Analysis and design of discrete-time linear systems with nested actuator saturations

Bin Zhou
Center for Control Theory and Guidance Technology, Harbin Institute of Technology, P.O. Box 416, Harbin, 150001, China

Abstract

This paper is concerned with the analysis and design of discrete-time linear systems subject to nested saturation functions. By utilizing a new compact convex hull representation of the saturation nonlinearity, a linear matrix inequalities (LMIs) based condition is obtained for testing the local and global stability of the considered nonlinear system. The estimation of the domain of attraction and the design of feedback gains such that the estimation of the domain of attraction for the resulting closed-loop system is maximized are then converted into some LMIs based optimization problems. Compared with the existing results on the same problems, the proposed solutions are less conservative as more slack variables are introduced into the conditions. A couple of numerical examples are worked out to validate the effectiveness of the proposed approach.

1. Introduction

Saturation nonlinearity existing in every practical control system can cause not only performance degradation but also instability of the overall system [1,2]. The presence of saturation nonlinearity makes the overall systems inherently nonlinear [1,3,4]. Therefore, various control problems for systems containing saturation nonlinearity have attracted much attention over the past several decades (see, for example, [5–11]). It is now clear that, for a linear system with input saturation, semi-global stabilization or global stabilization can be achieved if and only if it is asymptotically null controllable by bounded controls (ANCBC) [1,4,12]. For ANCBC linear systems, semi-global stabilization by saturated linear feedback is even possible if there are delays in the actuators [13,14]. If a linear system with input saturation is not ANCBC, only local stabilization can be achieved. In this case, the problem of estimating the domain of attraction and the problem of designing feedback laws such that the closed-loop system has a large estimation of domain of attraction have attracted considerable attention in recent years (see [15–20] and the references cited there). Recent advances on control systems with input saturation can be found in, for example, [3,7,21] and the references therein.

In some applications, the actuator of the control system may be subject to nested saturations. One example is linear systems with both limited amplitude and rate actuators [22]. Another example is that, as pointed out in [23], if a linear system is subject to both sensor and actuator amplitude saturations, then the closed-loop system is subject to nested saturations if a dynamic output feedback is applied on the systems. Another source of nested saturation functions is the recursive design of saturated linear feedback control of linear systems (see, for example, [4,24]). More introduction on this class of nonlinear systems can be found in [3,23,25]. Because of its important applications in practice, some results have been reported in the literature. For continuous-time linear systems with nested saturation functions, linear matrix inequalities (LMIs) based approaches were proposed in [23,25] by using respectively the generalized sector bound characterization and convex hull characterization of saturation nonlinearity. The discrete-time counterpart of the results in [25] was established in [26]. However, some of the coefficient matrices in [25,26] are assumed to be diagonal, which is not always met in practice [23]. Very recently, with a new treatment of the saturation nonlinearity introduced in [27], the results in [25] were generalized in [2] by removing the restrictions on the coefficient matrices.

The aim of the present paper is to extend our work in [2] to a discrete-time setting, namely, we consider the analysis and design
of discrete-time linear systems with nested saturations. Differently from [26], we will not impose the condition that some of the coefficient matrices are diagonal on the considered systems. The extension of the results in [2] to the discrete-time case is not trivial since the terms of the increment of the Lyapunov function cannot be examined by directly using the methods in [27,28] or by the inequality-based treatment of saturation nonlinearity proposed in [2] because of a quadratic term of the saturation nonlinearity. To deal with this difficulty, we propose a new convex hull characterization of the saturation nonlinearity that can be implied by early literature [2,3,27,28]. This new convex hull representation can be seen as a generalization of the one proposed in [7] and used in [25,26]. LMIs based conditions are then established to test both the local stability and global stability of the closed-loop system. The problem of estimation of domain of attraction and the problem of designing feedback gains such that the domain of attraction is maximized are then converted into some LMIs based optimization problems. Similarly to the results in [2] and differently from the results given in [26], the proposed conditions contain more slack variables and are thus less conservative than the results in [26]. Some numerical examples are worked out to validate the effectiveness of the proposed results.

The remainder of this paper is organized as follows. The problem formulation and some preliminary results are introduced in Section 2. In Section 3, we will present the solutions to the analysis and design of discrete-time linear systems with nested saturation functions. A couple of numerical examples are given in Section 4 to show the effectiveness of the proposed approaches and Section 5 concludes the paper.

2. Problem formulation and preliminaries

In this paper, we consider a discrete-time linear system subject to nested saturation in its control input

$$x(k + 1) = Ax(k) + B \text{sat}(F_1 x(k)) + E_1 \text{sat}(F_2 x(k)) + \cdots + E_p \text{sat}(F_p x(k)),$$

where $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$ are system matrices, $F_i \in \mathbb{R}^{m \times n}, i \in \{1, 2, \ldots, p\}$ are feedback gains, $E_i \in \mathbb{R}^{m \times \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \·
side of (10). To see this, consider \( m = 3 \) and denote
\[
D_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},
\]
\[
D_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad D_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},
\]
\[
D_5 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad D_6 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},
\]
\[
D_7 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad D_8 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.
\]
and the \( i \)-th row of \( v_j \) as \( v_{ij}, j \in I [1, 8], i = 1, 2, 3 \). Then (10) can be written as
\[
\text{sat}(u) \in \text{co} \left\{ D_j u + e_i v : i \in I [1, 2^m] \right\}.
\] (11)

It follows that the variables \( v_{11}, v_{12}, v_{13}, v_{21}, v_{22}, v_{23}, v_{31}, v_{32}, v_{33} \) and \( v_{33} \) will not appear in (11). Hence the utilization of Lemma 1 can lead to a lower computational burden than that of Lemma 2, especially when \( m \) is relatively large.

