# Discretisation of sparse linear systems: An optimisation approach☆

M. Souza[b,*,1], J.C. Geromel[b], P. Colaneri[c], R.N. Shorten[a,d,1]

[a] University College Dublin, Dublin, Ireland

[b] School of Electrical and Computer Engineering, Univeristy of Campinas, Campinas, SP, Brazil

[c] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, and IEIIT-CNR, Milan, Italy

[d] IBM Research Ireland, Dublin, Ireland

Received 7 July 2014 Received in revised form 11 January 2015
Accepted 28 March 2015 Available online 11 May 2015

## 1. Introduction

Discretisation of continuous-time linear systems is a well established procedure, due to its key role in digital control engineering [1] and sampled-data systems [2]. Nevertheless, the requirement for novel discretisation methods is still emerging in several areas. Examples of such areas include networked control systems [3] and large scale collaborative optimisation problems such as those found in intelligent transportation systems (ITS) applications [4]. The basic objective in these new application areas is that one seeks preserve a certain property of interest. In this paper, we will consider the problem of realising discretisation algorithms that preserve sparsity constraints.

Large-scale dynamical systems usually present structural characteristics, which are fundamental to describe their behaviour [5]. Indeed, these systems usually derive from the dynamical interaction of several interconnected subsystems, which can model industrial settings [6], automated highway systems (AHS) [7], structural dynamics [8] and network flow problems [9]. Thus, these structures arise not only due to physical properties of the system being modelled, but also due to communication and costs limitations. For instance, an AHS may only allow communication between neighbouring vehicles, which builds up a sparsity pattern in its continuous-time state dynamics.

The sparsity patterns presented by large scale systems are usually obtained for their continuous-time formulation. However, the discrete-time versions of these models are the ones that will be either implemented or simulated and, as it will be further discussed in the sequel, the classical discretisation methods usually destroy this sparsity pattern. To avoid this, discretisation methods based on Euler's forward approximation to the exponential can be adopted [10,11]. Unfortunately, these approximations are usually good only for small values of the sampling period.

This paper provides novel discretisation techniques for sparse linear systems. We break free from the classical approach of approximating the matrix exponential and recast the problem in the setting of convex optimisation, which can be solved efficiently with the existing methods. Error bounds are provided for special classes of sparse matrices that arise in several practical applications.

The notation is standard. Capital letters denote matrices and small letters denote vectors and scalars. For matrices and vectors, $(')$ denotes transpose and, for a block-structured symmetric matrix, $(\star)$ denotes each of its symmetric blocks. The sets of real, nonnegative real, positive real and natural numbers are indicated as $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_+^*$ and $\mathbb{N}$. For symmetric matrices, $X \succ 0$ denotes that $X$ is positive definite. For square matrices, $\mathbf{tr}(\cdot)$ denotes the trace function and

☆ This work was in part supported by Science Foundation Ireland grant 11/PI/1177; Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP/Brazil) (2012/02781-7); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/Brazil) (303887/2014-1).

* Corresponding author.
*E-mail address:* souza@dsce.fee.unicamp.br (M. Souza).

[1] M. Souza and R.N. Shorten were with the Hamilton Institute, National University of Ireland Maynooth, when this work was completed.

$\sigma_{\max}(\cdot)$ represent its maximum singular value. Block diagonal matrices are defined by its blocks using the notation **diag**$(\cdot)$, as usual. For a real matrix $A$, its *spectral* and its *Frobenius norms* will be denoted by $\|A\|_2 = \sigma_{\max}(A)$ and $\|A\|_F = \sqrt{\mathbf{tr}(A'A)}$. Finally, for a real function $f$ of one variable, $f^{(n)}$ denotes its $n$-th order derivative.

## 2. Discretisation of sparse linear systems

### 2.1. Problem statement

In this paper, we consider a continuous-time, linear, time-invariant (LTI) autonomous system given by

$$\dot{x}(t) = Ax(t), \qquad x(0) = x_0, \tag{1}$$

in which $x : \mathbb{R}_+ \to \mathbb{R}^n$ is its state. In the classical discretisation problem, the discrete-time realisation

$$x[k+1] = Mx[k], \qquad x[0] = x_0, \tag{2}$$

must be determined to ensure that $x(kh) \approx x[k]$ for all $k \in \mathbb{N}$, where $h \in \mathbb{R}_+^*$ is the *discretisation step* or *sampling period*. It is a well known fact [12] that, whenever $M = e^{hA} = \sum_{k=0}^{\infty}(hA)^k/k!$, the discrete-time LTI system (2) is such that $x(kh) = x[k]$ for all $k \in \mathbb{N}$. Hence, whenever this exact approach can be adopted, the discretisation problem is readily solved from the computation of the matrix exponential, [13]. However, in some applications, one seeks to determine $M$ that approximates $e^{hA}$ and preserves some specific properties, such as sparsity.

In what follows, we assume that $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ is a sparse matrix, whose specific sparsity pattern is defined by the set $\mathcal{S} \subset \mathbb{R}^{n\times n}$. Formally, one can consider the set $\mathcal{I}_{\mathcal{S}} \subset \{1,\dots,n\}^2$ composed of pairs $(i,j)$ such that $a_{ij}$ is allowed to be nonzero and, therefore, define $\mathcal{S}$ as the set that contains all matrices $S = (s_{ij}) \in \mathbb{R}^{n\times n}$ such that $s_{ij} = 0$ whenever $(i,j) \notin \mathcal{I}_{\mathcal{S}}$. Due to its definition, note that $\mathcal{S}$ is a subspace of $\mathbb{R}^{n\times n}$. However, it is of interest to observe that $A \in \mathcal{S}$ does not ensure that $e^{hA} \in \mathcal{S}$ for some $h > 0$. In fact, the discretisation procedure $A \mapsto e^{hA}$ generally destroys structural properties of the original continuous-time system. This phenomenon, which ensures $x(kh) = x[k], \forall k \in \mathbb{N}$, creates direct dependencies between state variables that do not exist in the original continuous-time dynamics. Hence, considering another subspace $\mathcal{R} \subset \mathbb{R}^{n\times n}$ that defines a sparsity pattern, our main goal is to determine $M \in \mathcal{R}$ such that $M \approx e^{hA}$ for some $h > 0$ given. It is often desirable that $\mathcal{R} = \mathcal{S}$, but this may be relaxed in some situations, where we will allow $\mathcal{R} \supset \mathcal{S}$. For example, in ITS applications, local inter-vehicle communication may be possible. Moreover, $\mathcal{R}$ also presents a set $\mathcal{I}_{\mathcal{R}} \subset \{1,\dots,n\}^2$ composed of the nonzero positions allowed by its structure and, since $\mathcal{R}$ may relax some constraints imposed by $\mathcal{S}$, it follows that $\mathcal{I}_{\mathcal{R}} \supset \mathcal{I}_{\mathcal{S}}$. This property can be exploited not only to improve the quality of the approximation to the matrix exponential but also to make the optimisation feasible in some situations.

