Invariance of symmetric convex sets for discrete-time saturated systems

Mirko Fiacchini, Sophie Tarbouriech, Christophe Prieur

Invariance of symmetric convex sets for discrete-time saturated systems

Mirko Fiacchini, Sophie Tarbouriech and Christophe Prieur

***Abstract*— The characterization of invariance and contractiveness for discrete-time saturated linear systems is considered. The geometrical approach used to analyze the problem leads to conditions valid for generic symmetric convex sets. The application of the results to the ellipsoidal case generalizes known results and leads to computational improvements.**

I. Introduction

Invariance has become fundamental for the analysis and design of control systems. The importance of invariant sets in control is due to stability and robustness implicit properties of these regions of the state space. Many results regarding invariance and related topics have been provided in literature: see, for instance, the notable pioneering contribution [4], the works [10], [15], concerning the maximal invariant set, and [18] regarding the minimal one. The problem of obtaining invariant sets for discrete-time nonlinear systems is dealt with using ellipsoids in [16], parallelotopes in [7], and polytopes in [1], [8]. Invariance of polytopes for continuous-time nonlinear systems has been considered in [9]. A recent monograph on invariance and set-theory in control is [5].

Among the nonlinear systems, particular interest has been devoted to the saturated linear ones, as saturation is a very common nonlinearity, potentially present in every real plant. The computation of invariant ellipsoids for saturated linear systems has been addressed in the works [2], [12]–[14], [22]. Alternatively, methods to obtain polytopic invariant sets are proposed for saturated systems in [3], [11], [17].

The main purpose of this paper is to characterize geometrically invariance and λ-contractiveness for discrete-time saturated linear systems. Using properties of support functions and convex analysis, conditions for a generic symmetric convex set Ω to be invariant and λ-contractive will be stated. In particular, the condition is posed to ensure that every scaled set $\alpha\Omega$, with α positive and smaller than one, is λ-contractive. It is worth recalling that this condition determines implicitly a local Lyapunov function. Such general condition is then applied to the ellipsoidal case. The geometrical approach provides a deeper insight on the problem, which permits to recover and to generalize well established results. In particular, it will be shown that computational improvements are achieved by carefully considering the geometrical structure of the problem.

M. Fiacchini and S. Tarbouriech are with CNRS; LAAS; 7 avenue du colonel Roche, F-31077 Toulouse, France, Université de Toulouse; UPS, INSA, INP, ISAE, UT1, UTM, LAAS; F-31077 Toulouse, France. {fiacchini, sophie.tarbouriech}@laas.fr.

C. Prieur is with Department of Automatic Control, Gipsa-lab, Domaine universitaire, 961 rue de la Houille Blanche, BP 46, 38402 Grenoble Cedex, France. christophe.prieur@gipsa-lab.grenoble-inp.fr.

The paper is organized as follows: Section II presents the problem statement. Section III provides the characterization of invariance for symmetric convex set for saturated systems. In Section IV the ellipsoidal case is presented and compared with existing methods. In Section V two numerical examples are detailed. The paper ends with a section of conclusions.

Notation

The set of positive integers smaller than or equal to the integer $n \in \mathbb{N}$ is denoted as $\mathbb{N}_n$, i.e. $\mathbb{N}_n = \{x \in \mathbb{N} : 1 \leq x \leq n\}$. Given $A \in \mathbb{R}^{n\times m}$, A_i with $i \in \mathbb{N}_n$ denotes its i-th row, $A_{(j)}$ with $j \subseteq \mathbb{N}_m$ its j-th column. Given a symmetric matrix $P \in \mathbb{R}^{n\times n}$, notation $P > 0$ ($P \geq 0$) means that P is positive (semi-)definite, as usual. Given $D \subseteq \mathbb{R}^n$ and a scalar $\alpha \geq 0$, denote the set $\alpha D = \{\alpha x \in \mathbb{R}^n : \ x \in D\}$. The interior of a set D is denoted as $\text{int}(D)$, its boundary is ∂D. Given $P \in \mathbb{R}^{n\times n}$ with $P > 0$, define the ellipsoid $\mathscr{E}(P) = \{x \in \mathbb{R}^n : x^T P x \leq 1\}$.

II. Problem statement

Consider the discrete-time saturated linear system

$$x^+ = f(x) = Ax + B\varphi(Kx), \tag{1}$$

where $x \in \mathbb{R}^n$ is the current state, $x^+ \in \mathbb{R}^n$ is the successor and the saturated feedback control is given by $u = \varphi(Kx) \in \mathbb{R}^m$. Function $\varphi : \mathbb{R}^m \to \mathbb{R}^m$ denotes the saturation function, i.e. $\varphi_i(y) = \text{sgn}(y_i)\min\{|y_i|, 1\}$, for every $i \in \mathbb{N}_m$. A useful tool when dealing with convex closed sets is the support function.

Definition 1: Given a set $D \subseteq \mathbb{R}^n$, the support function of D evaluated at $\eta \in \mathbb{R}^n$ is $\phi_D(\eta) = \sup_{x\in D} \ \eta^T x$.

A geometrical meaning of the support function of D at η is the signed "distance" of the further point of D (or its closure) from the origin, along the direction η. See [19], [20] for properties of support functions. In particular, we recall below that set inclusion conditions can be given in terms of linear inequalities involving the support functions, see [19].

Property 1: Given $D, C \subseteq \mathbb{R}^n$, closed and convex, then $x \in D$ if and only if $\eta^T x \leq \phi_D(\eta)$, for all $\eta \in \mathbb{R}^n$, and $C \subseteq D$ if and only if $\phi_C(\eta) \leq \phi_D(\eta)$, for all $\eta \in \mathbb{R}^n$.

Invariance and λ-contractiveness of a closed convex set can be posed in terms of support functions, since their definitions involve set inclusion relations, see [5].

Definition 2: A set $D \subseteq \mathbb{R}^n$ is an invariant set for the system $x^+ = f(x)$ with $x \in X$ if $D \subseteq X$ and $f(x) \in D$, for all $x \in D$.

Recall that any trajectory starting in an invariant set D remains confined in it.

Definition 3: A convex compact set $D \subseteq \mathbb{R}^n$ with $0 \in \text{int}(D)$ is said to be a λ-contractive set for the system

$x^+ = f(x)$ with $x \in X$ if $D \subseteq X$ and, for a suitable $\lambda \in [0,1]$, is such that $f(x) \in \lambda D$, for all $x \in D$.

