Preamble. This is a reprint of the article:

M. Schulze Darup and M. Mönnigmann. Positive invariance tests with efficient affine inclusions. In *Proc. of the 18th IFAC World Congress*, pp. 11116–11120, 2011.

The digital object identifier (DOI) of the original article is:

10.3182/20110828-6-IT-1002.03361

Positive invariance tests with efficient affine inclusions

Moritz Schulze Darup[†] *and* M. Mönnigmann[†]

Abstract

We analyze the computational complexity of several methods for automatically searching positive invariant (p.i.) sets of nonlinear autonomous continuous and discrete time systems. We show that p.i. detection can be improved upon by considering traits of the equations of the dynamical system such as monotonicity and convexity. Furthermore, we show that these traits can be taken into account by automated methods that apply to large system classes.

Keywords. positive invariance, domain of attraction, Lyapunov function, affine inclusions, interval matrices, interval arithmetics.

1 Introduction

Positively invariant (p.i.) sets play an important role in various problems and applications of control theory [1]. The present paper deals with methods for establishing p.i. on ellipsoids around a locally stable equilibrium of a nonlinear autonomous continuous time system

$$\dot{x}(t) = f(x(t)), \tag{1}$$

with appropriate initial conditions, where $f : \mathcal{X} \to \mathbb{R}^n$ is assumed to be a continuously differentiable function on an open $\mathcal{X} \subset \mathbb{R}^n$. The discussed methods can be extended to discrete time systems as briefly discussed in Sect. 5. Following a classical idea (see e.g. [3]), we split f into its affine and nonlinear contributions, and subsequently bound the nonlinear contributions with piecewise affine inclusions. In particular we are interested in computational approaches to calculating inclusions that are computationally efficient and that can be carried out automatically for any member of the system class (1), i.e. without human intervention and insight into special structures of the particular system at

[†] M. Schulze Darup and M. Mönnigmann are with Automatic Control and Systems Theory, Department of Mechanical Engineering, Ruhr-Universität Bochum, 44801 Bochum, Germany. E-mail: moritz.schulzedarup@rub.de.

hand. Specifically, we are interested in enlarging the set for which p.i. can be established, whenever such an enlargement comes at a reasonable computational effort. We stress that the treated computational methods in general provide conservative estimates of p.i. sets, which may, nevertheless, be practically relevant. P.i. sets are of practical importance in establishing the stability of model predictive control, for example. In this context an automated procedure for the proof of p.i. is useful even if the identified set is not the largest p.i. set.

From a technical point of view codelists and interval arithmetics are instrumental to the approaches discussed here. The presented methods are related to a second order approach that is based on bounds of the eigenvalues of Hessian matrices [7]. In the present paper, however, eigenvalue bounds for Hessian matrices are not used.

The paper is organized as follows. In Sect. 2 we introduce two variants for overestimating nonlinear, factorable functions by affine inclusions. Sections 3, 3.2, and 4 state sufficient conditions for p.i. of affine inclusions on ellipsoids, the detection of positive definiteness (p.d.) of matrices, and a simple algorithm for p.i. detection by p.d. detection, respectively. Section 5 briefly extents the proposed approach to discrete time systems and gives an example for a continuous and a discrete time system. Conclusions are given in Sect. 6.

2 Calculating affine inclusions for nonlinear functions

A continuously differentiable function $f_i : \mathcal{X} \to \mathbb{R}$ can be expanded into a Taylor series

$$f_i(x) = f_i(\breve{x}) + \sum_{j=1}^{n} a_{ij}\,(x_j - \breve{x}_j) + g_i(x - \breve{x}), \tag{2}$$

where $a_{ij} = \frac{\partial f_i(x)}{\partial x_j}\big|_{\breve{x}}$ and g_i has no affine contributions. Without restriction we assume $\breve{x} = 0$ is an equilibrium of (1). Since $f_i(\breve{x}) = 0$ the linearization of (1) reads

$$\dot{x}(t) = A\,x(t), \tag{3}$$

where A is the Jacobian with elements a_{ij}.