### 2.2. Mathematical preliminaries

The following auxiliary results and definitions are extensively used throughout. The matrix exponential can be computed via numerical methods based on *Padé approximants* [13,14]. Two particular cases of approximants to the matrix exponential are Tustin's formula

$$e^{sA} \approx T(sA) \triangleq \left(I - \frac{s}{2}A\right)^{-1}\left(I + \frac{s}{2}A\right) \tag{3}$$

and Taylor's polynomial of order $\lambda$, centred at the origin,

$$e^{sA} \approx R_\lambda(sA) \triangleq \sum_{k=0}^{\lambda}\frac{s^k}{k!}A^k. \tag{4}$$

Padé approximants are widely adopted in discretisation methods [11,15]. It is also worth pointing out that Tustin's approximant plays a key role in control theory [12].

The following theorem [13] provides an error bound for the approximation of a matrix function.

**Theorem 1.** *If $f(\cdot)$ has the Taylor series representation $f(z) = \sum_{k=0}^{\infty}\alpha_k z^k$ in an open disk containing the eigenvalues of $A \in \mathbb{C}^{n\times n}$, then*

$$\left\| f(A) - \sum_{k=0}^{\lambda}\alpha_k A^k \right\|_2 \leq \frac{n\|A\|_2^{\lambda+1}}{(\lambda+1)!}\max_{0\leq s\leq 1}\left\|f^{(\lambda+1)}(sA)\right\|_2. \tag{5}$$

We are particularly interested in the case $f(\cdot) \equiv \exp(\cdot)$, which implies

$$\left\| e^A - \sum_{k=0}^{\lambda}\frac{A^k}{k!} \right\|_2 \leq \frac{n\|A\|_2^{\lambda+1}}{(\lambda+1)!}e^{\|A\|_2}. \tag{6}$$

Additionally, it is also possible to obtain bounds for the Frobenius norm and for any Padé approximant to the exponential; see [13,14].

### 2.3. Discretisation as an optimisation problem

Now we focus on the main problem stated before, which can be analysed, in a simple manner, as a projection problem. Indeed, given a continuous-time system with realisation (1) and a step size $h \in \mathbb{R}_+^*$, we wish to determine $M^\star \in \mathcal{R} \subset \mathbb{R}^{n\times n}$ such that $M^\star$ is the "closest" element of $\mathcal{R}$ to $e^{hA}$, with respect to the metric $\delta$. Thus, its general formulation is

$$M^\star = \arg\inf_{M\in\mathcal{R}} \delta\left(M, e^{hA}\right), \tag{7}$$

in which $\delta$ provides the notion of distance between the approximation $M$ and the exact discrete-time matrix $e^{hA}$, for any $h \in \mathbb{R}_+^*$ given. Thus, from the computational viewpoint, it represents the *error* yielded by the approximation. Note that, whenever $\delta$ is induced by a matrix norm, the optimisation problem (7) is convex; see [16].

In this paper, we are particularly interested in two widely adopted norms in approximation problems (see [17]): the spectral norm and the Frobenius norm. For the first case, it is possible to show [18,19] that, for a given sampling period $h \in \mathbb{R}_+^*$, there exists $\sigma \in \mathbb{R}_+^*$ such that the error bound $\|M - e^{hA}\|_2 < \sigma$ holds if, and only if, the linear matrix inequality (LMI)

$$\begin{pmatrix} \sigma^2 I & \star \\ M - e^{hA} & I \end{pmatrix} \succ 0 \tag{8}$$

is satisfied. Accordingly, for the Frobenius norm case, the error bound $\|M - e^{hA}\|_F < \sigma$ holds if, and only if, there exists $W \succ 0$ such that the LMIs

$$\mathbf{tr}(W) < \sigma^2, \qquad \begin{pmatrix} W & \star \\ M - e^{hA} & I \end{pmatrix} \succ 0 \tag{9}$$

hold. Hence, whenever $\delta$ is induced by $\|\cdot\|_2$, the best approximation in $\mathcal{R}$ to the matrix exponential can be obtained solving the convex optimisation problem

$$(M^\star, \sigma^\star) = \arg\inf_{M\in\mathcal{R},\sigma}\{\sigma : (8)\}. \tag{10}$$

Similarly, for the Frobenius norm, the best approximation in $\mathcal{R}$ to the matrix exponential can be obtained solving

$$(M^\star, \sigma^\star) = \arg\inf_{M\in\mathcal{R},\sigma}\{\sigma : (9)\}. \tag{11}$$

In both cases, $\sigma^\star = \delta(M^\star, e^{hA})$ is the optimal value for the error yielded by the approximation. Additionally, both optimisation problems are convex, as expected.

**Remark 1.** The element-wise characteristics of the Frobenius norm allows us to obtain analytically the optimal solution of the problem (11). Indeed, the solution $M^\star = (m_{ij}^\star) \in \mathcal{R}$ to (11) is such that

$$m_{ij}^\star = \begin{cases} \phi_{ij}, & \text{if } (i,j) \in \mathcal{I}_{\mathcal{R}} \\ 0, & \text{if } (i,j) \notin \mathcal{I}_{\mathcal{R}} \end{cases} \tag{12}$$

in which $e^{hA} = (\phi_{ij}) \in \mathbb{R}^{n\times n}$. Note also that the optimal error is given by $\left\|M^\star - e^{hA}\right\|_F^2 = \sum_{(i,j)\notin\mathcal{I}_{\mathcal{R}}} \phi_{ij}^2$.

Thus, the optimal approximation with respect to the Frobenius induced metric is very simple to obtain; one just has got to neglect the elements that are not allowed in the feasibility pattern. This simple technique may not provide good approximations in some cases, since it is completely element-wise and, hence, we will only consider the 2-norm formulation in the sequel.

An important theoretical advantage in adopting the spectral norm is based on the following result, which ensures *consistency* and *order of accuracy* for the approximation yielded by the solution to (10) (see [20] for details).