The following corollary, which is originally proposed in [7], is a consequence of Lemma 1 or Lemma 2.

**Corollary 1.** Let \( m \geq 1 \) be a given integer and \( u, v \in \mathbb{R}^m \) be two given vectors. If \( \|v\|_\infty \leq 1 \), then
\[
\text{sat}(u) \in \text{co} \left\{ D_j u + e_i v : i \in I [1, 2^m] \right\}.
\] (12)

It is clear to see that the convex hull representation in (8) is less conservative than the convex hull representation in (12). Lemma 1 is also equivalent to Lemma 1.3 in [3]. However, compared with Lemma 1.3 in [3], Lemma 1 is in the same compact form as in Corollary 1. Moreover, if \( m = 1 \), then Lemma 1 is equivalent to Corollary 1. This is reasonable since it is shown in [19] that the convex hull representation in Corollary 1 leads to no conservatism at all in the estimation of domain of attraction for single input linear system (i.e., \( m = 1 \)) subject to saturated linear feedback.

### 3. Main results

The linearized system of (1) is given by
\[
x(k + 1) = (A + BF)x(k),
\] (13)
where \( F \) is determined by
\[
F = F_1 + E_2 F_2 + (E_2 E_3) F_3 + \cdots + \left( \prod_{i=2}^{p-1} E_i \right) F_p.
\] (14)

Similarly to [26], we should assume that \( A + BF \) is Schur stable. The we have the following results on testing the contractive invariance of an ellipsoid for the discrete-time nonlinear system (1).

**Theorem 1.** Let \( \mathbb{R}^m = m.2^m-1, i \in I [1, p] \). Then \( \mathbb{P}(P) \subseteq \mathcal{L}(H_i), i \in I [1, p], \) are satisfied. Here
\[
\mathbb{P} \triangleq \left\{ I \left[ 1, 2^m \right] \times I \left[ 1, 2^m \right] \times \cdots \times I \left[ 1, 2^m \right] \right\},
\] (16)
and \( f(i_1, i_2, \ldots, i_p) \) is defined as
\[
f(i_1, i_2, \ldots, i_p) = D_{i_1} F_1 + D_{i_2} H_1 + \sum_{k=1}^{p} \left( \prod_{j=1}^{k} D_{i_j} E_{j+1} \right)
\]
\[
\times \left( D_{i_{k+1}} F_{k+1} + D_{i_{k+1}} H_{k+1} \right).
\] (17)

**Proof.** For \( i = 1, 2, \ldots, p - 1 \), we denote
\[
u_i(k) = \text{sat}(F_1 x(k) + E_2 u_1(k)),
\] (19)
and \( \nu_p = \text{sat}(F_p x(k)) \). Then the discrete-time nonlinear system (1) can be written as
\[
x(k + 1) = Ax(k) + Bu_1(k).
\] (20)

Let \( v_i \in \mathbb{R}^2 \) be such that \( \|v_i\|_\infty \leq 1 \). Then, by Lemma 1, we know that
\[
u_i(k) = \text{co} \left\{ D_{i_1} F_1 x(k) + D_{i_2} u_2(k) + D_{i_3} u_3(k) \right\} : i_1 \in I [1, 2^m] \}
\]
\[
\times u_2(k) : i_2 \in I [1, 2^m] \right\}.
\] (22)

Notice the fact that, if
\[
a \in \text{co} \{ \nu_i : i \in I [1, J] \}, \quad b \in \text{co} \{ \nu_j : j \in I [1, J] \},
\] (23)
where \( V_i, \in I [1, J] \), are some matrices with appropriate dimensions, \( I \geq 1 \) and \( J \geq 1 \) are some integers, and \( \phi_i, \phi_j \) are some vectors, then
\[
a \in \text{co} \{ \nu_i + \nu_j : (i, j) \in I [1, J] \times I [1, J] \}.
\] (24)

Then we get from (21) and (22) that
\[
u_i(k) = \text{co} \left\{ D_{i_1} F_1 x(k) + D_{i_2} u_2(k) + D_{i_3} u_3(k) \right\} : i_1, i_2 \in I [1, 2^m] \}
\]
\[
\times u_3(k) : i_2 \in I [1, 2^m] \right\}.
\] (25)