Since λ-contractiveness induces invariance, guaranteeing λ-contractiveness of a set implicitly ensures also invariance.

The property of λ-contractiveness of a compact convex set can be used to impose a local Lyapunov function. In particular, we are interested in a condition on convex compact set $\Omega \subseteq \mathbb{R}^n$, with $0 \in \text{int}(\Omega)$, whose satisfaction ensures that every set $\alpha\Omega$, with $\alpha \in [0,1]$, is λ-contractive, that is $f(x) \in \lambda\,\alpha\Omega$, for all $x \in \alpha\Omega$, with $\lambda \in [0,1]$. This, with $\lambda < 1$, would imply that there exists a local Lyapunov function defined on Ω, whose level sets are $\alpha\Omega$ with $\alpha \in [0,1]$. Hence, it is necessary to characterize λ-contractiveness of sets $\alpha\Omega$, for all $\alpha \in [0,1]$, in terms of support functions. First we introduce the Minkowski function of a convex, compact set $D \subseteq \mathbb{R}^n$ with $0 \in \text{int}(D)$, at $x \in \mathbb{R}^n$, that is defined as

$$\Psi_D(x) = \min_{\alpha \geq 0} \{\alpha \in \mathbb{R} : \ x \in \alpha D\}.$$

The geometric meaning of the Minkowski function of $D \subseteq \mathbb{R}^n$ at $x \in \mathbb{R}^n$ is close to the concept of distance of x from the origin. In fact, given D and $x \in \mathbb{R}^n$, the value of $\Psi_D(x)$ is how much the set D should be scaled for x to be on its boundary, that is such that $x \in \partial(\Psi_D(x)D)$. Then $x \in \partial\Omega(x)$, where

$$\Omega(x) = \Psi_\Omega(x)\Omega. \tag{2}$$

The set $\Omega(x)$ is useful to determine the condition for the set $\alpha\Omega$ to be λ-contractive for the saturated system (1). Such condition is given by a (possibly uncountable) set of nonconvex constraints, as stated in the following proposition.

Proposition 1: Given the system (1), the convex, compact set Ω with $0 \in \text{int}(\Omega)$ is such that $\alpha\Omega$ is λ-contractive for every $\alpha \in [0,1]$, with $\lambda \in [0,1]$, if and only if

$$\eta^T f(x) \leq \lambda\, \phi_{\Omega(x)}(\eta), \quad \forall x \in \Omega, \ \forall \eta \in \mathbb{R}^n. \tag{3}$$

Proof: By definition, the set $\alpha\Omega$ is λ-contractive for every $\alpha \in [0,1]$ if and only if $x^+ \in \lambda\Omega(x)$, for all $x \in \Omega$. This is equivalent, by Property 1, to (3). ■

III. Invariance for symmetric convex sets

One key concept that will be used in the following is convexity. First, we define the following functions on $\mathbb{R}^m$

$$\begin{aligned} \check{\varphi}_i(y) &= \max\{y_i, -1\} = \begin{cases} y_i & \text{if } y_i \geq -1, \\ -1 & \text{if } y_i < -1, \end{cases} \\ \hat{\varphi}_i(y) &= \min\{y_i, 1\} = \begin{cases} y_i & \text{if } y_i \leq 1, \\ 1 & \text{if } y_i > 1, \end{cases} \end{aligned} \tag{4}$$

for every $i \in \mathbb{N}_m$, whose convexity related properties are stated in the following.

Property 2: Functions $\check{\varphi}_i : \mathbb{R}^m \to \mathbb{R}$ and $\hat{\varphi}_i : \mathbb{R}^m \to \mathbb{R}$, in (4), are convex and concave, respectively, and such that

$$\hat{\varphi}_i(y) \leq \varphi_i(y) \leq \check{\varphi}_i(y), \tag{5}$$

for all $y \in \mathbb{R}^m$ and for every $i \in \mathbb{N}_m$.

Proof: Convexity of $\check{\varphi}_i(\cdot)$ over $\mathbb{R}^m$ follows directly from the fact that the pointwise maximum of convex functions is convex, see [6]. Analogously, $\hat{\varphi}_i$ is concave on $\mathbb{R}^m$ since any pointwise minimum of concave functions is concave. Furthermore, for any $i \in \mathbb{N}_m$, we have three possible cases: if $y_i > 1$ then $\hat{\varphi}_i(y) = \varphi_i(y) \leq \check{\varphi}_i(y)$; if $|y_i| \leq 1$ then $\hat{\varphi}_i(y) = \varphi_i(y) = \check{\varphi}_i(y)$; if $y_i < -1$ then $\hat{\varphi}_i(y) \leq \varphi_i(y) = \check{\varphi}_i(y)$. In any case, the relation (5) holds. ■

The bounding functions $\hat{\varphi}(\cdot)$ and $\check{\varphi}(\cdot)$ are used to determine an upper bounding function of $\eta^T f(x)$ for any $\eta \in \mathbb{R}^n$, with $f(\cdot)$ characterizing the saturated system (1).

Definition 4: Define the function $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as

$$F(x,\eta) = \eta^T A x + \sum_{i \in \mathbb{N}_m} v_i(x,\eta),$$

where, for every $i \in \mathbb{N}_m$ and with $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^n$,

$$v_i(x,\eta) = \begin{cases} \eta^T B_{(i)} \check{\varphi}_i(Kx) & \text{if } \eta^T B_{(i)} \geq 0, \\ \eta^T B_{(i)} \hat{\varphi}_i(Kx) & \text{if } \eta^T B_{(i)} < 0. \end{cases} \tag{6}$$

Function $F(\cdot,\eta)$ is a convex upper bound of $\eta^T f(\cdot)$, for any $\eta \in \mathbb{R}^n$, and it permits to pose λ-contractiveness conditions in terms of convex constraints.

Proposition 2: Given the system (1), the function $F(\cdot,\cdot)$, as in Definition 4, is such that

$$\eta^T f(x) \leq F(x,\eta), \quad \forall x \in \mathbb{R}^n, \ \forall \eta \in \mathbb{R}^n, \tag{7}$$

and $F(\cdot,\eta)$ is convex on $\mathbb{R}^n$, for every $\eta \in \mathbb{R}^n$.

