambda_{\min}(R)\|u^*_{x_0}\|^2_{L_2} &\le \int_{t_0}^{T} u^*_{x_0}(t)^T R u^*_{x_0}(t)\,dt \le J_{x_0}(u^*_{x_0})\\ &\le J_{x_0}(0) = x(T)^T P x(T) + \int_{t_0}^{T} x(t)^T Q x(t)\,dt,\end{aligned}$$

where $x(t) = e^{(t-t_0)A}x_0$. We also have

$$\begin{aligned}J_{x_0}(0) &\le \|P\|\cdot\|x_0\|^2 e^{2\lambda(T-t_0)} + \|Q\|\cdot\|x_0\|^2\int_{t_0}^{T} e^{2\lambda(t-t_0)}\,dt\\ &= \|x_0\|^2\delta(\lambda),\end{aligned}$$

from which 1) follows.

2) From $x^*_{x_0}(t) = e^{(t-t_0)A}x_0 + \int_{t_0}^{t} e^{(t-\gamma)A}Bu^*_{x_0}(\gamma)\,d\gamma$ it follows that

$$\begin{aligned}\|x^*_{x_0}(t)\| &\le \|x_0\|\max_{t\in[t_0,T]} e^{\lambda(t-t_0)}\\ &\quad + \|B\|\cdot\|u^*_{x_0}\|_{L_2}\left(\int_{t_0}^{t} e^{2\lambda(t-\gamma)}\,d\gamma\right)^{1/2}\\ &\le \|x_0\|\max\{1, e^{\lambda(T-t_0)}\}\\ &\quad + \|B\|\cdot\|x_0\|\left(\frac{\delta(\lambda)\eta(\lambda)}{\lambda_{\min}(R)}\right)^{1/2}.\end{aligned}$$

3) We have from (4),

$$\begin{aligned}\|\psi^*_{x_0}(t)\| &\le 2\|P\|e^{\lambda(T-t)}\|x^*_{x_0}\|_\infty\\ &\quad + 2\|x^*_{x_0}\|_\infty\|Q\|\int_{t}^{T} e^{\lambda(\gamma-t)}\,d\gamma\\ &\le 2\|x_0\|\varepsilon(\lambda)\Bigg(\|P\|\max\{1, e^{\lambda(T-t_0)}\}\\ &\quad + \|Q\|\sqrt{T-t_0}\left(\int_{t}^{T} e^{2\lambda(\gamma-t)}\,d\gamma\right)^{1/2}\Bigg).\end{aligned}$$

□

Appendix E. PROOF OF THEOREM 6

Proof. We first show that if $(u_0, \dots, u_{L-1}) \in \overline{U}^L$ then $\sum_{j=0}^{L-1} f_j(t)u_j \in U$ for all $t \in [t_0, T]$. Let $c = \sqrt{\frac{T-t_0}{L}}$. Since $\sum_{j=0}^{L-1} f_j(t)u_j = \sum_{j=0}^{L-1}(cf_j(t))\frac{u_j}{c}$ with $\frac{u_j}{c} \in U$ and since $0 \in U = -U$ is convex it suffices to show that $\sum_{j=0}^{L-1} cf_j(t) = 1$ for all $t \in [t_0, T]$, which is true.

Next we show that if $0 \le l < L$ then $(\langle f_l, u_1^*\rangle, \dots, \langle f_l, u_m^*\rangle)^T \in \overline{U}$, where $\langle\cdot,\cdot\rangle$ denotes the usual scalar product in L_2 and $u^*_{x_0} = (u_1^*, \dots, u_m^*) \in L_2([t_0, T], U)$. To this end we show that if $h : [a, b] \to U$ is continuous then $\int_a^b h(t)\,dt \in (b-a)U$: If $N \in \mathbb{N}^+$ and $r_k^{(N)} := a + \frac{k}{N}(b-a)$ then the sequence $h^{(N)} = \sum_{k=0}^{N-1}\mathbf{1}_{[r_k^{(N)}, r_{k+1}^{(N)})}\cdot h(r_k^{(N)})$ converges in L_1 to h, and $\frac{1}{b-a}\int_a^b h^{(N)}(t)\,dt = \frac{1}{b-a}\sum_{k=0}^{N-1}\frac{b-a}{N}h(r_k^{(N)}) \in U$ since U is convex. Thus $\frac{1}{b-a}\int_a^b h(t)\,dt = \lim_{N\to\infty}\frac{1}{b-a}\int_a^b h^{(N)}(t)\,dt \in U$, since U is closed. Therefore $(\langle f_l, u_1^*\rangle, \dots, \langle f_l, u_m^*\rangle)^T = \frac{1}{c}\int_{T_l}^{T_{l+1}} u^*_{x_0}(t)\,dt = \frac{c}{T_{l+1}-T_l}\int_{T_l}^{T_{l+1}} u^*_{x_0}(t)\,dt \in cU = \overline{U}$.