By repeating the above process we finally have
\[
u_i(k) = \text{co} \left\{ D_{i_1} F_1 x(k) + D_{i_2} u_2(k) + \sum_{k=1}^{p-1} \left( \prod_{j=1}^{k} D_{i_j} E_{j+1} \right)
\]
\[
\times \left( D_{i_{k+1}} F_{k+1} x(k) + D_{i_{k+1}} u_{k+1}(k) \right) : (i_1, i_2, \ldots, i_p) \in \Pi \right\},
\] (26)
where \( v_i \in \mathbb{R}^m, i \in I [1, p] \), are such that \( \|v_i\|_\infty \leq 1 \). If we set \( v_i = H_i x(k), i \in I [1, p] \), then we can write (26) as
\[
u_i(k) = \text{co} \{ f(i_1, i_2, \ldots, i_p) x(k) : (i_1, i_2, \ldots, i_p) \in \Pi \}.
\] (27)
Consequently, for any $x(k) \in \mathcal{Z}(H_i)$, $i \in I[1, p]$, the discrete-time nonlinear system (20) can be written as
\[
x(k+1) = \text{co}(A + Bf_i(i_1, i_2, \ldots, i_p)) \times x(k) : (i_1, i_2, \ldots, i_p) \in \Pi.
\] (28)

As $V(x(k)) = x^T(k)Px(k)$ is a convex function, it follows from (15) and (28) that the time-shift of $V(x(k))$ along the trajectories of the discrete-time nonlinear system (20) satisfies
\[
\nabla V(x(k)) \triangleq x^T(k + 1)Px(k + 1) - x^T(k)Px(k)
\leq \max_{(i_1, i_2, \ldots, i_p) \in \Pi} \left\{ x^T(k) \left( A + Bf_i(i_1, i_2, \ldots, i_p) \right)^T \right. \\
\left. \times P \left( A + Bf_i(i_1, i_2, \ldots, i_p) \right) - P \right\} x(k)
\leq 0, \ \forall x(k) \in \mathcal{E}(P) \subseteq \mathcal{Z}(H_i) \setminus \{0\}.
\] (29)

The proof is finished. ■

**Remark 3.** Theorem 1 can be regarded as the discrete-time version of our recent results on the following continuous-time linear systems subject to nested saturation functions [2]
\[
\dot{x}(t) = Ax(t) + B \text{sat}(F_1x(t)) + E_2 \text{sat}(F_2x(t))
+ E_3 \text{sat}(F_3x(t) + \cdots + E_p \text{sat}(F_px(t)) \cdots)),
\] (30)
in which the coefficient matrices are as defined as for system (1). Compared with the results given in [2], the results in Theorem 1 are in compact forms due to the usage of the convex hull representation of the saturation nonlinearity given in Lemma 1. In fact, with the help of Lemma 1, the main inequality in [2] can be simplified as
\[
(A + Bf_i(i_1, i_2, \ldots, i_p))^T P
+ P \left( A + Bf_i(i_1, i_2, \ldots, i_p) \right) < 0, \ \forall (i_1, i_2, \ldots, i_p) \in \Pi.
\] (31)

The following corollary can be obtained immediately by setting $H_i = 0$, $i \in I[1, p]$ in (15).

**Corollary 2.** The discrete-time nonlinear system (1) is globally exponentially stable if there exists a matrix $P > 0$ such that

\[
(A + Bf_i(i_1, i_2, \ldots, i_p))^T P \left( A + Bf_i(i_1, i_2, \ldots, i_p) \right)
- P < 0, \ \forall (i_1, i_2, \ldots, i_p) \in \Pi.
\] (32)

where $f_i(i_1, i_2, \ldots, i_p)$ is defined as
\[
f_i(i_1, i_2, \ldots, i_p) \triangleq D_{F_i} F_1 + \sum_{k=1}^{p-1} \left( \prod_{j=1}^{k} D_{E_{j+1}} \right) D_{h_k} F_{k+1}.
\] (33)

If the convex hull representation of the saturation nonlinearity given in Corollary 1 is used in the proof of Theorem 1, the following result can be obtained.

**Corollary 3.** The ellipsoid $\mathcal{E}(P) = \{x : x^T P x \leq 1\} \subseteq \mathcal{Z}(H_i)$, $i \in I[1, p]$, such that
\[
(A + Bf_i(i_1, i_2, \ldots, i_p))^T P \left( A + Bf_i(i_1, i_2, \ldots, i_p) \right)
- P < 0, \ \forall (i_1, i_2, \ldots, i_p) \in \Pi.
\] (34)

and $\mathcal{E}(P) \subseteq \mathcal{Z}(H_i)$, $i \in I[1, p]$, are satisfied. Here $f_i(i_1, i_2, \ldots, i_p)$ is obtained by replacing $D_{\mathcal{E}_i}^{-1}$ with $D_{h_i}^{-1}$ in $f_i(i_1, i_2, \ldots, i_p)$ defined in (17).

Since the conditions in Theorem 1 contain more free variables than those in Corollary 3, Theorem 1 is claimed to be less conservative than Corollary 3, which will be validated by our numerical examples given in Section 4.