Proof: Convexity of function $F(\cdot,\eta)$ is due to the fact that it is the sum of functions convex in x, for every $\eta \in \mathbb{R}^n$. In fact, $\eta^T A x$ is linear and terms $v_i(\cdot,\eta)$ are convex in x by definition, see (6), and from Property 2. Moreover, we have

$$\begin{cases} \eta^T B_{(i)} \varphi_i(Kx) \leq \eta^T B_{(i)} \check{\varphi}_i(Kx) & \text{if } \eta^T B_{(i)} \geq 0, \\ \eta^T B_{(i)} \varphi_i(Kx) \leq \eta^T B_{(i)} \hat{\varphi}_i(Kx) & \text{if } \eta^T B_{(i)} < 0, \end{cases}$$

for every $i \in \mathbb{N}_m$, which means that $\eta^T B_{(i)} \varphi_i(x) \leq v_i(x,\eta)$. Then, condition (7) follows. ■

Function $F(\cdot,\cdot)$ admits an alternative representation, more suitable to pose the condition for invariance in terms of linear matrix inequalities (LMI). The equivalence of the two representations is stated and proved below.

Proposition 3: Given the system (1), function $F(\cdot,\cdot)$, as in Definition 4, is such that, for every $x \in \mathbb{R}^n$ and every $\eta \in \mathbb{R}^n$,

$$F(x,\eta) = \eta^T A x + \sum_{i \in \mathbb{N}_m} \max\{\eta^T B_{(i)} K_i x, -|\eta^T B_{(i)}|\}. \tag{8}$$

Proof: It is sufficient to prove that

$$v_i(x,\eta) = \max\{\eta^T B_{(i)} K_i x, -|\eta^T B_{(i)}|\}, \tag{9}$$

for every $i \in \mathbb{N}_m$ with $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^n$, where $v_i(\cdot,\cdot)$ is defined in (6). From (4), it follows that

$$v_i(x,\eta) = \begin{cases} \eta^T B_{(i)} \max\{K_i x, -1\} & \text{if } \eta^T B_{(i)} \geq 0, \\ \eta^T B_{(i)} \min\{K_i x, 1\} & \text{if } \eta^T B_{(i)} < 0, \end{cases}$$

and then

$$v_i(x,\eta) = \begin{cases} \eta^T B_{(i)} \max\{K_i x, -1\} & \text{if } \eta^T B_{(i)} \geq 0, \\ -\eta^T B_{(i)} \max\{-K_i x, -1\} & \text{if } \eta^T B_{(i)} < 0. \end{cases}$$

Since $a \max h(x) = \max a h(x)$ for every $h(\cdot)$ and every $a \geq 0$, we have that

$$v_i(x,\eta) = \begin{cases} \max\{\eta^T B_{(i)} K_i x, -\eta^T B_{(i)}\} & \text{if } \eta^T B_{(i)} \geq 0, \\ \max\{\eta^T B_{(i)} K_i x, \eta^T B_{(i)}\} & \text{if } \eta^T B_{(i)} < 0, \end{cases}$$

which is equivalent to (9). ■

Before presenting the main contribution of the paper, some definitions are introduced to simplify the notation. Given the system (1), the state $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^n$ define

$$\begin{aligned} I^+(x) &= \{i \in \mathbb{N}_m : \ K_i x > 1\}, \\ I^-(x) &= \{i \in \mathbb{N}_m : \ K_i x < -1\}, \\ I^0(x) &= \{i \in \mathbb{N}_m : \ |K_i x| \leq 1\}, \end{aligned} \tag{10}$$

and

$$\begin{aligned} E^+(\eta) &= \{i \in \mathbb{N}_m : \ \eta^T B_{(i)} > 0\}, \\ E^-(\eta) &= \{i \in \mathbb{N}_m : \ \eta^T B_{(i)} < 0\}, \\ E^0(\eta) &= \{i \in \mathbb{N}_m : \ \eta^T B_{(i)} = 0\}. \end{aligned} \tag{11}$$

Clearly $I^+(x) \cup I^-(x) \cup I^0(x) = \mathbb{N}_m$ and $E^+(\eta) \cup E^-(\eta) \cup E^0(\eta) = \mathbb{N}_m$. Moreover we define

$$\begin{aligned} \mathscr{J}(\Omega) = \{J \subseteq \mathbb{N}_m : \ & \exists x \in \Omega, \eta \in \mathbb{R}^n \text{ s.t.} \\ & i \in J \ \Leftrightarrow \ \eta^T B_{(i)} K_i x < -|\eta^T B_{(i)}|\} \cup \{\emptyset\}. \end{aligned} \tag{12}$$

Notice that $J \in \mathscr{J}(\Omega)$ if and only if $J = \emptyset$ or there exist $x \in \Omega$ and $\eta \in \mathbb{R}^n$ such that $i \in J$ if and only if $i \in I^+(x)$ and $i \in E^+(\eta)$ or $i \in I^-(x)$ and $j \in E^-(\eta)$, for all $i \in \mathbb{N}_m$. It is worth illustrating the geometrical meaning of sets $J \in \mathscr{J}(\Omega)$, empty set apart. Consider the terms in the summation in (8). For every $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^n$, it follows, from (9), that $v_i(x,\eta) > \eta^T B_{(i)} K_i x$ if and only if $\eta^T B_{(i)} K_i x < -|\eta^T B_{(i)}|$. Hence, the set of indices $J \subseteq \mathbb{N}_m$ is in $\mathscr{J}(\Omega)$ if and only if there exists $x \in \Omega$ and $\eta \in \mathbb{R}^n$ such that $v_i(x,\eta) = -|\eta^T B_{(i)}| > \eta^T B_{(i)} K_i x$ for all (and only those) $i \in J$. Roughly speaking, we can think to elements J of $\mathscr{J}(\Omega)$ as the possible sets of indices such that the i-term in summation in (8) "saturates" if and only if $i \in J$. Then, given $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^n$, and denoting

$$\begin{aligned} J(x,\eta) = & \ \Big(I^+(x) \cap E^-(\eta)\Big) \cup \Big(I^-(x) \cap E^+(\eta)\Big), \\ \bar{J}(x,\eta) = & \ I^0(x) \cup \Big(I^+(x) \cap E^+(\eta)\Big) \cup \\ & \cup \Big(I^-(x) \cap E^-(\eta)\Big) \cup E^0(\eta), \end{aligned} \tag{13}$$

we have that $J(x,\eta) \cup \bar{J}(x,\eta) = \mathbb{N}_m$ and $\mathscr{J}(\Omega)$ is the set of all possible $J(x,\eta)$ for every $x \in \Omega$ and $\eta \in \mathbb{R}^n$ (and the empty set).