**Theorem 2.** *Consider the continuous-time dynamical system (1) and the optimisation problem (10). If the approximant $R_\lambda(hA)$, $\lambda \geq 1$, to the matrix exponential $e^{hA}$ is feasible to (10), then the discrete-time iteration $x[k+1] = M^\star x[k]$, with $x[0] = x_0$, is consistent with the differential equation (1) and it has order of accuracy of, at least, $\lambda$.*

**Proof.** The proof follows from the concept of truncation error for numerical methods for ODEs. At the time instant $t_k = kh \in [0,T]$, $k \in \mathbb{N}$, for some $T > 0$ fixed, the truncation error $\tau_k(h)$ is defined as $\tau_k(h) = (1/h)(x(t_{k+1}) - x[k+1])$, that is, it is related to the error caused by a single iteration, assuming that the true solution at the point $t_k$ is known. Hence, for this case, we have that

$$\begin{aligned} \|\tau_k(h)\|_2 &= (1/h)\|e^{hA}x(t_k) - M^\star x(t_k)\|_2 \\ &\leq (1/h)\|e^{hA} - M^\star\|_2 \|x(t_k)\|_2 \\ &\leq (1/h)\|e^{hA} - R_\lambda(hA)\|_2 \|x(t_k)\|_2 \\ &\leq nh^\lambda \|A\|_2^{\lambda+1} e^{\|hA\|_2} \|x(t_k)\|_2/(\lambda+1)!, \end{aligned} \tag{13}$$

where we have used the results of Theorem 1 and the fact that $R_\lambda(hA)$ is feasible to (10). Since the solution of (1) is bounded for any $t_k \in [0,T]$, $k \in \mathbb{N}$, there exists a constant $K$ such that $\|\tau_k(h)\|_2 \leq Kh^\lambda$ for $h \in \mathbb{R}_+^*$ sufficiently small and, hence, the method is consistent and has order of accuracy of at least $\lambda$. The proof is complete. □

Some final remarks must be made in this section. First, the results stated in Theorem 2 provide important theoretical consequences for the discretisation method studied in this paper. Since in almost all situations the matrix $R_1(hA) = I + hA$, $h \in \mathbb{R}_+^*$, is feasible to $\mathcal{R}$, we guarantee that the discrete-time iteration (2), with $M = M^\star$ obtained from (7), is consistent and has order of accuracy of at least 1. As it will be discussed in the sequel, the feasibility of $R_\lambda$ for $\lambda \geq 2$ is ensured for special cases of $\mathcal{S}$ and $\mathcal{R}$. Moreover, it is important to note that the main difficulty in this problem is to adequately choose $\mathcal{R}$ in order to provide an optimal solution $M^\star$ such that the optimal error $\sigma^\star$ is sufficiently small. This problem is not fully addressed here. For some special classes of matrices, to be analysed in the sequel, some results provide a basis for this choice. Furthermore, it is important to remember that, in many applications, $\mathcal{R}$ is specified by, for instance, inter-agent communication patterns.

**Remark 2.** In the literature, two other methods can also be adopted in the discretisation problem with structural constraints. These are *Euler's approximation*, which uses $M \approx I + hA$, and the recently introduced *Mixed Euler-ZOH* (mE-ZOH) method, discussed in [10,11]. The latter is based on the approximation $M \approx I + D(h)A$, where

$$D(h) = \mathbf{diag}\left(\int_0^h e^{a_{11}t}dt, \ldots, \int_0^h e^{a_{nn}t}dt\right). \tag{14}$$

It is clear that both approximations are less expensive from a computational viewpoint when compared to our approach, since both do not require the explicit evaluation of $e^{hA}$. However, the approximations provided by these methods are usually most effective for very small values of $h \in \mathbb{R}_+^*$. Based on these methods, we can impose the structure $M = I + DA$ in (10) or (11) and obtain optimisation problems with the diagonal matrix $D$ as a decision variable.

## 3. Special sparsity patterns

In this section, we discuss some special sparsity patterns that arise in many practical applications. These patterns are then exploited to provide error bounds on the approximations yielded by (10). Moreover, we briefly analyse the relaxation of the structure of $\mathcal{R}$ to provide smaller values for the error.

### 3.1. Band matrix systems

First, we study band matrix systems, which present a peculiar structure that is intrinsic to several applications. For instance, tridiagonal systems are often used to model interaction between species [21], binary distillation systems [22] and some queueing systems [23]. Systems of this class present the following realisation

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2), \\ \dot{x}_i &= f_i(x_{i-1}, x_i, x_{i+1}), \quad i = 2, \ldots, n-1 \\ \dot{x}_n &= f_n(x_{n-1}, x_n) \end{aligned} \tag{15}$$

evolving from the initial condition $x(0) = x_0 \in \mathbb{R}^n$. Moreover, this structure implies the linearised model obtained near an equilibrium point has a tridiagonal dynamic matrix, which is a particular case of a band matrix.

Following [13], we say $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ is a $(b_u, b_\ell)$-band matrix if all the nonzero elements of $A$ have to satisfy the condition $-b_u \leq i - j \leq b_\ell$. The constants $b_u$ and $b_\ell$ are usually called *upper bandwidth* and *lower bandwidth*, respectively. Diagonal, tridiagonal, pentadiagonal, Hessenberg and triangular matrices can be seen as special cases of band matrices and, thus, this class of matrices plays an important role in a wide range of applications.

Band matrices present several important properties, due to their structure. For instance, if $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ is a $(b_u, b_\ell)$-band matrix, it is easy to show (see [18]) that $A^k$ is a $(k\cdot b_u, k\cdot b_\ell)$-band matrix, for every $k \in \mathbb{N}$. This fact implies the following properties hold, for $b_u, b_\ell > 0$ and a given scalar $h \in \mathbb{R}_+^*$:

(i) The exponential $e^{hA}$ is, in general, a full matrix.
(ii) Taylor's polynomial approximant of order $\lambda$, $R_\lambda(hA)$, is a $(\lambda b_u, \lambda b_\ell)$-band matrix.

Some comments concerning the previous results must be made at this point. First, the "fill-up" effect caused by the exponential is emphasised in (i), but the same occurs to general Padé approximants to the exponential and, in particular, to Tustin's formula, whence we conclude these methods do not consider adequately the sparsity pattern presented by such systems. Finally, it is remarkable that the truncated Taylor series preserves a band sparsity pattern, but with a larger bandwidth. Moreover, it is clear that, as $\lambda$ increases, the quality of the approximation improves, but the yielded matrix will have a large number of nonzero entries.