We next use Theorem 1 to estimate the domain of attraction for the discrete-time nonlinear system (1), namely, to solve the remaining part of Item 1 in Problem 1. Similarly to the argument given in [7], the largeness of the ellipsoid $\mathcal{E}(P)$ with respect to a shape reference set $\mathcal{X}_k \subseteq \mathbb{R}^n$ can be measured by the scalar $\alpha$ which is the maximal number such that $\alpha \mathcal{X}_k \subseteq \mathcal{E}(P)$ is satisfied, where $\alpha \mathcal{X}_k \triangleq \{x \in \mathcal{X}_k \}$.

where $x_i \in \mathbb{R}^r, i \in I[1, l]$ are some given vectors, then $\alpha \mathcal{X}_k \subseteq \mathcal{E}(P)$ is equivalent to [2]
\[
\begin{bmatrix} \gamma x_j^* \\ \lambda_n^* \end{bmatrix} \geq 0, \ j \in I[1, l],
\] (36)
where $\gamma = 1/\alpha^2$. If the reference set is chosen as the ellipsoid, namely, $\mathcal{X}_k = \mathcal{E}(R)$, where $R > 0$, then $\alpha \mathcal{X}_k \subseteq \mathcal{E}(P)$ is equivalent to [2]
\[
\begin{bmatrix} \gamma R^* \\ \lambda_n^* \end{bmatrix} \geq 0,
\] (37)
where $\gamma = 1/\alpha^2$. Then the best estimate of the domain of attraction for the discrete-time nonlinear system (1) can be formulated as the following constrained optimization problem
\[
\sup_{P > 0, \text{sym}, \text{H} \in I[1, p]} \alpha \mathcal{X}_k \subseteq \mathcal{E}(P)
\] (38)
s.t. (a) $\alpha \mathcal{X}_k \subseteq \mathcal{E}(P)$,
(b) $(A + Bf_i(i_1, i_2, \ldots, i_p))^T P \left( A + Bf_i(i_1, i_2, \ldots, i_p) \right)$
\[- P < 0, \ \forall (i_1, i_2, \ldots, i_p) \in \Pi.
\] (c) $\mathcal{E}(P) \subseteq \mathcal{Z}(H_i), i \in I[1, p].$

By denoting $\gamma = 1/\alpha^2$, $Q = P^{-1}$ and $G_i = H_i P^{-1}, i \in I[1, p]$, and applying the Schur complement, the optimization problem (38) can be transformed into the following one
\[
\inf_{Q > 0, \text{sym}, \text{H} \in I[1, p]} \gamma \mathcal{X}_k \subseteq \mathcal{E}(P)
\] (39)
s.t. (a) $\begin{bmatrix} \gamma R^* \\ \lambda_n^* \end{bmatrix} \geq 0, \text{ or } \begin{bmatrix} \gamma x_j^* \\ \lambda_n^* \end{bmatrix} \geq 0, j \in I[1, l],
\] (b) $\begin{bmatrix} AQ + Bf_i(i_1, i_2, \ldots, i_p) \\ Q \end{bmatrix}^* \geq 0, \ \forall (i_1, i_2, \ldots, i_p) \in \Pi,
\] (c) $\begin{bmatrix} \gamma G_{ij}^* \\ \lambda_n^* \end{bmatrix} \geq 0, \ j \in I[1, p], i \in I[1, l].$

where $G_{ij}$ is the $j$-th row of $G_i$, and, for $(i_1, i_2, \ldots, i_p) \in \Pi, \mathcal{F}(i_1, i_2, \ldots, i_p)$ is defined as
\[
\mathcal{F}(i_1, i_2, \ldots, i_p) = D_{\mathcal{E}_i} F_1 + \sum_{k=1}^{p-1} \left( \prod_{j=1}^{k} D_{E_{j+1}} \right) D_{h_k} F_{k+1}.
\] (40)

The resulting best estimation of the domain of attraction can be recovered as $\mathcal{E}(Q^{-1})$. The corresponding approach for estimating
the domain of attraction for the discrete-time nonlinear system (1) by using Corollary 3 can be obtained as well and the details are omitted for brevity.

Similarly, solutions to Item 2 of Problem 1 by using Theorem 1 can be obtained by solving the following LMI based optimization problem

\[
\begin{align*}
\inf_{Q \succ 0, Z, L_i, i \in \{1, p\}} & \gamma \\
\text{s.t.} & (a) \text{ and } (c) \text{ in (39)} \\
& (b') A^TQ + QB \succ 0, \\
& \forall (i_1, i_2, \ldots, i_p) \in \mathcal{P}, \\
& \text{where, for } (i_1, i_2, \ldots, i_p) \in \mathcal{P}, \mathcal{F}(i_1, i_2, \ldots, i_p) \text{ is defined as} \\
& \mathcal{F}(i_1, i_2, \ldots, i_p) = D_{i_1}Z_1 + SD_{i_1}G_1 + \sum_{k=1}^{p-1} \left( \prod_{j=1}^{k} D_{i_j}E_{j+1} \right) \\
& \times \left( D_{i_{k+1}}Z_{k+1} + SD_{i_{k+1}}G_{k+1} \right). 
\end{align*}
\]

The resulting best estimation of the domain of attraction can be recovered as \( \delta (Q^{-1}) \) and the best feedback gains are recovered as \( F_i = ZQ^{-1}, i \in \{1, p\} \). Corresponding approaches by using Corollary 3 can be stated as well and will not be included here.