Theorem 1: Given the system (1), and the symmetric convex compact set $\Omega \subseteq \mathbb{R}^n$, with $0 \in \text{int}(\Omega)$, if for every $J \in \mathscr{J}(\Omega)$ and every $i \in J$, there exists $\sigma_i^J(x) \in \mathbb{R}$ such that $|\sigma_i^J(x)| \leq 1$ and

$$\begin{aligned} \eta^T A x + \sum_{i \in \bar{J}} \eta^T B_{(i)} K_i x + \sum_{i \in J} \sigma_i^J(x) \eta^T B_{(i)} \leq \\ \leq \lambda \phi_{\Omega(x)}(\eta), \quad \forall \eta \in \mathbb{R}^n, \ \forall x \in \Omega, \end{aligned} \tag{14}$$

then $\alpha\Omega$ is λ-contractive, with $\lambda \in [0,1]$, for every $\alpha \in [0,1]$.

Proof: First notice that, from Proposition 2, a sufficient condition for $\alpha\Omega$ to be λ-contractive, for all $\alpha \in [0,1]$, is

$$F(x,\eta) \leq \lambda \phi_{\Omega(x)}(\eta), \quad \forall x \in \Omega, \ \forall \eta \in \mathbb{R}^n. \tag{15}$$

We have to prove that (14) implies (15). Fix $x \in \Omega$ and $\eta \in \mathbb{R}^n$ and denote $I^+ = I^+(x)$, $I^- = I^-(x)$, $I^0 = I^0(x)$ and $E^+ = E^+(\eta)$, $E^- = E^-(\eta)$. From Proposition 3, we have that

$$\begin{aligned} F(x,\eta) = \eta^T A x + \sum_{i \in I^0} \eta^T B_{(i)} K_i x + \sum_{i \in I^+ \cap E^+} \eta^T B_{(i)} K_i x + \\ + \sum_{i \in I^- \cap E^-} \eta^T B_{(i)} K_i x + \sum_{i \in I^+ \cap E^-} \eta^T B_{(i)} + \sum_{i \in I^- \cap E^+} (-\eta^T B_{(i)}), \end{aligned}$$

is a valid representation of $F(x,\eta)$ for all $x \in \Omega$ such that $I^+(x) = I^+$, $I^-(x) = I^-$ and for every $\eta \in \mathbb{R}^n$ such that $E^+(\eta) = E^+$ and $E^-(\eta) = E^-$. From definitions (11) and (13), posing $\bar{J} = \bar{J}(x,\eta)$, we have that condition

$$\begin{aligned} F(x,\eta) = \eta^T A x + \sum_{i \in \bar{J}} \eta^T B_{(i)} K_i x + \sum_{i \in I^+ \cap E^-} \eta^T B_{(i)} + \\ + \sum_{i \in I^- \cap E^+} (-\eta^T B_{(i)}) \leq \lambda \phi_{\Omega(x)}(\eta), \end{aligned}$$

implies the satisfaction of (3) for any $\eta \in \mathbb{R}^n$ such that $\eta^T B_{(i)} > 0$ if $i \in E^+$ and $\eta^T B_{(i)} < 0$ if $i \in E^-$. Applying the S-procedure we find the following equivalent condition

$$\begin{aligned} \eta^T A x + \sum_{i \in \bar{J}} \eta^T B_{(i)} K_i x + \sum_{i \in \bar{J} \cap E^+} \tau_i \eta^T B_{(i)} + \sum_{i \in \bar{J} \cap E^-} (-\tau_i \eta^T B_{(i)}) + \\ + \sum_{i \in I^+ \cap E^-} (1 - \tau_i) \eta^T B_{(i)} + \sum_{i \in I^- \cap E^+} (\tau_i - 1) \eta^T B_{(i)} \leq \lambda \phi_{\Omega(x)}(\eta), \end{aligned}$$

for $\tau_i = \tau_i(x) \geq 0$ for all $i \in E^+ \cup E^-$. Thus, if there exist $\sigma_i(x) \in \mathbb{R}$ for all $i \in E^+ \cup E^-$ such that

$$\eta^T A x + \sum_{i \in \bar{J}} \eta^T B_{(i)} K_i x + \sum_{i \in E^+ \cup E^-} \sigma_i(x) \eta^T B_{(i)} \leq \lambda \phi_{\Omega(x)}(\eta), \tag{16}$$

for all $\eta \in \mathbb{R}^n$, and

$$\begin{cases} \sigma_i(x) = \tau_i(x) \geq 0, & \text{if } i \in \bar{J} \cap E^+, \\ \sigma_i(x) = -\tau_i(x) \leq 0, & \text{if } i \in \bar{J} \cap E^-, \\ \sigma_i(x) = 1 - \tau_i(x) \leq 1, & \text{if } i \in I^+ \cap E^-, \\ \sigma_i(x) = \tau_i(x) - 1 \geq -1, & \text{if } i \in I^- \cap E^+, \end{cases} \tag{17}$$

then $F(x,\eta) \leq \lambda \phi_{\Omega(x)}(\eta)$ for all $\eta \in \mathbb{R}^n$ such that $\eta^T B_{(i)} > 0$ if $i \in E^+$ and $\eta^T B_{(i)} < 0$ if $i \in E^-$.

Consider now the point $\bar{x} = -x$ and $\bar{\eta} = -\eta$, clearly $\bar{x} \in \Omega$ by symmetry of Ω. Following a logical process analogous to the one illustrated above, and since

$$\begin{aligned} & I^+ = I^+(x) = I^-(-x) = I^-(\bar{x}), \\ & I^- = I^-(x) = I^+(-x) = I^+(\bar{x}), \\ & \bar{J} = \bar{J}(x,\eta) = \bar{J}(-x,-\eta) = \bar{J}(\bar{x},\bar{\eta}), \\ & E^+ = E^+(\eta) = E^-(-\eta) = E^-(\bar{\eta}), \\ & E^- = E^-(\eta) = E^+(-\eta) = E^+(\bar{\eta}), \end{aligned}$$

we can determine a condition in terms of x and η ensuring that $F(\bar{x},\bar{\eta}) \leq \lambda \phi_{\Omega(\bar{x})}(\bar{\eta})$ for all $\bar{\eta} \in \mathbb{R}^n$ such that $\bar{\eta}^T B_{(i)} < 0$ if $i \in E^+$ and $\bar{\eta}^T B_{(i)} > 0$ if $i \in E^-$. Such condition is the existence of $\sigma_i(-x) \in \mathbb{R}$ for all $i \in E^+ \cup E^-$ such that

$$\eta^T A x + \sum_{i \in \bar{J}} \eta^T B_{(i)} K_i x + \sum_{i \in E^+ \cup E^-} \sigma_i(-x) \eta^T B_{(i)} \leq \lambda \phi_{\Omega(x)}(\eta), \tag{18}$$

for all $\eta \in \mathbb{R}^n$, and

$$\begin{cases} \sigma_i(-x) \leq 0, & \text{if } i \in \bar{J} \cap E^+, \\ \sigma_i(-x) \geq 0, & \text{if } i \in \bar{J} \cap E^-, \\ \sigma_i(-x) \geq -1, & \text{if } i \in I^+ \cap E^-, \\ \sigma_i(-x) \leq 1, & \text{if } i \in I^- \cap E^+. \end{cases} \tag{19}$$