Now we focus again on the discretisation problem. The results developed in this section altogether allow us to state the following theorem, which provides an important error bound to the approximation yielded by (10).

**Theorem 3.** *Assume $A \in \mathcal{S} \subset \mathbb{R}^{n\times n}$, where $\mathcal{S}$ is the set composed of all $(b_u, b_\ell)$-band matrices and let $h \in \mathbb{R}_+^*$ be given. If $\mathcal{R} \subset \mathbb{R}^{n\times n}$ is the set composed of all $(\lambda b_u, \lambda b_\ell)$-band matrices, $\lambda \geq 1$, then the solution $(M^\star, \sigma^\star)$ to the convex optimisation problem (10) is such that*

$$\sigma^\star \leq \frac{n(h\|A\|_2)^{\lambda+1}}{(\lambda+1)!} e^{h\|A\|_2}. \tag{16}$$

*Furthermore, the iteration (2), with $M = M^\star$, is consistent with (1) and has order of accuracy of at least $\lambda$.*

**Proof.** First, note that the definition of $\mathcal{R}$ implies that Taylor's approximant $R_\lambda(hA)$ satisfies the constraints of the optimisation problem (10) for some $\sigma \in \mathbb{R}_+^*$. Therefore, it is clear that $\sigma^\star = \|M^\star - e^{hA}\|_2 \leq \|R_\lambda(hA) - e^{hA}\|_2$. Finally, considering the error bound (6), we conclude that (16) holds. The rest of the statement follows from Theorem 2. The proof is complete. □

It is worth noting that the special structure presented by band matrices allows us to obtain error bounds for our approximation. As it has been shown in the proof of Theorem 3, this result is completely based on the feasibility of $R_\lambda(hA)$ with respect to the constraints of the optimisation problem (10). Thus, the approximations yielded by the optimal solution of (10) are always at least as good as the ones provided by Taylor's polynomial approximant of order $\lambda$.

### 3.2. Arrowhead matrix systems

Now we briefly discuss another special class of sparse matrices, known as *arrowhead matrices*, which may arise in network flow control and optimisation problems [9]. Indeed, in this reference, the authors consider a continuous-time primal–dual update for optimisation variables associated with a network flow control problem. The decentralised structure of the algorithm must be considered by the discretisation method applied to this problem. In the special case of single link networks, the dynamic matrix of the linearised model has only nonzero elements on its main diagonal and on its last row and column, which is the structure of an arrowhead matrix.

Formally, an arrowhead matrix $A \in \mathbb{R}^{n\times n}$ can be written, without any loss of generality, in the form

$$A = D + uv' + wu', \tag{17}$$

where $D \in \mathbb{R}^{n\times n}$ is diagonal, $v, w \in \mathbb{R}^n$ are given vectors and $u = [0 \quad \cdots \quad 0 \quad 1]' \in \mathbb{R}^n$. As it happens with band matrices, the matrix exponential of an arrowhead matrix is, in general, full. Indeed, we have $A^2 = AD + DA - D^2 + (uv' + wu')^2$ and, due to the last term, $A^2$ is filled-up and sparsity is lost for $e^A$. However, Euler's approximation

$$e^{hA} \approx I + hA = (I + hD) + h(uv' + wu') \tag{18}$$

also has the arrowhead structure and, thus, is feasible for the optimisation problem (10), whenever $\mathcal{S}$ and $\mathcal{R}$ are sets that contain this class of matrices. Hence, whenever $M = M^\star$ given by (10), we can ensure (2) is consistent with (1) with order of accuracy of at least 1. Thus, in this case, considering that $(\sigma^\star, M^\star)$ is the solution to (10), we have

$$\sigma^\star = \|M^\star - e^{hA}\|_2 \leq (1/2)nh^2\|A\|_2^2 e^{h\|A\|_2} \tag{19}$$

where we used the results of Theorem 1. As before, the structure constraint imposed by $\mathcal{R}$ can be relaxed.

## 4. Stability and positivity preservation

In almost all situations, sparsity preservation is just one constraint of many. For example, one is often interested in the quality of the solution and other qualitative properties, such as stability and positivity. In this section we show how these properties can be incorporated in the discretisation procedure; specifically, we consider preservation of asymptotical stability and positivity in case of positive systems. Both properties have already been addressed for special classes of systems in [10,11].

### 4.1. Preserving stability

It is widely known the mapping $A \mapsto e^{hA}$ preserves asymptotic stability for any $h \in \mathbb{R}_+^*$, as a consequence of the dynamic behaviour of the considered system. Whenever approximations are adopted, one has to be careful with the choice of the step $h \in \mathbb{R}_+^*$, since only some methods can preserve stability for all $h \in \mathbb{R}_+^*$. We now indicate how to preserve Lyapunov stability in our optimisation setting. We begin with the following theorem.

**Theorem 4.** *Let the Hurwitz stable[3] matrix $A \in \mathcal{S} \subset \mathbb{R}^{n\times n}$ and the sampling period $h \in \mathbb{R}_+^*$ be given. Let $M$ be a matrix in $\mathcal{R} \subset \mathbb{R}^{n\times n}$ and $\sigma$ be a scalar in $\mathbb{R}_+^*$. The following statements are equivalent:*

(i) *$M$ is Schur stable[4] and the error bound $\|M - e^{hA}\|_2 < \sigma$ holds.*
(ii) *There exist matrices $S = S' \in \mathbb{R}^{n\times n}$ and $G \in \mathbb{R}^{n\times n}$ such that*

$$\begin{pmatrix} \sigma^2 I & \star \\ M - e^{hA} & I \end{pmatrix} \succ 0, \qquad \begin{pmatrix} G + G' - G'SG & \star \\ M & S \end{pmatrix} \succ 0. \tag{20}$$

**Proof.** First, note that $\|M - e^{hA}\|_2 < \sigma$ is clearly equivalent to the first inequality of (20). Thus, we first observe that the Schur stability of $M$ implies the second inequality of (20) holds for $G = S^{-1} \succ 0$. Conversely, we also observe that Schur complement applied to the second inequality of (20) yields $S^{-1} \succeq G + G' - G'SG \succ M'S^{-1}M$, which implies $M$ is Schur stable, completing the proof. □

Taking into account the statement of Theorem 4, the best approximation in terms of the metric induced by the spectral norm can be determined by solving

$$(M^\star, \sigma^\star, S^\star, G^\star) = \arg \inf_{M\in\mathcal{R},\sigma,S>0,G} \{\sigma : (20)\}, \tag{21}$$

which, albeit non-convex, possesses a remarkable property; namely, whenever $G \in \mathbb{R}^{n\times n}$ is fixed, the constraints become LMIs with respect to the remaining variables and, consequently, the problem becomes convex. This property can be exploited to provide a sequential convex optimisation method that calculates a suboptimal solution as a result of a convergent sequence of intermediate solutions with decreasing costs

$$(M_{\ell+1}, \sigma_{\ell+1}, S_{\ell+1}) = \arg \inf_{M\in\mathcal{R},\sigma,S>0,G_\ell=S_\ell^{-1}} \{\sigma : (20)\}, \tag{22}$$

where $G$ is fixed as the previous value for $S^{-1}$.