We anticipate that it will be crucial to obtain bounds on $g_i(x)$ for all $x \in \mathcal{B}$, where $\mathcal{B} \subset \mathcal{X}$ is a compact hyperrectangle that contains the equilibrium $\breve{x} = 0$. The desired bounds can be written in any of the following three forms.

$$\begin{aligned} g_i(x) \in \sum_{j=1}^{n} [w_{ij}]\,x_j &= \sum_{j=1}^{n} [\underline{\alpha}_{ij}(x_j), \overline{\alpha}_{ij}(x_j)] \\ &= [\underline{\beta}_i(x), \overline{\beta}_i(x)] \end{aligned} \tag{4}$$

with

$$[\underline{\alpha}_{ij}(x_j), \overline{\alpha}_{ij}(x_j)] = \begin{cases} [\underline{w}_{ij}\,x_j, \overline{w}_{ij}\,x_j] & \text{if } x_j \geq 0 \\ [\overline{w}_{ij}\,x_j, \underline{w}_{ij}\,x_j] & \text{if } x_j < 0, \end{cases} \tag{5}$$

$$\underline{\beta}_i(x) = \sum_{j=1}^{n} \underline{\alpha}_{ij}(x_j) \text{ and } \overline{\beta}_i(x) = \sum_{j=1}^{n} \overline{\alpha}_{ij}(x_j), \tag{6}$$

where $i \in \mathcal{N} := \{1, \ldots, n\}$, $j \in \mathcal{N}$, and where $[w_{ij}]$ is a shorthand notation for the real interval $[w_{ij}] = [\underline{w}_{ij}, \overline{w}_{ij}] \subset \mathbb{R}$. Equation (4) and subsequent equations involve arithmetic operations on intervals. These operations are carried out according to standard interval arithmetics (IA) rules (see e.g. [5]).

We frequently need bounds on $g_i(x)$ and its derivatives on hyperrectangles $\mathcal{B}$. For ease of reference these bounds are summarized in Conds. 1.

Conditions 1: *Let $i \in \mathcal{N}$ be arbitrary. Assume there exist intervals (i) $[g_i]$, (ii) $[(\nabla g_i)_j]$, (iii) $\left[(\nabla^2 g_i)_{jj}\right]$, such that (i) $g_i(x) \in [g_i]$, (ii) $(\nabla g_i(x))_j \in [(\nabla g_i)_j]$, (iii) $(\nabla^2 g_i(x))_{jj} \in [(\nabla g_i)_{jj}]$ for all $x \in \mathcal{B}$ and all $j \in \mathcal{N}$.*

Lemmata 1 and 2 state how $g_i(x)$ can be bounded on $\mathcal{B}$ by affine functions.

Lemma 1: *Let $i \in \mathcal{N}$ be arbitrary. Assume Conds. 1(ii) holds. Then, for all $x \in \mathcal{B}$ and $x' \in \mathcal{B}$, $g_i(x) \in g_i(x') + \sum_{j=1}^n [(\nabla g_i)_j]\ (x_j - x'_j)$.*

Proof. Let $x \in \mathcal{B}$ and $x' \in \mathcal{B}$ be arbitrary. According to the mean value theorem there exists a $\xi \in \mathcal{B}$ on the line between x and x' such that $g_i(x) = g_i(x') + \sum_{j=1}^n (\nabla g_i(\xi))_j\ (x_j - x'_j)$. Since $(\nabla g_i(\xi))_j \in [(\nabla g_i)_j]$ for all $\xi \in \mathcal{B}$, the claim holds. ■

Note that bounds of the form $g_i(x) \in \sum_{i=1}^n [(\nabla g)_j]\, x_j$ as introduced in Eq. (4) result from Lemma 1 for the particular choice $x' = 0$.

If at least one partial derivative of $g_i(x)$ is known to be nonnegative or nonpositive for all $x \in \mathcal{B}$, then an upper bound $g_i(x) \leq \overline{g}_i$ can be replaced by a tighter affine bound as stated in the following lemma. The corresponding lower bound is omitted here for brevity.