Remark 4. Numerical experience show that the feedback gains and/or the eigenvalues of the linearized closed-loop system are badly scaled if the feedback gains \( F_i \) are jointly optimized (a similar phenomenon was discovered in [25]). To avoid this problem, we may add an additional requirement on the eigenvalues of the closed-loop system, for example, \( |\lambda_i(A + BF)| \leq \delta \) where \( \delta < 1 \) is a prescribed scalar. Such a requirement can be easily transformed into the following LMI based constraint

\[
\begin{align*}
\begin{bmatrix}
\frac{1}{\delta} A(Q + B\mathcal{F}_c) & Q \\
0 & \text{\ast}
\end{bmatrix} & > 0,
\end{align*}
\]

where, for \( (i_1, i_2, \ldots, i_p) \in \mathcal{P}, \mathcal{F}_c \) is defined as

\[
\mathcal{F}_c = Z_1 + E_2Z_2 + (E_2E_3)Z_3 + \cdots + \left( \prod_{i=2}^{p} E_i \right) Z_p.
\]

Remark 5. For any set \( \mathcal{K} \subseteq \{1, p\} \), both Theorem 1 and Corollary 3 can be adopted to design \( F_i, i \in \mathcal{K} \), such that the resulting estimation of domain of attraction is maximized with respect to the reference set \( \mathcal{X}_K \). The details are not included here for the sake of brevity.

Remark 6. Similarly to the above development, to test the global stability of system (1), we need only to solve the feasibility of the following set of LMIs

\[
\begin{align*}
\begin{bmatrix}
A^TQ + B\mathcal{F}_c & Q \\
0 & \text{\ast}
\end{bmatrix} & > 0, \\
\forall (i_1, i_2, \ldots, i_p) \in \mathcal{P},
\end{align*}
\]

where, for \( (i_1, i_2, \ldots, i_p) \in \mathcal{P}, \mathcal{F}_c \) is defined as

\[
\mathcal{F}_c = D_{i_1}F_1Q + \sum_{k=1}^{p-1} \left( \prod_{j=1}^{k} D_{i_j}E_{j+1} \right) \\
\times \left( D_{i_{k+1}}Z_{k+1} + SD_{i_{k+1}}G_{k+1} \right).
\]

Notice that a necessary condition for the feasibility of (45) is that both \( A + BF \), where \( F \) is defined in (14), are Schur stable, which is reasonable. Moreover, the problem of global stabilizing controller design can be solved in a similar way and is omitted for brevity.

The improvement of the proposed approach is achieved by introducing more additional decision variables. Let \( M \) and \( N \) denote respectively the total row size and the total number of scalar decision variables of an LMI system. For the comparison purpose, we list the values of \( M \) and \( N \) associated with different approaches in Table 1. For simplicity, we have assumed that \( m_i = m, i \in \{1, p\} \).

<table>
<thead>
<tr>
<th>Approach</th>
<th>( M ) (total row size of LMIs)</th>
<th>( N ) (total number of decision variables)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theorem 1 in [26]</td>
<td>( n(q + 1)^m + (n + 1)mq )</td>
<td>( mnq + n(n + 1) )</td>
</tr>
<tr>
<td>Corollary 3</td>
<td>( n2^m + (n + 1)mq )</td>
<td>( mnq + n(n + 1) )</td>
</tr>
<tr>
<td>Theorem 1</td>
<td>( n2^m + (n + 1)mq2^{p-1} )</td>
<td>( mnq2^{p-1} + n(n + 1) )</td>
</tr>
</tbody>
</table>

Table 1. The total row size of LMI and number of decision variables in different approaches.
and $E_2$ is a diagonal matrix given by
\[
E_2 = \begin{bmatrix} 13 & 0 \\ 0 & -1 \end{bmatrix},
\]
which is such that $F = F_1 + E_2 F_2$.

By respectively applying Theorem 1 in [26], Theorem 1 and Corollary 3 in this paper and by choosing $R = I_2$, the resulting best estimate of the domain of attraction can be obtained as $\delta(P)$, in which the positive definite matrix $P$ is respectively given by
\[
p_{\text{Theorem 1 in [26]}} = \begin{bmatrix} 2.0439 & 0.98296 \\ 0.98296 & 3.2407 \end{bmatrix},
p_{\text{Theorem 1}} = \begin{bmatrix} 1.7413 & 1.1527 \\ 1.1527 & 2.7932 \end{bmatrix}. \tag{51}
\]

We notice that Theorem 1 in [26] and Corollary 3 in this paper led to the same results. These ellipsoids are recorded in Fig. 1 from which it follows that Theorem 1 leads to the largest estimation of the domain of attraction. The computation time required by Theorem 1 in [26], Theorem 1 and Corollary 3 (Lenovo T430, Intel Core i5-3320M CPU @ 2.60 GHz) are respectively, 0.109 s, 0.203 s, and 0.193 s, which again indicates a trade-off between the computational complexity and the performance improvement.

If we set $F_1$ to be unknown but such that $F = F_1 + E_2 F_2$ is prescribed as in (48), we get the same results as the above. This is because $F_1$ and $F_2$ are related linearly if $F$ is prescribed.