Now we complete the orthonormal system $(f_j)_{j=0}^{L-1}$ of $L_2([t_0, T], \mathbb{R})$ to some complete orthonormal basis $(f_j)_{j\ge 0}$ (we make a specific choice below).

Then we have, by Theorem 1,

$$\begin{aligned}0 &\le J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\bar{u}_j^*\right) - J_{x_0}(u^*_{x_0})\\ &\le J_{x_0}\left(\sum_{j=0}^{L-1} f_j(t)\begin{pmatrix}\langle f_j, u_1^*\rangle\\ \dots\\ \langle f_j, u_m^*\rangle\end{pmatrix}\right)\\ &\quad - J_{x_0}\left(\sum_{j=0}^{\infty} f_j(t)\begin{pmatrix}\langle f_j, u_1^*\rangle\\ \dots\\ \langle f_j, u_m^*\rangle\end{pmatrix}\right)\\ &\le \left[\|x_0\|\eta_1(\lambda) + 2\|u^*_{x_0}\|_{L_2}\eta_2(\lambda)\right]\end{aligned}$$

$$\left\| \sum_{j=L}^{\infty} f_j(t) \begin{pmatrix} \langle f_j, u_1^* \rangle \\ \cdots \\ \langle f_j, u_m^* \rangle \end{pmatrix} \right\|_{L_2}$$
$$= \left[\|x_0\| \eta_1(\lambda) + 2\|u_{x_0}^*\|_{L_2} \eta_2(\lambda) \right]$$
$$\left(\sum_{j=L}^{\infty} \sum_{k=1}^{m} |\langle f_j, u_k^* \rangle|^2 \right)^{1/2}.$$

Now we choose the Haar system to complete $(f_j)_{j=0}^{L-1}$. It is well known that the functions $(h_p^{(q)})_{p \in \mathbb{N}_0, 1 \le q \le 2^p}$ given by:

$$h_p^{(q)}(t) = \begin{cases} 2^{p/2} & \text{if } t \in \left(\frac{q-1}{2^p}, \frac{q-1/2}{2^p} \right) \\ -2^{p/2} & \text{if } t \in \left(\frac{q-1/2}{2^p}, \frac{q}{2^p} \right) \\ 0 & \text{elsewhere} \end{cases}$$

together with $h_0^{(0)} = \mathbf{1}_{[0,1]}$ form an orthonormal basis in $L_2([0,1], \mathbb{R})$.

If $0 \le l < L$ let $F_l(t) = \frac{L}{T-t_0}(t - T_l)$. Then the functions $\sqrt{\frac{L}{T-t_0}} \cdot h_p^{(q)} \circ F_l$ $(p \in \mathbb{N}_0, 1 \le q \le 2^p)$ together with the constant function $\sqrt{\frac{L}{T-t_0}}$ form an orthonormal basis in $L_2([T_l, T_{l+1}], \mathbb{R})$, thus $(f_j)_{j=0}^{L-1}$ together with $\left(\sqrt{\frac{L}{T-t_0}} \cdot h_p^{(q)} \circ F_l \right) \cdot \mathbf{1}_{[T_l, T_{l+1})}$ $(p \in \mathbb{N}_0, 1 \le q \le 2^p, 0 \le l < L)$ form an orthonormal basis in $L_2([t_0, T], \mathbb{R})$.

With this choice we obtain

$$\sum_{j=L}^{\infty} \sum_{k=1}^{m} |\langle f_j, u_k^* \rangle|^2$$
$$= \sum_{k=1}^{m} \sum_{l=0}^{L-1} \sum_{p=0}^{\infty} \sum_{q=1}^{2^p} \left| \langle \sqrt{\frac{L}{T-t_0}} h_p^{(q)} \circ F_l, u_k^* \rangle_l \right|^2,$$

where $\langle \cdot, \cdot \rangle_l$ is the scalar product in $L_2([T_l, T_{l+1}], \mathbb{R})$.