Lemma 2: *Let $i \in \mathcal{N}$ be arbitrary. Assume Conds. 1 (i) and (ii) hold. Then, for all $x \in \mathcal{B}$,*

$$g_i(x) \leq \overline{g}_i + \overline{\delta}_i(x) \leq \overline{g}_i \tag{7}$$

where $\overline{\delta}_i(x) = \sum_{j=1}^n \overline{\gamma}_{ij}(x_j)$ and

$$\overline{\gamma}_{ij}(x_j) = \begin{cases} \overline{\nabla g_{ij}}\ (x_j - \underline{x}_j), & \text{if} \quad \overline{\nabla g_{ij}} < 0, \\ \underline{\nabla g_{ij}}\ (x_j - \overline{x}_j), & \text{if} \quad \underline{\nabla g_{ij}} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Proof. The second inequality in Eq. (7) holds, because $\overline{\delta}_i(x) \leq 0$ for all $x \in \mathcal{B}$ by definition. To show the first relation in Eq. (7) assume there exists an $x'' \in \mathcal{B}$ such that $g_i(x'') > \overline{g}_i + \overline{\delta}_i(x'')$ and show that a contradiction results. Let $\mathcal{I} = \{j \in \mathcal{N} \,|\, \underline{\nabla g_{ij}} > 0\}$ and $\mathcal{D} = \{j \in \mathcal{N} \,|\, \overline{\nabla g_{ij}} < 0\}$, then the assumption implies

$$g_i(x'') > \overline{g}_i + \sum_{j \in \mathcal{I}} \underline{\nabla g_{ij}}\ (x''_j - \overline{x}_j) + \sum_{j \in \mathcal{D}} \overline{\nabla g_{ij}}\ (x''_j - \underline{x}_j) \tag{9}$$

On the other hand, lemma 1 yields $g_i(x'') \in g_i(x') + \sum_{j=1}^n [(\nabla g_i)_j]\ (x''_j - x'_j)$. In particular this holds for the choice

$$x'_j = \begin{cases} \overline{x}_j, & \text{if } j \in \mathcal{I}, \\ \underline{x}_j, & \text{if } j \in \mathcal{D}, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

which results in the upper bound

$$g_i(x'') \leq g_i(x') + \sum_{j \in \mathcal{I}} \underline{\nabla g_{ij}}\ (x''_j - \overline{x}_j) + \sum_{j \in \mathcal{D}} \overline{\nabla g_{ij}}\ (x''_j - \underline{x}_j). \tag{11}$$

Combining Eqs. (9) and (11) yields $g_i(x') > \overline{g}_i$, which is a contradiction, since $x' \in \mathcal{B}$ and $g_i(x) \leq \overline{g}_i$ for all $x \in \mathcal{B}$ by assumption. ■

2.1 Using function traits to tighten affine inclusions

The affine inclusions introduced in the previous section can be improved upon, if information on properties such as monotonicity and convexity are available. It turns out to be convenient to discuss these properties not for g_i, but for the auxiliary functions $h_{ij} : \mathcal{L}_j \times \mathcal{B}_j \to \mathbb{R}$ with $\mathcal{L}_j = [x_j] \subset \mathbb{R}$, $\mathcal{B}_j = \{x \in \mathcal{B} \,|\, x_j = 0\} \subset \mathbb{R}^n$ and

$$h_{ij}(x_j, z_j) = g_i(x_j\, e_j + z_j), \tag{12}$$

where e_j represents the j-th unit vector. The desired properties can be related to bounds on the second order derivatives of g_i as in the following Lemma 3, which we state without proof.