**Proposition 5.** *Assume (20) is feasible for some $G_0 = S_0^{-1} \succ 0$ fixed. The iterative method defined by (22) generates a convergent sequence such that $\sigma_{\ell+1} \leq \sigma_\ell$, $\forall \ell \in \mathbb{N}$.*

[3] A matrix whose eigenvalues are in $\{s \in \mathbb{C} : \mathrm{Re}(s) < 0\}$.
[4] A matrix whose eigenvalues are in $\{z \in \mathbb{C} : |z| < 1\}$.

For the proof, the reader can refer to the similar result stated in [24]. Note that the most challenging aspect of this algorithm is the choice of $G_0 = S_0^{-1}$. Indeed, suppose there is a feasible solution $M$ that is sufficiently close to $e^{hA}$. We can adopt $S_0 = X^{-1}$, with $X \succ 0$ being a Lyapunov matrix for $e^{hA}$, that is, $(e^{hA})'X(e^{hA}) - X = -Q \prec 0$, for a given $Q \succ 0$. This method provides good results; in most cases, no conservatism is added, since it finds an optimal solution. Furthermore, the number of iterations, empirically speaking, is usually found to be small.

### 4.2. Preserving positivity

Positivity of LTI systems is a very important property; see [23,25]. Formally, the continuous-time LTI system (1) is said to be positive if, for any given initial state $x(0) = x_0 \geq 0$, we have $x(t) \geq 0,\ \forall t \in \mathbb{R}_+$. Equivalently, the discrete-time, LTI system (2) is positive if, for any initial state $x[0] = x_0 \geq 0$, we have $x[k] \geq 0,\ \forall k \in \mathbb{N}$. Thus, it is unacceptable to provide a discrete-time approximation to a continuous-positive system which does not preserve this property, due to the physical meaning the state variables may have in real world applications. Hence, we now discuss on the preservation of positivity in our approach.

It is a well-known fact that (1) is positive if, and only if, $A$ is a Metzler matrix, that is, all of its off-diagonal elements are nonnegative. Moreover, (2) is positive if, and only if, $M$ is a nonnegative matrix, that is, all of its elements are nonnegative. Let us denote by $\mathcal{M}$ the set of all Metzler matrices and $\mathcal{M}_d$ the set of all the nonnegative matrices. It is clear that, if $A \in \mathcal{M}$, then $M = e^{hA} \in \mathcal{M}_d,\ \forall h \in \mathbb{R}_+^*$. Therefore, we should seek $M$ that approximates the exponential $e^{hA}$, satisfies some structure constraint imposed by $\mathcal{R}$ and is a nonnegative matrix. Additionally, since $\mathcal{M}_d$ is a convex cone in $\mathbb{R}^{n\times n}$, the optimal approximation considering the metric induced by the 2-norm can be obtained from

$$(M^\star, \sigma^\star) = \arg \inf_{M\in\mathcal{R}\cap\mathcal{M}_d,\sigma} \{\sigma : (8)\}, \tag{23}$$

which is convex due to the convexity of the new feasible set for $M$ is $\mathcal{R} \cap \mathcal{M}_d$. Clearly this formulation does not add any conservatism to the previous conditions. Note that (23) can be combined with the sequential procedure developed previously to ensure the Schur stability of $M^\star$.

**Remark 3.** Based on the mE-ZOH structure, if we adopt $M = I + DA$ and ensure $M \in \mathcal{M}_d$, then $M$ is Schur stable for every diagonal matrix $D \succ 0$ whenever $A \in \mathcal{M}$ is Hurwitz (see [11]). Therefore, in this case, the optimisation problem (23) always provides a stable, nonnegative approximation.

## 5. Related problems

### 5.1. Lyapunov function preservation

One of the many properties of Tustin's approximation is the preservation of Lyapunov functions: for any Hurwitz stable matrix $A$ and any $h \in \mathbb{R}_+^*$, if $P \succ 0$ is such that $A'P + PA \prec 0$ then $T(hA)'PT(hA) - P \prec 0$ holds (see [26]). Hence, if one needs to preserve a Lyapunov matrix $P \succ 0$ in our discretisation setting, the problem to be solved can be formulated as

$$\inf_{M\in\mathcal{R},\ P\succ 0} \left\{\delta\left(M, e^{hA}\right) : A'P + PA \prec 0,\ M'PM - P \prec 0\right\}, \tag{24}$$

which is convex whenever $\delta$ is induced by a norm and $P \succ 0$ such that $A'P + PA \prec 0$ is fixed. Nevertheless, if $P \succ 0$ is considered as a variable, this problem can be solved by the sequential method presented previously.

This problem can be generalised to the context of quadratic stability of switched systems; see [27,28]. Given a set of $N$ Hurwitz matrices $\mathcal{A}_c$, a matrix $P \succ 0$ is a common Lyapunov matrix (CLM) for $\mathcal{A}_c$ if $A'P + PA \prec 0$ for all $A \in \mathcal{A}_c$. Similarly, given a set of $N$ Schur matrices $\mathcal{A}_d$, a matrix $P \succ 0$ is a CLM for $\mathcal{A}_d$ if $M'PM - P \prec 0$ for all $M \in \mathcal{A}_d$. These concepts are essential to define *quadratic stability* (QS) of switched systems (see [27]). As a clear consequence from the LTI case, whenever we consider a set $\mathcal{A}_c$ of Hurwitz matrices with an associated CLM $P \succ 0$, the first order diagonal Padé approximant yields a set $\mathcal{A}_d$ of Schur matrices with the same CLM.