**Example 2.** In this example, we consider a discrete-time nonlinear system in the form of (1) in which $E_i$ are not diagonal. For simplicity, we let $p = 2$, $(A, B)$ be given by (47), and $F$ be given by (48). Moreover, we let
\[
E_2 = \begin{bmatrix} 6 & 1 \\ 5 & -2 \\ 13 & 3 \\ 10 & 2 \end{bmatrix}. \tag{54}
\]

In this case, the approach given in [26] is no longer applicable. If the matrix $F_1$ is fixed as in (49), then, by respectively using Theorem 1 and Corollary 3 in this paper, the resulting optimal feedback gains $F_1$ are respectively obtained as
\[
p_{\text{Theorem 1 in [26]}} = \begin{bmatrix} 2.4129 & 2.9244 \\ 0.71738 & 2.6320 \end{bmatrix},
p_{\text{Theorem 1}} = \begin{bmatrix} 2.6356 & 3.0330 \\ 0.75380 & 2.7856 \end{bmatrix}. \tag{52}
\]

and the positive definite matrices $P$ in the best estimation of the domain of attraction $\delta(P)$ are respectively obtained by
\[
p_{\text{Theorem 1 in [26]}} = \begin{bmatrix} 2.0467 & 0.9819 \\ 0.9819 & 3.2407 \end{bmatrix},
p_{\text{Theorem 1}} = \begin{bmatrix} 1.4678 & 1.1265 \\ 1.1265 & 2.8018 \end{bmatrix}. \tag{53}
\]

We again notice that Theorem 1 in [26] and Corollary 3 in this paper lead to the same results. These ellipsoids are plotted in Fig. 2.

From this figure we again see that Theorem 1 proposed in this paper leads to the largest estimation of the domain of attraction, which validates the effectiveness of the proposed approach. The computation time required by Theorem 1 in [26], Theorem 1 and Corollary 3 (Lenovo T430, Intel Core i5-3320M CPU @ 2.60 GHz) are respectively, 0.129 s, 0.242 s, and 0.193 s, which again indicates a trade-off between the computational complexity and the performance improvement.

We finally emphasize that, if we set $F_2$ to be unknown and such that $F = F_1 + E_2 F_2$ is prescribed as in (48), we get the same results as the above. This is because $F_1$ and $F_2$ are related linearly if $F$ is prescribed.
Corollary 3. The computation times required by Theorem 1 and Corollary 3 (Lenovo T430, Intel Core i5-3320M CPU @ 2.60 GHz) are, respectively, 0.351 s and 0.292 s, which also indicates a trade-off between the computational complexity and the performance improvement.

If the feedback gains $F_1$ and $F_2$ are jointly optimized, then by using respectively Theorem 1 and Corollary 3 in this paper and by taking Remark 4, where $\delta = 0.5$, into consideration, the best feedback gains $F_i, i = 1, 2,$ are respectively obtained as

$$
P_{\text{Theorem 1}} = \begin{bmatrix} 2.4109 & 3.0049 \\ 1.1218 & 2.9690 \end{bmatrix}, \quad P_{\text{Corollary 3}} = \begin{bmatrix} 2.4475 & 3.3983 \\ 1.1479 & 2.7622 \end{bmatrix}. \tag{56}
$$

and the positive definite matrices $P$ in the best estimation of the domain of attraction $\delta(P)$ are respectively obtained by

$$
P_{\text{Theorem 1}} = \begin{bmatrix} 1.4672 & 1.1265 \\ 1.1265 & 2.8016 \end{bmatrix}, \quad P_{\text{Corollary 3}} = \begin{bmatrix} 2.0444 & 0.9834 \\ 0.9834 & 3.2402 \end{bmatrix}. \tag{57}
$$

These ellipsoids are also recorded in Fig. 3. It follows that Theorem 1 again leads to a much less conservative result than Corollary 3. The computation times required by Theorem 1 and Corollary 3 (Lenovo T430, Intel Core i5-3320M CPU @ 2.60 GHz) are, respectively, 0.291 s and 0.212 s, which also indicates a trade-off between the computational complexity and the performance improvement.

Example 3. We consider a nonlinear discrete-time system in the form of (1) with $p = 2$,

$$
A = \begin{bmatrix} -0.4 & 0.2 & 0.4 \\ 1.5 & 0.5 & -0.4 \\ 0.2 & -0.3 & 0.2 \end{bmatrix}, \quad B = \begin{bmatrix} 0.3 & 0 \\ -0.1 & 1 \end{bmatrix}, \tag{61}
$$

$$
F_1 = \begin{bmatrix} -0.9 & -0.2 & -0.2 \\ -1.1 & 0.4 & 0.2 \end{bmatrix}, \quad F_2 = \begin{bmatrix} -0.4 & 0.3 & 0.3 \\ 1.1 & -0.5 & -0.5 \end{bmatrix}, \tag{62}
$$

and a non-diagonal matrix

$$
E_2 = \begin{bmatrix} -1 & 0 \\ 0.5 & 0.8 \end{bmatrix}. \tag{63}
$$

Then by solving the LMIs in (45) we obtain the following feasible positive definite solution

$$
P = \begin{bmatrix} 0.2598 & 0.041 & -0.0441 \\ 0.041 & 0.0702 & -0.0002 \\ -0.0441 & -0.0002 & 0.0753 \end{bmatrix}. \tag{64}
$$

It follows from Corollary 2 that this nonlinear discrete-time system is globally exponentially stable. For the purpose of illustration, the state trajectories of this nonlinear discrete-time system with the initial condition $x(0) = [-7.5, -8]^T \times 10^2$ are recorded in Fig. 4 from which we clearly see that this system is asymptotically stable.