Lemma 3: *Assume Conds. 1 (ii) and (iii) hold. Then, for all $z_j \in \mathcal{B}_j$, the function $h_{ij} : \mathcal{L}_j \times \mathcal{B}_j \to \mathbb{R}$ is convex (concave) on $\mathcal{L}_j$, if $\underline{\nabla^2 g_i}_{jj} \geq 0$ ($\overline{\nabla^2 g_i}_{jj} \leq 0$).*

Apart from the properties treated in Lemma 3, the particular form of $\mathcal{B}$ affects the affine inclusions. We anticipate that domains of the particular form $\mathcal{L}_j \in \{[0, \Delta x_j], [-\Delta x_j, \Delta x_j], [-\Delta x_j, 0]\}$ with $\Delta x_j > 0$ will be needed in Sect. 3 and 4. We introduce the trait variables $\mathtt{xv}_{ij}$ and $\mathtt{lr}_j$ defined in Tab. 1 in order to be able to state the traits and domain of a function h_{ij} in a compact fashion.

Table 1: Shorthand notation for function traits and domains.

$\mathtt{xv}_{ij}$	meaning	$\mathtt{lr}_j$	meaning
1	$0 \leq \underline{\nabla^2 g_i}_{jj}$, i.e. h_{ij} is convex	1	$[0, \Delta x_j]$
0	$\underline{\nabla^2 g_i}_{jj} < 0 < \overline{\nabla^2 g_i}_{jj}$	0	$[-\Delta x_j, \Delta x_j]$
-1	$\overline{\nabla^2 g_i}_{jj} \leq 0$, i.e. h_{ij} is concave	-1	$[-\Delta x_j, 0]$

Proposition 1 states our main result on tightening affine inclusions based on trait information.

Proposition 1: *Assume Conds. 1 (i)–(iii) hold. Let $h_{ij} : \mathcal{L}_j \times \mathcal{B}_j \to \mathbb{R}$, $j = 1, \ldots, n$ be defined as in Eq. (12), and assume $\mathcal{L}_j \in \{[0, \Delta x_j], [-\Delta x_j, \Delta x_j], [-\Delta x_j, 0]\}$ for all $j \in \mathcal{N}$. Assume convexity properties $\mathtt{cv}_{ij}$ are known for all h_{ij}, $j = 1, \ldots, n$ and let $[w_{ij}^*]$ be defined as in Tab. 2. Then $g_i(x) \in \sum_{j=1}^n [w_{ij}^*]\, x_j$ for all $x \in \mathcal{B}$ and $[w_{ij}^*] \subseteq [(\nabla g_i)_j]$ for all $j \in \mathcal{N}$.*

We use the following lemma to prove proposition 1.

Lemma 4: *Assume Conds. 1 (i) to (iii) hold. Let $\overline{\beta}_i(x)$ and $\overline{\delta}_i(x)$ be defined as in Eq. (4) and in Lemma 2, respectively. Assume that, for every $j \in \mathcal{N}$, either $[w_{ij}] = [(\nabla g_i)_j]$ or $[w_{ij}] = [w_{ij}^*]$ as defined in Tab. 2. Then $\overline{\delta}_i(x) \leq \overline{\beta}_i(x)$ for all $x \in \mathcal{B}$.*

Proof. By definition $\overline{\delta}_i(x) = \sum_{j=1}^n \overline{\gamma}_{ij}(x_j)$ and $\overline{\beta}_{ij}(x) = \sum_{j=1}^n \overline{\alpha}_{ij}(x_j)$. The claim therefore holds, if, for all $x_j \in \mathcal{L}_j$ and all $j \in \mathcal{N}$,

$$\overline{\gamma}_{ij}(x_j) \leq \overline{\alpha}_{ij}(x_j). \tag{13}$$

It suffices to show that (13) holds for arbitrary $j \in \mathcal{N}$ and all $x_j \in \mathcal{L}_j$. Consider the cases

	$x_j < 0$	$0 \leq x_j$
$0 < \underline{w}_{ij} \leq \overline{w}_{ij}$	(1a)	(1b)
$\underline{w}_{ij} \leq 0 \leq \overline{w}_{ij}$	(2a)	(2b)
$\underline{w}_{ij} \leq \overline{w}_{ij} < 0$	(3a)	(3b)

Recall that $\overline{\gamma}_{ij}(x_j) \leq 0$ for all $x_j \in \mathcal{L}_j$ by the definition of $\overline{\gamma}_{ij}$ given in Eq. (8). On the other hand, the definition of $\overline{\alpha}_{ij}(x_j)$ in Eq. (5) implies $\overline{\alpha}_{ij}(x_j) \geq 0$ in the cases (1b), (2a), (2b), and (3a). In these cases the claim therefore holds, and (1a) and (3b) remain to be considered.