Condition (18) is obtained by replacing $\bar{x}$ with $-x$ and $\bar{\eta}$ with $-\eta$ in $F(\bar{x},\bar{\eta}) \leq \lambda \phi_{\Omega(\bar{x})}(\bar{\eta})$. Notice that conditions (16)-(17) and (18)-(19), which are imposed for $x \in \Omega$, are substantially the same. The only difference is on the constraints (17) on variables $\sigma_i(x)$ and (19) on $\sigma_i(-x)$, with $i \in E^+ \cap E^-$. Then they are both satisfied if and only if there exists $\sigma_i^J(x)$ for $i \in \big(I^+ \cap E^-\big) \cup \big(I^- \cap E^+\big) = J(x,\eta)$ such that $|\sigma_i^J(x)| \leq 1$

and (14) holds at x. Since such condition has to be posed for every $x \in \Omega$ and every $\eta \in \mathbb{R}^n$ and by definition of $\mathscr{I}(\Omega)$, the theorem is proved. ■

Theorem 1, concerning generic symmetric convex compact sets, is particularized in what follows to ellipsoids.

IV. Invariance for ellipsoids

In this section we focus on a relaxed condition for λ-contractiveness of ellipsoidal sets $\alpha\Omega$ for any $\alpha \in [0,1]$, based on convex constraints. The aim of computational tractability of the related problem is achieved by restricting the choice of functions $\sigma_i^J(x)$ to linear functions. Given the ellipsoid $\Omega = \mathscr{E}(P)$, with $P \in \mathbb{R}^{n\times n}$ symmetric and positive definite, and $x \in \mathbb{R}^n$, the Minkowski function is $\Psi_\Omega(x) = \sqrt{x^T P x}$, and, since $\alpha\Omega = \{x \in \mathbb{R}^n : x^T P x \le \alpha^2\}$, then

$$\Omega(x) = \Psi_\Omega(x)\Omega = \{y \in \mathbb{R}^n : y^T P y \le x^T P x\}. \tag{20}$$

First we provide a characterization of λ-contractiveness of ellipsoids $\alpha\Omega$, with $\alpha \in [0,1]$.

Proposition 4: Given the system (1), the ellipsoid $\Omega = \mathscr{E}(P)$, with $P \in \mathbb{R}^{n\times n}$ and $P > 0$, is such that $\alpha\Omega$ is λ-contractive, with $\lambda \in [0,1]$, for any $\alpha \in [0,1]$ if and only if

$$\eta^T f(x) \le \lambda \sqrt{x^T P x}\sqrt{\eta^T P^{-1}\eta}, \quad \forall x \in \Omega,\ \forall \eta \in \mathbb{R}^n.$$

Proof: From Proposition 1, we have only to prove that $\phi_{\Omega(x)}(\eta) = \sqrt{x^T P x}\sqrt{\eta^T P^{-1}\eta}$ for all $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^n$. Recall that $\phi_{\mathscr{E}(Q)}(\eta) = \sqrt{\eta^T Q^{-1}\eta}$, for every $Q > 0$ and any $\eta \in \mathbb{R}^n$, see [5]. Then, defining $\tilde{P}(x) = (x^T P x)^{-1} P$, we have

$$\begin{aligned}\phi_{\Omega(x)}(\eta) &= \sup_{y\in\Omega(x)} \eta^T y = \sup_{y^T P y \le x^T P x} \eta^T y = \sup_{y^T \tilde{P}(x) y \le 1} \eta^T y \\ &= \sqrt{\eta^T \tilde{P}(x)^{-1}\eta} = \sqrt{x^T P x}\sqrt{\eta^T P^{-1}\eta},\end{aligned}$$

which proves the proposition. ■

Condition for invariance provided in Theorem 1 involves an infinite number of constraints in $x \in \Omega$, not necessarily convex, one for any $\eta \in \mathbb{R}^n$. In the ellipsoidal case, the explicit dependence on η can be removed, as illustrated in the following, to obtain a formulation of the condition involving only the state x. In this section we denote, with a slight abuse of notation, $\mathscr{I}(P) = \mathscr{I}(\mathscr{E}(P))$.

Corollary 1: Given the system (1), and the ellipsoid $\Omega = \mathscr{E}(P)$, with $P \in \mathbb{R}^{n\times n}$ and $P > 0$, if for every $J \in \mathscr{I}(P)$ and every $i \in J$, there exists $\sigma_i^J(x) \in \mathbb{R}$ such that $|\sigma_i^J(x)| \le 1$ and

$$M(x,J)^T P\, M(x,J) \le \lambda^2 x^T P x, \quad \forall x \in \mathscr{E}(P), \tag{21}$$

where $M(x,J) = Ax + \sum_{i\in J} B_{(i)} K_i x + \sum_{i\in J} B_{(i)} \sigma_i^J(x)$, then $\alpha\Omega$ is λ-contractive, with $\lambda \in [0,1]$, for every $\alpha \in [0,1]$.