5. Conclusion

This paper has proposed new solutions to the problems of analysis and design of discrete-time linear systems subject to nested saturations. By using a compact form of convex hull representation of the saturation nonlinearity, sufficient conditions testing both the local stability and global stability of the considered nonlinear systems are obtained in terms of linear matrix inequalities (LMIs).
The analysis and design problems are then transformed into some LMI-based optimization problems. Compared with the existing result on the same problems, the new results are less conservative as the proposed optimization problems contain more free decision parameters. Numerical examples have validated the effectiveness of the proposed approaches.

Acknowledgments

The author would like to thank Professor Zongli Lin for his encouragement in the revising of this paper. The author would also like to thank the Associate Editor and the anonymous reviewers for their valuable suggestions which have helped to improve the quality of this paper considerably.

Appendix. Proof of Lemma 1

We use the idea found in [2] to prove this lemma. To this end, we first introduce some preliminary results. We divide the set \( D_m \) into two sets \( D_m^+ \) and \( D_m^- \) so that

\[
(a). D_m^+ \bigcup D_m^- = D_m, \quad (b). D \in D_m^+ \iff D^+ \in D_m^-.
\]

It follows that both \( D_m^+ \) and \( D_m^- \) have \( 2^{m-1} \) elements and \( D_m \cap D_m = \varnothing \). Clearly, such a decomposition is not unique (actually, there are \( 2^m \) kinds of decomposition). For notational simplicity, we label the elements in \( D_m \) as \( D_i, i \in [1, 2^{m-1}] \). Thus the elements in \( D_m \) can be labeled as \( D_i, i \in \{1, 2^{m-1}\} \).

**Lemma 3 ([27]).** Let \( a, b \in \mathbb{R} \) be two arbitrary scalars. Then

\[
\text{asat}(b) \leq \max \{ab, -|a|\}.
\]

**Lemma 4.** Let \( D_m \) be such that the decomposition (65) holds and the elements in \( D_m \) be labeled as \( D_i, i \in [1, 2^{m-1}] \). Let \( u \in \mathbb{R}^m \) and \( v \in \mathbb{R}^{2^m} \) be two given vectors and such that \( \|v\|_{\infty} \leq 1 \). Then, for any vector \( \varphi \in \mathbb{R}^m \), there holds

\[
\langle \varphi, \text{sat}(u) \rangle \leq \max_{k \in [1, 2^{m-1}]} \left\{ \langle \varphi, D_k u + (2^{m-1-k} \otimes D_k) v \rangle \right\},
\]

\[
\langle \varphi, D_k^T u + (2^{m-1-k} \otimes D_k) v \rangle,
\]

where \( (x, y) = x^T y \) denotes the inner product of the vectors \( x \) and \( y \).

**Proof.** Denote the \( k \)-th element of \( u \) by \( u_k \) and the \( k \)-th column of \( \varphi \) by \( \varphi_k \). Let

\[
|\varphi| = \left[ |\varphi_1| \quad |\varphi_2| \ldots |\varphi_m| \right]^T.
\]

Then by using Lemma 3 and denoting \( I_m \in \mathbb{R}^m \) a vector whose elements are 1, we get

\[
\langle \varphi, \text{sat}(u) \rangle = \sum_{k=1}^m \varphi_k \text{sat}(u_k)
\]

\[
\leq \max_{k=1}^m \{\varphi_k u_k, -|\varphi_k|\}
\]

\[
= \max_{k \in [1, 2^{m-1}]} \left\{ \langle \varphi, D_k u \rangle - |\varphi| \|D_k I_m\| \right\}
\]

\[
= \max_{k \in [1, 2^{m-1}]} \left\{ \langle \varphi, D_k u \rangle - |\varphi| \|D_k I_m\| \right\}, \quad (70)
\]

in which (69) can be verified by a simple and direct calculation. Assume that \( D_k = \text{diag}(\alpha_j, j \in [1, m]) \) has \( \varphi_k \) nonzero diagonal elements \( \alpha_j = \alpha_{j_1} = \cdots = \alpha_{j_m} = 1 \), and \( D_k = \text{diag}(\beta_j, j \in [1, m]) \) has \( \psi_k \) nonzero diagonal elements \( \beta_j = \beta_{j_1} = \cdots = \beta_{j_m} = 1 \). It is clear that \( \psi_k + \phi_k = m \) and

\[
\{j_1, j_2, \ldots, j_m\} \bigcup \{j_1, j_2, \ldots, j_m\} = [1, m].
\]

Observe that, for two arbitrary scalars \( u, v \in \mathbb{R} \), if \( |v| \leq 1 \), then

\[
-|u| \leq -|uv| \leq uv.
\]

With this, for any \( v_k \in \mathbb{R} \) such that \( |v_k| \leq 1, k \in [1, 2^{m-1}], j \in [1, m] \), we have

\[
-|\varphi| D_k^T I_m = -\sum_{k=1}^m |\varphi_k| v_k^T I_m
\]

\[
-|\varphi| D_k I_m = -\sum_{k=1}^m |\varphi_k| v_k I_m
\]