Case (1a): First assume $\underline{w}_{ij} = \underline{w}^*_{ij}$ and note that the assumption $\underline{w}_{ij} \geq 0$ implies $\underline{w}^*_{ij} = \underline{\nabla g_i}_j$ in all rows in Tab. 2. This implies $\underline{w}_{ij} = \underline{w}^*_{ij} = \underline{\nabla g_i}_j(x_j)$ and

$$\overline{\alpha}_{ij}(x_j) = \underline{\nabla g_i}_j x_j \geq \underline{\nabla g_i}_j (x_j - \overline{x}_j) = \overline{\gamma}_{ij}(x_j),$$

where the relations hold by definition of $\overline{\alpha}_{ij}$ in Eq. (5), because $\underline{\nabla g_i}_j = \underline{w}_{ij} > 0$ and $\overline{x}_j \geq 0$ imply $-\underline{\nabla g_i}_j \, \overline{x}_j \leq 0$, and by definition of $\overline{\gamma}_{ij}(x_j)$ in Eq. (8), respectively.

Case (3b) can be proved similarly. ■

Table 2: Rules for $[w^*_{ij}]$.

$\mathtt{xv}_{ij}$	$\mathtt{lr}_j$	$[w^*_{ij}]$
1	1	$\left[\underline{\nabla g_i}_j, \min\left(\frac{\overline{g}_i}{\Delta x_j}, \overline{\nabla g_i}_j\right)\right]$
1	0	$\left[\max\left(\frac{\underline{g}_i - \overline{g}_i}{\Delta x_j}, \underline{\nabla g_i}_j\right), \min\left(\frac{\overline{g}_i - \underline{g}_i}{\Delta x_j}, \overline{\nabla g_i}_j\right)\right]$
1	−1	$\left[\max\left(\frac{-\overline{g}_i}{\Delta x_j}, \underline{\nabla g_i}_j\right), \overline{\nabla g_i}_j\right]$
0	$\{1, 0, -1\}$	$[(\nabla g_i)_j]$
−1	1	$\left[\max\left(\frac{\underline{g}_i}{\Delta x_j}, \underline{\nabla g_i}_j\right), \overline{\nabla g_i}_j\right]$
−1	0	$\left[\max\left(\frac{\underline{g}_i - \overline{g}_i}{\Delta x_j}, \underline{\nabla g_i}_j\right), \min\left(\frac{\overline{g}_i - \underline{g}_i}{\Delta x_j}, \overline{\nabla g_i}_j\right)\right]$
−1	−1	$\left[\underline{\nabla g_i}_j, \min\left(\frac{-\underline{g}_i}{\Delta x_j}, \overline{\nabla g_i}_j\right)\right]$

Proof of Prop. 1. Consider the case $\mathtt{xv}_{ij} = \mathtt{lr}_j = 1$. In this case $h_{ij}(x_j, z_j)$ is convex on $\mathcal{L}_j$ for arbitrary but fixed $z_j \in \mathcal{B}_j$ according to Lemma 3. By $[w_{ij}] = [(\nabla g_i)_j]$, $j = 1, \ldots, n$ denote bounds introduced in Eq. (4) on the components of $\nabla g_i(x)$ on $\mathcal{B}$. By $[\alpha_{ij}(x_j)]$ and $[\beta_i(x)]$ denote the corresponding bounds introduced in Eqs. (5) and (6), respectively.