Proof: We prove that conditions (14) and (21) are equivalent for the case of $\Omega = \mathscr{E}(P)$. For every $J \in \mathscr{I}(P)$, the condition in (14) can be posed, from Proposition 4, as

$$\eta^T M(x,J) \le \lambda \sqrt{x^T P x}\sqrt{\eta^T P^{-1}\eta}, \quad \forall \eta \in \mathbb{R}^n,\ \forall x \in \mathscr{E}(P). \tag{22}$$

Given $\eta \in \mathbb{R}^n$ with $\eta \neq 0$, define $\hat{\eta} = (\eta^T P^{-1}\eta)^{-1/2}\eta$, and notice that $\hat{\eta} \in \partial\mathscr{E}(P^{-1})$, in fact $\hat{\eta}^T P^{-1}\hat{\eta} = 1$. Thus, apart from the trivial case of $\eta = 0$, (22) is equivalent to

$$\hat{\eta}^T M(x,J) \le \lambda \sqrt{x^T P x}, \quad \forall \hat{\eta} \in \partial\mathscr{E}(P^{-1}),\ \forall x \in \mathscr{E}(P),$$

and then, since the supremum of a linear function over a bounded convex set is attained at its boundary, we have

$$\begin{aligned}\sup_{\hat{\eta}\in\partial\mathscr{E}(P^{-1})} M(x,J)^T\hat{\eta} &= \sup_{\hat{\eta}\in\mathscr{E}(P^{-1})} M(x,J)^T\hat{\eta} = \\ &= \phi_{\mathscr{E}(P^{-1})}(M(x,J)) \le \lambda\sqrt{x^T P x}, \quad \forall x \in \mathscr{E}(P).\end{aligned}$$

From the expression of the support function of $\mathscr{E}(P^{-1})$ at $M(x,J)$, (21) follows. ■

It can be proved, see [14], that conditions of Theorem 1 and Corollary 1 are also necessary, besides of sufficient, for λ-contractiveness of $\alpha\Omega$, for all $\alpha \in [0,1]$, with $m = 1$.

Notice that the condition for λ-contractiveness of ellipsoids $\alpha\Omega$, for all $\alpha \in [0,1]$, presented by Corollary 1, consists in possibly nonconvex constraints. This condition is relaxed to obtain a sufficient condition given by convex constraints, by assuming linearity of $\sigma_i^J(x)$, for all $J \in \mathscr{I}(P)$ and $i \in J$.

Proposition 5: Given the system (1), and the ellipsoid $\Omega = \mathscr{E}(P)$, with $P \in \mathbb{R}^{n\times n}$ and $P > 0$, if for every $J \in \mathscr{I}(P)$ and every $i \in J$, there exists $H(i,J) \in \mathbb{R}^{1\times n}$ such that

$$\begin{aligned}&H(i,J)P^{-1}H(i,J)^T \le 1, && \forall J \in \mathscr{I}(P),\ \forall i \in J, \\ &N(J)^T P N(J) \le \lambda^2 P, && \forall J \in \mathscr{I}(P),\end{aligned}$$

where $N(J) = A + \sum_{i\in J} B_{(i)} K_i + \sum_{i\in J} B_{(i)} H(i,J)$, then $\alpha\Omega$ is λ-contractive, with $\lambda \in [0,1]$, for every $\alpha \in [0,1]$.

Proof: The property follows directly from Corollary 1 imposing linearity of functions $\sigma_i^J(x)$, that is $\sigma_i^J(x) = H(i,J)x$, for all $J \in \mathscr{I}(P)$ and every $i \in J$. ■

Comparison and computational considerations

The main improvements of the proposed results are illustrated by comparison with some existing methods. First, we recall the main result of the work [14].

Theorem 2: Given the system (1), and the ellipsoid $\Omega = \mathscr{E}(P)$, with $P \in \mathbb{R}^{n\times n}$ and $P > 0$, if there exists $L \in \mathbb{R}^{m\times n}$ such that

$$\begin{aligned}&L_i P^{-1} L_i^T \le 1, && \forall i \in \mathbb{N}_m, \\ &N(J)^T P N(J) \le \lambda P, && \forall J \subseteq \mathbb{N}_m,\end{aligned} \tag{23}$$

where $N(J) = A + \sum_{i\in J} B_{(i)} K_i + \sum_{i\in J} B_{(i)} L_i$, then $\alpha\Omega$ is λ-contractive, with $\lambda \in [0,1]$, for every $\alpha \in [0,1]$.

Notice that there are analogies with results presented in Proposition 5, but also important differences. First of all notice that the second condition in (23) is imposed for every possible subset of $\mathbb{N}_m$ while our result involves only the appropriately selected subsets of $\mathbb{N}_m$, that we denoted $\mathscr{I}(P)$. The presence of further constraints in the condition of [14] implies that our proposal is less conservative. Moreover the matrices $H(i,J)$ in Proposition 5 are replaced with the rows of a single matrix L in Theorem 2. Any solution obtained by the method in Theorem 2 can be recovered by posing $H(i,J) = L_i$, for every J and $i \in J$, in our condition.

The work [2] provides an improved version of the sufficient condition for λ-contractiveness, although for continuous-time systems. An analogous criterion can be formulated for the discrete-time case.

Theorem 3: Given the system (1), and the ellipsoid $\Omega = \mathscr{E}(P)$, with $P \in \mathbb{R}^{n\times n}$ and $P > 0$, if for every $J \subseteq \mathbb{N}_m$ and every $i \in J$, there exists $G(i,J) \in \mathbb{R}^{1\times n}$ such that

$$\begin{array}{ll} G(i,J)P^{-1}G(i,J)^T \leq 1, & \forall J \subseteq \mathbb{N}_m,\ \forall i \in J, \\ N(J)^T P N(J) \leq \lambda P, & \forall J \subseteq \mathbb{N}_m, \end{array}$$

where $N(J) = A + \sum_{i\in\overline{J}} B_{(i)}K_i + \sum_{i\in J} B_{(i)}G(i,J)$, then $\alpha\Omega$ is λ-contractive, with $\lambda \in [0,1]$, for every $\alpha \in [0,1]$.

The method proposed in Theorem 3 is more general than that one in Theorem 2, since it introduces more variables in place of matrix L, reducing the conservativeness, see [2] (although for the continuous-time case). On the other hand, the constraints still involve every subset of $\mathbb{N}_m$, as for [14].

Remark 1: The fact that a quadratic constraint is imposed for every $J \subseteq \mathbb{N}_m$ in spite of for $J \in \mathscr{I}(P)$, implies that, also in this case, any solution satisfying condition of Theorem 3 fulfils condition of Proposition 5 too. This leads to a smaller or equal degree of conservativeness of our result with respect to Theorem 3 (and thus also to Theorem 2). Nevertheless, we have not been able yet to find a solution of condition in Proposition 5 which does not satisfy also Theorem 3 (that is, to prove that our approach is "strictly" less conservative than latter). On the other hand, the numerical benefits of our approach are evident, as the LMIs involved in condition of Theorem 3 might be much more than those of our condition. This fact is illustrated by Examples 1 and 2.