In our optimisation approach, if one seeks to find the discrete-time set of matrices $\mathcal{A}_d$ such that presents the same CLM $P \succ 0$ as $\mathcal{A}_c$, the problem to be solved is

$$\inf_{M_i\in\mathcal{R},P\succ 0} \left\{\sum_{i=1}^{N} \delta\left(M_i, e^{hA_i}\right) : A_i'P + PA_i \prec 0,\ M_i'PM_i - P \prec 0\right\}, \tag{25}$$

for $i = 1, \ldots, N$, which is only convex if $\delta$ is induced by a norm and the CLM $P \succ 0$ for $\mathcal{A}_c$ is fixed. In the more general case where $P \succ 0$ is a variable, the problem is nonconvex and, thus, an adaptation of the sequential procedure detailed before has to be adopted.

### 5.2. Robust discretisation

Another interesting problem that can be addressed is a "robust" discretisation problem. In some applications, one cannot assume that the discretisation step $h \in \mathbb{R}_+^*$ is constant, due to uneven data rates present in real world scenarios. Therefore, one can be interested in the determination of a "robust" discrete-time matrix $M$ that approximates adequately $e^{hA}$ for all $h \in [h_\star, h^\star]$, in which this interval bounds the uncertainties on the step size. This problem can be formulated as

$$\inf_{M\in\mathcal{R}} \sup_{h\in[h_\star,h^\star]} \delta\left(M, e^{hA}\right), \tag{26}$$

in which the matrix to be found minimises the maximum error with respect to all values of $h \in [h_\star, h^\star]$. Considering $\delta$ is induced by the 2-norm, this problem is equivalent to

$$\inf_{M\in\mathcal{R},\sigma} \left\{\sigma : \left\|M - e^{hA}\right\|_2 < \sigma,\ h \in \left[h_\star, h^\star\right]\right\} \tag{27}$$

which is similar to the basic problem presented in this paper. It is clear that, due to the nonlinear dependence of $e^{hA}$ on the discretisation step $h \in [h_\star, h^\star]$, this problem is difficult to solve as is. However, the continuity of the constraints with respect to $h \in \mathbb{R}_+^*$ allows one to split the interval $[h_\star, h^\star]$ with a large enough number $N$ of evenly spaced points $h_i$, with $h_1 = h_\star$ and $h_N = h^\star$, and impose (10) to each of them simultaneously. Hence, the convex optimisation problem to be solved is

$$(M^\star, \sigma^\star) = \arg \inf_{M\in\mathcal{R},\sigma} \left\{\sigma : \left\|M - e^{h_iA}\right\|_2 < \sigma,\ i = 1, \ldots, N\right\}. \tag{28}$$

Note that the optimal cost provides the bound for the worst-case error obtained by the robust approximation in the considered interval. Moreover, since the problem can be stated in terms of LMIs, it can be readily solved even for large values of $N$ [19].

### 5.3. Stochastic matrices and Markov processes

The final application to be analysed considers the discretisation of Markov processes. Large-scale, sparse, stochastic models arise in

queue theory and in its applications [29,23,25] and it is essential that discretisation methods preserve the special characteristics presented by these systems. A finite dimensional, continuous-time Markov processes can be modelled as a linear autonomous system with realisation (1), in which $A$ is a Metzler matrix and is such that $\pi' A = 0$, where $\pi = [1 \quad \cdots \quad 1]'$. Thus, $A$ is the transpose of the *transition rate matrix* associated with the process. Similarly, in discrete-time, a finite dimensional Markov process (or chain) can be modelled as a linear autonomous system with realisation (2), where $M$ is a nonnegative matrix and satisfies $\pi' M = \pi'$. Therefore, in the discrete-time case, $M$ is the transpose of the *stochastic matrix* associated with the Markov chain.

Discretisation methods can be applied to continuous-time Markov processes in order to obtain their corresponding discrete-time chains. It can be shown that, if $Q$ is a transition rate matrix, then $P = e^{hQ}$ is a stochastic matrix for any $h \in \mathbb{R}_+^*$. Thus, in this setting, the approximation to be obtained is the solution to

$$\inf_{M \in \mathcal{R} \cap \mathcal{M}_d} \left\{ \delta\left(M, e^{hA}\right) : \pi' M = \pi' \right\}, \tag{29}$$

which is clearly convex whenever $\delta$ is induced by a norm. Thus, the approximation yielded by (29), applied to a continuous-time Markov process, always yields a well-posed Markov chain, in the sense that $M$ is always the transpose of a stochastic matrix.

## 6. Numerical examples

**Example 1** (*Queueing System*)**.** In this example, we consider a simple queueing system, described in [23]. Under Markovian hypotheses, a system with 2 servers and a queue with capacity of 3 clients can be modelled as (1) with

$$A = \begin{pmatrix} -\lambda & \mu & & & \\ \lambda & -\lambda-\mu & 2\mu & & \\ & \lambda & -\lambda-2\mu & 2\mu & \\ & & \lambda & -\lambda-2\mu & 2\mu \\ & & & \lambda & -2\mu \end{pmatrix},$$

where $\lambda, \mu \in \mathbb{R}_+^*$ are parameters related to the rate of arrivals and the time needed to perform the service. At any given instant $t \in \mathbb{R}_+$, $x_i(t)$ represents the probability of having $i$ users in the system. Thus, this system is a continuous-time Markovian process and, in order to be discretised, its characteristics have to be taken into account. For instance, for $\lambda = \mu = 1$ and $h = 0.25$ s, we solve (29) with the metric induced by the spectral norm and obtain

$$M^\star = \begin{pmatrix} 0.8221 & 0.1833 & & & & \\ 0.1779 & 0.6622 & 0.3077 & & & \\ & 0.1545 & 0.5663 & 0.2834 & & \\ & & 0.1260 & 0.5498 & 0.2899 & \\ & & & 0.1668 & 0.5494 & 0.3215 \\ & & & & 0.1607 & 0.6785 \end{pmatrix},$$

which yields an error of $\sigma^\star = 0.0895$. In this case, Euler's method fails to ensure positivity for all $h \in \mathbb{R}_+^*$ and its modified version, although preserving distribution properties and always providing a nonnegative approximation, does not ensure $\pi'(I + D(h)A) = \pi'$, where $\pi = [1 \quad \cdots \quad 1]'$, for $h \in \mathbb{R}_+^*$. Thus, the approximation obtained is adequate.