With the substitution $s = F_l(t)$ we have

$$\left| \langle \sqrt{\frac{L}{T-t_0}} h_p^{(q)} \circ F_l, u_k^* \rangle_l \right|^2$$
$$= \left(\frac{L}{T-t_0} \right) \left| \int_{T_l}^{T_{l+1}} h_p^{(q)}(F_l(t)) u_k^*(t)\, dt \right|^2$$
$$= \left(\frac{T-t_0}{L} \right) \left| \int_{\frac{q-1}{2^p}}^{\frac{q-1/2}{2^p}} 2^{p/2} u_k^* \left(\frac{T-t_0}{L} s + T_l \right) ds \right.$$
$$\left. - \int_{\frac{q-1/2}{2^p}}^{\frac{q}{2^p}} 2^{p/2} u_k^* \left(\frac{T-t_0}{L} s + T_l \right) ds \right|^2$$
$$= 2^p \left(\frac{T-t_0}{L} \right) \left| \int_{\frac{q-1/2}{2^p}}^{\frac{q}{2^p}} \left(u_k^* \left(\frac{T-t_0}{L} s + T_l \right) \right. \right.$$
$$\left. \left. - u_k^* \left(\frac{T-t_0}{L} \left(s + \frac{1}{2^{p+1}} \right) + T_l \right) \right) ds \right|^2$$
$$\le 2^p \left(\frac{T-t_0}{L} \right) \|u_k^*\|_{Lip}^2 \left| \int_{\frac{q-1/2}{2^p}}^{\frac{q}{2^p}} \frac{T-t_0}{L} \frac{1}{2^{p+1}}\, ds \right|^2$$
$$= \frac{1}{2^{3p+4}} \left(\frac{T-t_0}{L} \right)^3 \|u_k^*\|_{Lip}^2.$$

Therefore

$$\sum_{j=L}^{\infty} \sum_{k=1}^{m} |\langle f_j, u_k^* \rangle|^2$$
$$\le \left(\frac{T-t_0}{L} \right)^3 \sum_{k=1}^{m} \sum_{l=0}^{L-1} \sum_{p=0}^{\infty} \frac{1}{2^{3p+4}} \sum_{q=1}^{2^p} \|u_k^*\|_{Lip}^2$$
$$\le \left(\frac{T-t_0}{L} \right)^3 \sum_{l=0}^{L-1} \sum_{p=0}^{\infty} \frac{1}{2^{3p+4}} \sum_{q=1}^{2^p} \|u_{x_0}^*\|_{Lip}^2$$
$$= \frac{(T-t_0)^3}{L^2} \|u_{x_0}^*\|_{Lip}^2 \cdot \frac{1}{12},$$

thus

$$0 \le J_{x_0} \left(\sum_{j=0}^{L-1} f_j(t) \bar{u}_j^* \right) - J_{x_0}(u_{x_0}^*)$$
$$\le \frac{1}{L} (T-t_0)^{3/2} \cdot \frac{\|u_{x_0}^*\|_{Lip}}{2\sqrt{3}} \left[\|x_0\| \eta_1(\lambda) + 2\|u_{x_0}^*\|_{L_2} \eta_2(\lambda) \right].$$

□

Appendix F. PROOF OF THEOREM 8

Proof. We have, by Taylor's theorem (see Kantorovich and Akilov (1964)), $J_{x_0}(u_{x_0}^* + h) = J_{x_0}(u_{x_0}^*) + DJ_{x_0}(u^*)(h) + \frac{1}{2} D^2 J_{x_0}(u_{x_0}^*)(h,h)$, where DJ_{x_0} denotes the Frechet-derivative of J_{x_0}. Letting $h = u - u_{x_0}^*$ we obtain $DJ_{x_0}(u_{x_0}^*)(u - u_{x_0}^*) \ge 0$, otherwise $u - u_{x_0}^*$ would be a descent direction in $u_{x_0}^*$.

Moreover, it follows from (3) that

$$\frac{1}{2} D^2 J_{x_0}(u_{x_0}^*)(h,h) = \langle h, Rh \rangle - \langle h, \mathcal{F}h \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the usual scalar product in L_2 and $\mathcal{F}$ is the negative Fredholm operator:

$$\mathcal{F} : L_2([t_0, T], \mathbb{R}^m) \to L_2([t_0, T], \mathbb{R}^m)$$
$$\mathcal{F}z(t) = \int_{t_0}^{T} B^T k(t, \beta) B z(\beta)\, d\beta,$$

where $k(t, \beta)$ is as defined in (5). Since $\langle h, \mathcal{F}h \rangle \le 0$ we obtain $0 \le \frac{1}{2} D^2 J_{x_0}(u_{x_0}^*)(h,h) \le \varepsilon$ and also $0 \le \langle h, Rh \rangle \le \varepsilon$. We have $h(t)^T R h(t) \ge \|h(t)\|^2 \lambda_{\min}(R)$ for all $t \in [t_0, T]$, and hence

$$\varepsilon \ge \int_{t_0}^{T} \|h(t)\|^2 \lambda_{\min}(R)\, dt = \lambda_{\min}(R) \|h\|_{L_2}^2.$$

□