Remark 2: The set $\mathscr{I}(\Omega)$ can be determined by means of a finite number of quadratic programming problems. In fact, considering $\Omega = \mathbb{R}^n$, the solution of a quadratic programming problem in $2n$ variables can determine whether the set $J \in \mathbb{N}_m$ belongs to $\mathscr{I}(\mathbb{R}^n)$ or not. Consequently, 2^m quadratic $2n$-dimensional optimization problems can be posed to define $\mathscr{I}(\mathbb{R}^n)$, which is such that $\mathscr{I}(\Omega) \subseteq \mathscr{I}(\mathbb{R}^n)$.

It is also worth stressing that methods presented in [14] and [2] have been compared with the relaxed, computational oriented, results of Proposition 5. Considering generic functions $\sigma_i^J(x)$ for every $J \in \mathscr{I}(P)$ and every $i \in J$, as in Theorem 1 and Corollary 1, leads to more general theoretical results and provides a deeper insight on the problem.

V. Numerical examples

We provide here two numerical examples to illustrate the benefits of the proposed approach.

Example 1: This simple example has the only purpose of comparing the result obtained considering the constraints for $J \subseteq \mathbb{N}_m$ in spite of for $J \in \mathscr{I}(P)$. Although the example is rather artificial, it provides an insight on how, also for low dimensional systems, the results can be affected by the improper choice of sets $J \subseteq \mathbb{N}_m$. We consider a case in which the cardinality of $\mathscr{I}(P)$ is much smaller than 2^m and compare the results obtained using Theorem 3 and Proposition 5.

Consider the system (1) with $n = 2$ and $m = 7$, where

$$A = \begin{bmatrix} 0.8876 & -0.5555 \\ 0.5555 & 1.5542 \end{bmatrix},$$

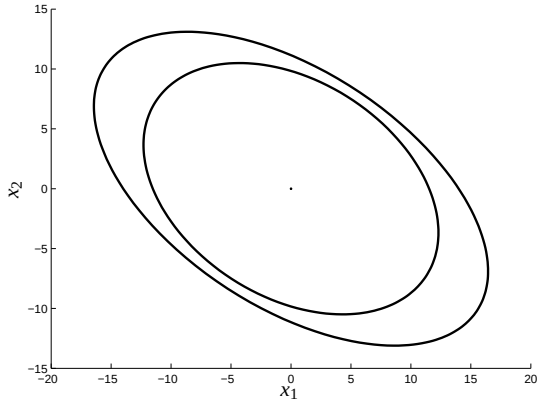

Fig. 1. Ellipsoidal estimations obtained with Theorem 3 (inner) and Proposition 5 (outer).

whose eigenvalues are $1.2209 \pm 0.4444i$, and

$$B = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & -1 \\ 0 & -1 \\ 0 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix}^T, \quad K = \begin{bmatrix} 0.1847 & -0.1136 \\ -0.1847 & 0.1136 \\ -0.0988 & -0.2734 \\ -0.0988 & -0.2734 \\ 0.0988 & 0.2734 \\ 0.0858 & -0.3870 \\ 0.2835 & 0.1598 \end{bmatrix},$$

and $\lambda = 1$. Matrix $K \in \mathbb{R}^{7\times 2}$ is the LQR gain with $Q = I_n$ and $R = I_m$ and the eigenvalues of $A + BK$ are $0.1681 \pm 0.0764i$. Notice the particular structures of matrices B and K and consider for instance $B_{(i)}$ and K_i for $i = 1$ and $i = 2$. It is evident that there is not $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^n$ such that

$$\begin{cases} \eta^T B_{(1)} K_1 x < -|\eta^T B_{(1)}|, \\ \eta^T B_{(2)} K_2 x \geq -|\eta^T B_{(2)}|, \end{cases}$$

simply because $B_{(1)} = -B_{(2)}$ and $K_1 = -K_2$. From definition of $\mathscr{I}(P)$, see (12), none of the elements $J \subseteq \mathbb{N}_m$ such that $1 \in J$ and $2 \notin J$ belongs to $\mathscr{I}(P)$, for every positive definite matrix $P \in \mathbb{R}^{n\times n}$. Similarly, if $1 \notin J$ and $2 \in J$, then $J \notin \mathscr{I}(P)$, for all $P \in \mathbb{R}^{n\times n}$. Hence, if $J \in \mathscr{I}(P)$ then either $1 \in J$ and $2 \in J$, or $1 \notin J$ and $2 \notin J$. Many other subsets of $\mathbb{N}_m$ do not belong to $\mathscr{I}(P)$. Similar considerations can be posed on the third, forth and fifth elements of B and K. Finally we find that, among the $2^7 = 128$ sets $J \in \mathbb{N}_m$, only 7 (or less) of them compose $\mathscr{I}(P)$. This implies simpler optimization problems (and then lower numerical sensibility) besides of potentially smaller degree of conservativeness of the results.

An ellipsoidal estimation of the domain of attraction has been computed using Proposition 5 and maximizing the scaling factor β such that $\beta\Gamma \subseteq \Omega$, where Γ is a given polytope. The optimal solution is the outer ellipsoid in Figure 1. To remove the dependency on P of the set $\mathscr{I}(P)$, we considered the degenerate ellipsoid $\mathscr{E}(P) = \mathbb{R}^n$ in definition (12). Then, ellipsoidal estimations are computed using Theorem 3 and employing two semi-definite programming solvers in MATLAB. The solution obtained with *SEDUMI* solver is the same as that one obtained using Proposition 5. On the contrary, *SDPT3* solver provides as optimal solution the inner ellipsoid depicted in Figure 1. Hence the optimal solutions of both methods seem to be the same, but the

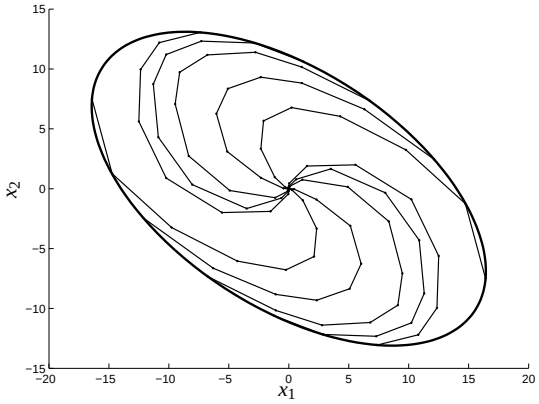

Fig. 2. Set Ω and trajectories of the system starting on its boundary.

higher computational burden required by Theorem 3 affected numerically the solver, leading to a suboptimal solution.