**Example 2** (*Reactor–Separator Process*)**.** This example considers the linearised model for a reactors–separator process. It was described in [30] and analysed from the discretisation viewpoint in [10]. The dynamic matrix is Hurwitz and presents the block structure

$$A = \begin{pmatrix} A_{11} & & A_{13} \\ A_{21} & A_{22} & \\ & A_{32} & A_{33} \end{pmatrix},$$

in which each block is a $4 \times 4$ real matrix; all of the submatrices are given in [10]. Note that such structure is rather usual in distributed control applications. Hence, our main goal here is to obtain a discrete-time approximant that preserves stability and presents the same block pattern. To this end, our set $\mathcal{R}$ only allows nonzero valuers for the entries that correspond to nonzero blocks of $A$. Following [10], we analyse (see Fig. 1) the spectral radius of the approximant yielded by the optimal solution to (10) (dashed line) and compare it with the spectral radius of the exact solution $e^{hA}$ (solid line) and with the spectral radius of the approximant obtained by the mE-ZOH method (dot-dashed line), for $h \in (0, 0.5]$. It is clear that, for small $h \in (0, 0.5]$, both approximations are good but the optimal solution to (10) is more adequate when this is not the case.

**Example 3** (*Network Flow Control*)**.** We consider an example related to the network flow control problem studied in [9]. The resource allocation is done with a fair utility function $U_i(x_i) = \log(x_i)$ in a single link (with cost $\lambda$ and capacity $c = 2$) and four sources network, with flows $x_i$, $i = 1, \ldots, 4$. In this case, the routing matrix is $R = [1 \quad 1 \quad 1 \quad 1]$. For illustration, the feedback gains $K = 0.01\, I$ and $\Gamma = 10$ are adopted. The linearised model for the dynamics of the variables $x_i$ and $\lambda$, around the optimal solution to this problem ($x_i^\star = 0.5$, $i = 1, \ldots, 4$, $\lambda^\star = 2$), is given by

$$\frac{d}{dt}\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} KU''(x^\star) & -KR' \\ \Gamma R & 0 \end{pmatrix}\begin{pmatrix} x \\ \lambda \end{pmatrix},$$

where $U''(x^\star) = \mathbf{diag}(U_i''(x_i^\star))$. The dynamic matrix $A$ of this system has eigenvalues very close to the imaginary axis. We consider $\mathcal{R} = \mathcal{S}$ for this case. For instance, if we take $h = 0.2$ s, the optimal solution $M^\star$ of problem (10) is

$$M^\star = \begin{pmatrix} 0.9881 & & & & -0.0020 \\ & 0.9881 & & & -0.0020 \\ & & 0.9881 & & -0.0020 \\ & & & 0.9881 & -0.0020 \\ 1.9867 & 1.9867 & 1.9867 & 1.9867 & 0.9920 \end{pmatrix},$$

with the associated error $\sigma^\star = \|M^\star - e^{hA}\|_2 = 0.0040$. Furthermore, we analyse the behaviour of some methods applied to this example and plot the eigenvalues loci parameterised by $h \in (0, 1]$. Blue dots represent the eigenvalues of $e^{hA}$, whilst the ones of the approximations are denoted by green or red circles, the first being used when the approximation yielded is Schur stable and the latter on the contrary. These curves are shown in Fig. 2, which is organised as follows. Subplot (a) shows the optimal solution obtained by solving (10), which yields stable solutions for $h \in (0, 0.41]$; (b) illustrates the approximation given by the mE-ZOH method, which provides stable solutions for $h \in (0, 0.09]$ (if we impose $M = I + DA$ in (10) and optimise $D$, the results are only slightly better); (c) considers the optimal solution obtained by solving (11), which is stable for $h \in (0, 0.27]$; (d) shows the stable solution of (21) obtained with the iterative procedure with $Q = I$. This example shows the quality of the approximations obtained.

## 7. Conclusions

In this paper, we have addressed the discretisation problem for sparse linear systems. Several practical dynamical systems, mainly those that involve the interaction of a large number of independent agents, present some sparsity pattern in their continuous-time formulation that is lost whenever classical discretisation procedures are adopted. Hence, the discretised models yielded by such techniques are not viable for simulation or implementation purposes, due to cost and communication constraints. To preserve the sparsity pattern, we formulate a convex optimisation problem

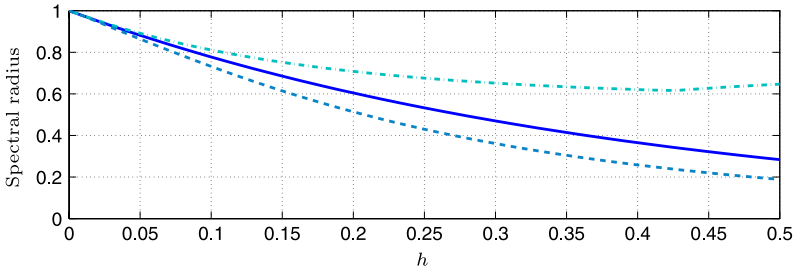
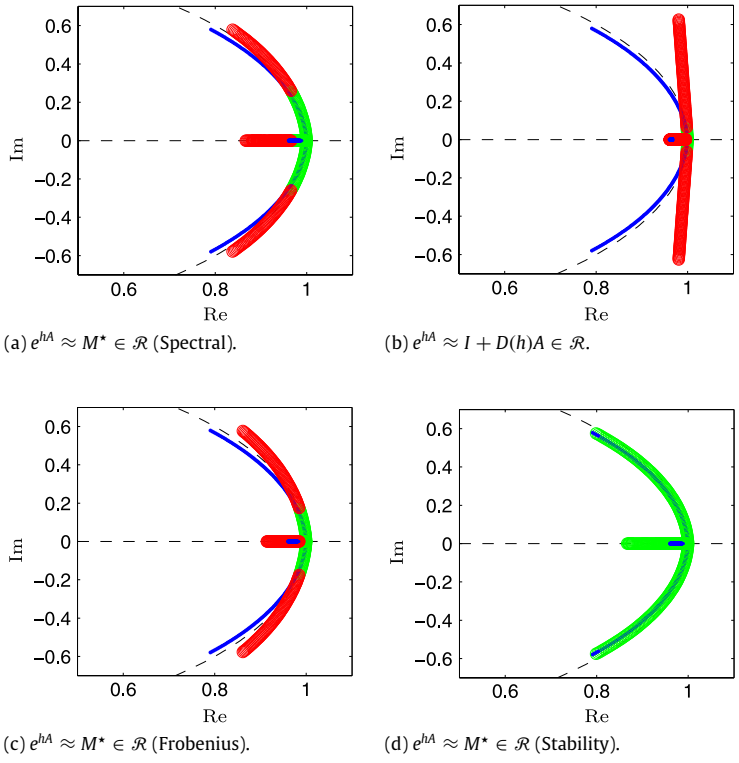



**Fig. 1.** Spectral radius for $h \in (0, 0.5]$.




(a) $e^{hA} \approx M^{\star} \in \mathcal{R}$ (Spectral). (b) $e^{hA} \approx I + D(h)A \in \mathcal{R}$.