Now denote the vector \( v \) as

\[
v = \left[v_1, v_2, \ldots, v_{2^{m-1}} \right]^T,
\]

\[
v_k = \left[v_{k,1}, v_{k,2}, \ldots, v_{k,m} \right]^T,
\]

which is such that \( \|v\|_{\infty} \leq 1 \) by definition. Then substituting (73)–(74) into (70) gives

\[
\langle \varphi, \text{sat}(u) \rangle \leq \max_{k \in [1, 2^{m-1}]} \left\{ |\varphi| D_k^T u + \sum_{k=1}^m |\varphi_k| v_k, |\varphi| D_k u + \sum_{k=1}^m |\varphi_k| v_k \right\}.
\]

\[
= \max_{k \in [1, 2^{m-1}]} \left\{ |\varphi| D_k^T u + \varphi \phi_k v_k + \phi \phi_k D_k u + \varphi \phi_k D_k v_k \right\}
\]

\[
= \max_{k \in [1, 2^{m-1}]} \left\{ |\varphi| D_k^T u + \varphi \phi_k (e_{2^{m-1-k}} \otimes D_k) v, \varphi \phi_k D_k u + \varphi \phi_k (e_{2^{m-1-k}} \otimes D_k) v \right\},
\]

which is equivalent to (67). This completes the proof. ■

**Lemma 5.** (Theorem 1.2.4 in [30]). Let \( \mathcal{X} \subset \mathbb{R}^n \) be a convex set. Then there is a vector \( x \in \mathbb{R}^n \) such that \( x \notin \mathcal{X} \) if and only if there exists a vector \( \alpha \in \mathbb{R}^n \) such that

\[
\langle \alpha, x \rangle > \langle \alpha, k \rangle, \quad \forall k \in \mathcal{X}.
\]

**Lemma 6.** Let \( p \) and \( m \) be two given integers. Assume that \( \eta \in \mathbb{R}^m \) and \( h_i \in \mathbb{R}^p, i \in [1, p] \), are some given vectors. Then

\[
\eta \in \text{co} \{h_i : i \in [1, p]\},
\]

if and only if, for any \( v \in \mathbb{R}^m \),

\[
(v, \eta) \leq \max_{i \in [1, p]} \{v, h_i\}.
\]

**Proof.** We first show that if (78) is satisfied, then (79) is true for any \( v \in \mathbb{R}^m \). By definition, there exist scalars \( \lambda_i \geq 0, i \in [1, p] \), such that

\[
\eta = \sum_{i=1}^p \lambda_i h_i, \quad \sum_{i=1}^p \lambda_i = 1.
\]

Then, for any \( v \in \mathbb{R}^m \), we can compute

\[
(v, \eta) = \sum_{i=1}^p \lambda_i (v, h_i)
\]

\[
\leq \sum_{i=1}^p \lambda_i \left( \max_{i \in [1, p]} \{v, h_i\} \right) = \max_{i \in [1, p]} \{v, h_i\}.
\]
We next show that, if (79) is true for any \( v \in \mathbb{R}^n \), then (78) must hold true. We show this by contradiction. Assume that \( \eta \notin \text{co}\{h_i : i \in [1, p]\} \). Then by Lemma 5, there exists a vector \( \alpha \in \mathbb{R}^n \) such that
\[
(\alpha, \eta) > (\alpha, h_i), \quad \forall h_i \in \text{co}\{h_i : i \in [1, p]\},
\]
which implies that \( (\alpha, \eta) > (\alpha, h_i) \), \( \forall i \in [1, p] \), or equivalently, \( (\alpha, \eta) > \max_{i\in[1,p]}(\alpha, h_i) \).

This contradicts with (79). Therefore we have \( \eta \notin \text{co}\{h_i : i \in [1, p]\} \) and the proof is finished. ■

We are now able to prove Lemma 1. By combining Lemmas 4 and 6, we get
\[
\text{sat}(u) \in \text{co}\left\{ D_k u + \left( e_{2^m-1-k} \otimes D_k \right) v_k, D_k u + \left( e_{2^m-1-k} \otimes D_k \right) v_k \in I_{[1, 2^{m-1}]} \right\},
\]
where \( D_k \) is decomposed such that (65) is satisfied, and we define \( D_{2^m-1-k} \) as \( D_k \in I_{[1, 2^{m-1}]} \), namely, we have labeled the elements in \( \phi_{2^m} \) as \( D_k \). Consequently, by the definition of the function \( f_m(k) \) in (6), we have
\[
f_m(k) = k + f_m(k + 2^{m-1}), \quad k \in I_{[1, 2^{m-1}]}.
\]
With this, we get from (84) that
\[
\text{sat}(u) \in \text{co}\left\{ D_k u + \left( e_{2^m-1-k} \otimes D_k \right) v_k, D_{2^m-1-k} u + \left( e_{2^m-1-f_m(k + 2^{m-1})} \otimes D_{2^m-1-k} \right) v_k \in I_{[1, 2^{m-1}]} \right\},
\]
which is just (8). This completes the proof.

References