The invariance of $\Omega = \mathscr{E}(P)$ is checked by computing the trajectories of the system, for different initial conditions, see Figure 2. Notice that every trajectory remains bounded inside the set Ω and converges to the origin.

Example 2: This example shows that the number of LMIs required to obtain the maximal λ-contractive ellipsoid can be consistently reduced by using the proposed method. In particular we show that the cardinality of $\mathscr{I}(\Omega)$ can be much smaller than the number of sets $J \subseteq \mathbb{N}_m$, also for generic systems. As seen in the previous example, the structure of matrices B and K can determine the set $\mathscr{I}(\mathbb{R}^n)$. For different values of $n \in \mathbb{N}$ and $m \in \mathbb{N}$, with $m \leq n$, we consider a matrix B whose first $p < m$ columns are randomly generated, and the other $m-p$ columns are linear combinations of the first p. That means, roughly speaking, that B has rank $p < m$. Notice that the elements of $\mathscr{I}(\mathbb{R}^n)$ do not depend on matrix A. The matrix K is then obtained as the solution of an LQR problem. The results are reported in Table I.

n	m	p	q	2^m
5	5	2	27	32
5	5	3	30	32
6	6	2	44	64
6	6	3	52	64
6	6	4	60	64
7	7	2	61	128
7	7	3	74	128
7	7	4	88	128
7	7	5	120	128
7	7	6	128	128

TABLE I

CARDINALITY q OF $\mathscr{I}(\mathbb{R}^n)$ COMPARED WITH 2^m.

Notice that the cardinality q of the set $\mathscr{I}(\mathbb{R}^n)$ can be much smaller than 2^m, the number of the subsets of $\mathbb{N}_m$. Then, at the price of some required pre-computation, the complexity of the optimization problem leading to the desired λ-contractive ellipsoid can be consistently reduced.

VI. CONCLUSIONS

A characterization of invariance and contractiveness for saturated linear systems is presented. In particular, conditions for invariance and contractiveness of symmetric convex sets are determined. The results have been applied to characterize and compute invariant ellipsoids. Future research directions concern the particularization of the presented results to polytopes and the extension to more general nonlinearities, as generalized saturated functions, for instance, see [21].

REFERENCES

[1] T. Alamo, A. Cepeda, M. Fiacchini, and E. F. Camacho. Convex invariant sets for discrete–time Lur'e systems. *Automatica*, 45:1066–1071, 2009.

[2] T. Alamo, A. Cepeda, and D. Limon. Improved computation of ellipsoidal invariant sets for saturated control systems. In *44th IEEE Conference on Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC '05*, pages 6216–6221, dec. 2005.

[3] T. Alamo, A. Cepeda, D. Limon, and E. F. Camacho. A new concept of invariance for saturated systems. *Automatica*, 42:1515–1521, 2006.

[4] D. P. Bertsekas. Infinite-time reachability of state-space regions by using feedback control. *IEEE Transactions on Automatic Control*, 17:604–613, 1972.

[5] F. Blanchini and S. Miani. *Set-Theoretic Methods in Control*. Birkhäuser, 2008.

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] M. Cannon, V. Deshmukh, and B. Kouvaritakis. Nonlinear model predictive control with polytopic invariant sets. *Automatica*, 39:1487–1494, 2003.

[8] M. Fiacchini, T. Alamo, and E. F. Camacho. On the computation of convex robust control invariant sets for nonlinear systems. *Automatica*, 46(8):1334–1338, 2010.

[9] M. Fiacchini, S. Tarbouriech, and C. Prieur. Polytopic control invariant sets for continuous-time systems: A viability theory approach. In *Proceedings of the American Control Conference, 2011. ACC'11*, pages 1218 –1223, San Francisco, CA, june 2011.

[10] E. G. Gilbert and K. Tan. Linear systems with state and control constraints: The theory and application of maximal output admissible sets. *IEEE Transactions on Automatic Control*, 36:1008–1020, 1991.

[11] J. M. Gomes da Silva and S. Tarbouriech. Polyhedral regions of local stability for linear discrete-time systems with saturating controls. *IEEE Transactions on Automatic Control*, 44:2081–2085, 1999.

[12] J. M. Gomes da Silva and S. Tarbouriech. Local stabilization of discrete-time linear systems with saturating controls: An LMI-based approach. *IEEE Transactions on Automatic Control*, 46:119–125, 2001.

[13] P.-O. Gutman and P. Hagander. A new design of constrained controllers for linear systems. *IEEE Transactions on Automatic Control*, 30(1):22–33, 1985.

[14] T. Hu, Z. Lin, and B. M. Chen. Analysis and design for discrete-time linear systems subject to actuator saturation. *Systems & Control Letters*, 45(2):97 – 112, 2002.

[15] I. Kolmanovsky and E. G. Gilbert. Theory and computation of disturbance invariant sets for discrete-time linear systems. *Mathematical Problems in Engineering*, 4:317–367, 1998.

[16] L. Magni, G. De Nicolao, L. Magnani, and R. Scattolini. A stabilizing model-based predictive control algorithm for nonlinear systems. *Automatica*, 37:1351–1362, 2001.

[17] B. E. A. Milani. Piecewise-affine Lyapunov functions for discrete-time linear systems with saturating controls. *Automatica*, 38(12):2177–2184, 2002.

[18] S. V. Raković, E. C. Kerrigan, K. I. Kouramas, and D. Q. Mayne. Invariant approximations of the minimal robust positively invariant set. *IEEE Transactions on Automatic Control*, 50:406–410, 2005.

[19] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, USA, 1970.

[20] R. Schneider. *Convex bodies: The Brunn-Minkowski theory*, volume 44. Cambridge University Press, Cambridge, England, 1993.

[21] S. Tarbouriech, I. Queinnec, T. Alamo, M. Fiacchini, and E. F. Camacho. Ultimate bounded stability and stabilization of linear systems interconnected with generalized saturated functions. *Automatica*, 47(7):1473–1481, 2011.

[22] B. Zhou, W. X. Zheng, and G.-R. Duan. An improved treatment of saturation nonlinearity with its application to control of systems subject to nested saturation. *Automatica*, 47(2):306–315, 2011.