(c) $e^{hA} \approx M^{\star} \in \mathcal{R}$ (Frobenius). (d) $e^{hA} \approx M^{\star} \in \mathcal{R}$ (Stability).

**Fig. 2.** Eigenvalues loci plot for $h \in (0, 1]$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

that yields an approximation to the exact discrete-time dynamical matrix that presents a desired structure. Error bounds were provided for band matrix and arrowhead matrix linear systems, which occur in several applications. We have also addressed related problems that arise in different areas of control theory. Academic examples illustrate the developed results. Future studies should focus on extending and adapting the developed theory to consider dynamical systems presenting inputs and outputs.

## References


[1] G.F. Franklin, J.D. Powell, M.L. Workman, Digital Control of Dynamic Systems, third ed., Prentice Hall, Englewood Cliffs, NJ, 1997.

[2] T. Chen, B.A. Francis, Optimal Sampled-Data Control Systems, Springer-Verlag, London, UK, 1995.

[3] M. Souza, G.S. Deaecto, J.C. Geromel, J. Daafouz, Self-triggered linear quadratic networked control, Optimal Control Appl. Methods 35 (5) (2014) 524–538.

[4] A. Schlote, F. Häusler, T. Hecker, A. Bergmann, E. Crisostomi, I. Radusch, R. Shorten, Cooperative regulation and trading of emissions using plug-in hybrid vehicles, IEEE Trans. Intell. Transp. Syst. 14 (4) (2013) 1572–1585.

[5] J. Lunze, Feedback Control of Large Scale Systems, in: Ser. Systems and Control Engineering, Prentice Hall, Upper Saddle River, NJ, 1992.

[6] F.-Y. Wang, D. Liu, Networked Control Systems: Theory and Applications, Springer-Verlag, London, UK, 2008.

[7] P. Seiler, R. Sengupta, Analysis of communication losses in vehicle control problems, in: Proc. of the Americal Control Conference, Arlington, VA, June 2001.

[8] R.W. Clough, J. Penzien, Dynamics of Structures, second ed., McGraw-Hill, New York, NY, 1993.

[9] J.T. Wen, M. Arcak, A unifying passivity framework for network flow control, IEEE Trans. Automat. Control 49 (2) (2004) 162–174.

[10] M. Farina, P. Colaneri, R. Scattolini, Block-wise discretization accounting for structural constraints, Automatica 49 (11) (2013) 3411–3417.

[11] P. Colaneri, M. Farina, S. Kirkland, R. Scattolini, R. Shorten, Positive linear systems: discretization with positivity and structural constraints, in: Hybrid Systems and Constraints, ISTE-Wiley, 2013, pp. 1–20. (Chapter).

[12] B.D.O. Anderson, J.B. Moore, Optimal Control: Linear Quadratic Methods, Dover Publications, Mineola, NY, 2007.

[13] G.H. Golub, C.F. Van Loan, Matrix Computations, third ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[14] G.A. Baker Jr., P. Graves-Morris, Padé Approximants, second ed., Cambridge University Press, Cambridge, UK, 1996.

[15] R.N. Shorten, M. Corless, S. Sajja, S. Solmaz, On Padé approximations, quadratic stability and discretization of switched linear systems, Systems Control Lett. 60 (9) (2011) 683–689.

[16] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, Cambridge, UK, 2004.

[17] I. Markovsky, Structured low-rank approximation and its applications, Automatica 44 (4) (2007) 891–909.

[18] M. Souza, J.C. Geromel, P. Colaneri, R.N. Shorten, Discretisation of sparse linear systems: An optimisation approach, January 2015. [Online]. Available: http://www.hamilton.ie/bob/Discretisation_Arxiv.pdf.

[19] S. Boyd, L.E. Ghaoui, E. Feron, V. Balakrishnan, Linear Matrix Inequalities in System and Control Theory, in: Studies in Applied Mathematics, SIAM, Philadelphia, PA, 1994.

[20] E. Süli, D.F. Mayers, An Introduction to Numerical Analysis, Cambridge University Press, Cambridge, UK, 2003.
[21] B. Fiedler, T. Gedeon, A Lyapunov function for tridiagonal competitive-cooperative systems, SIAM J. Math. Anal. 30 (3) (1999) 469–478.
[22] J. Lévine, P. Rouchon, Quality control of binary distillation columns via nonlinear aggregate models, Automatica 27 (3) (1991) 463–480.
[23] L. Farina, S. Rinaldi, Positive Linear Systems: Theory and Applications, John Wiley & Sons, New York, NY, 2000.
[24] M. Souza, A.R. Fioravanti, J.C. Geromel, $\mathcal{H}_2$ sampled-data filtering of linear systems, IEEE Trans. Signal Process. 62 (18) (2014) 4839–4846.
[25] D.G. Luenberger, Introduction to Dynamic Systems: Theory, Models and Applications, John Wiley & Sons, New York, NY, 1979.
[26] S. Barnett, Matrices in Control Theory, Robert E. Krieger Publishing Company, Florida, USA, 1984.
[27] S. Sajja, S. Solmaz, R. Shorten, M. Corless, Preservation of common quadratic Lyapunov functions and Padé approximations, in: Proc. of the 49th IEEE Conf. on Dec. and Contr., Atlanta, USA, December 2010, pp. 7334–7338.
[28] T. Mori, T. Nguyen, Y. Mori, H. Kokame, Preservation of Lyapunov functions under bilinear mapping, Automatica 42 (6) (2006) 1055–1058.
[29] A. Leon Garcia, Probability, Statistics and Random Processes for Electrical Engineering, third ed., Prentice Hall, Upper Saddle River, NJ, 2008.
[30] B.T. Stewart, A.N. Venkat, J.B. Wright, G. Pannocchia, Cooperative distributed model predictive control, Systems Control Lett. 59 (8) (2010) 460–469.