Since $h_{ij}(x_j, z_j)$ is convex on $\mathcal{L}_j$, we have, for all $\xi \in \mathcal{L}_j$, $\zeta \in \mathcal{L}_j$, and $z_j \in \mathcal{B}_j$,

$$h_{ij}(t\,\xi + (1-t)\,\zeta, z_j) \leq t\, h_{ij}(\xi, z_j) + (1-t)\, h_{ij}(\zeta, z_j) \tag{14}$$

for all $t \in [0, 1]$ (Jensen's inequality). Choosing $\xi = \Delta x_j$ and $\zeta = 0$ yields

$$h_{ij}(t\,\Delta x_j, z_j) \leq t\, h_{ij}(\Delta x_j, z_j) + (1-t)\, h_{ij}(0, z_j).$$

Using Eq. (12) this can be rewritten as

$$g_i(t\,\Delta x_j\, e_j + z_j) \leq t\,(g_i(\Delta x_j\, e_j + z_j) - g_i(z_j)) + g_i(z_j). \tag{15}$$

Below we show that the r.h.s. of Eq. (15) is bounded above on the interval $t \in [0, 1]$ according to

$$t\,(g_i(\Delta x_j\, e_j + z_j) - g_i(z_j)) + g_i(z_j) \leq t\,\overline{g}_i + \overline{\beta}_i(z_j). \tag{16}$$

Combining Eqs. (15) and (16) yields

$$g_i(t\,\Delta x_j\,e_j + z_j) \leq t\overline{g}_i + \overline{\beta}_i(z_j)$$

for all $t \in [0,1]$. Substituting $x_j = t\,\Delta x_j$ yields

$$g_i(x) = g_i(x_j\,e_j + z_j) \leq \frac{\overline{g}_i}{\Delta x_j}\,x_j + \overline{\beta}_i(z_j)$$

for all $x_j \in \mathcal{L}_j$, which proves the claim under the assumption that Eq. (16) holds.

It remains to prove Eq. (16). As a preparation consider the following relations

$$\begin{aligned} g_i(\Delta x_j\,e_j + z_j) &\leq \overline{g}_i + \overline{\delta}_i(\Delta x_j\,e_j + z_j) \leq \overline{g}_i + \overline{\delta}_i(z_j) \\ &\leq \overline{g}_i + \overline{\beta}_i(z_j), \end{aligned} \tag{17}$$

which hold according to Lemma 2, by definition of $\overline{\delta}_i$ in Lemma 2, and according to Lemma 4, respectively.

Turning to Eq. (16) again we first note that this relation obviously holds if the slope $m_{ij} := g_i(\Delta x_j\,e_j + z_j) - g_i(z_j)$ and the offset $g_i(z_j)$ of the line in t on the l.h.s. of Eq. (16) are not larger than the slope and offset on the r.h.s., i.e. $m_{ij} \leq \overline{g}_i$ and $g_i(z_j) \leq \overline{\beta}_i(z_j)$, where the latter relation holds by assumption. Without giving details we claim that Eq. (16) also holds for some values $m_{ij} \not\leq \overline{g}_i$, specifically if

$$\overline{g}_i \geq g_i(\Delta x_j\,e_j + z_j) - \overline{\beta}_i(z_j). \tag{18}$$

Note that this is a less strict condition than $m_{ij} \leq \overline{g}_i$, since

$$g_i(\Delta x_j\,e_j + z_j) - g(z_j) \geq g_i(\Delta x_j\,e_j + z_j) - \overline{\beta}_i(z_j)$$

due to $g_i(z_j) \leq \overline{\beta}_i(z_j)$. The condition (18) can be derived by setting the l.h.s. and r.h.s. of Eq. (16) equal and by showing that the intersection of the two lines in t occurs for a $t > 1$ if Eq. (18) holds. According to Eq. (17), the relation (18) holds, which proves Eq. (16). We claim the other cases stated in Tab. 2 without proof. Finally, we note that $[w_{ij}^*] \subseteq [(\nabla g_i)_j]$ holds for all cases listed in Tab. 2 by definition of $[w_{ij}^*]$. ■

3 Sufficient conditions for positive invariance on ellipsoids

Let $\varphi(t, x(0))$ denote the solution of (1) that passes through $x(0)$ at time $t = 0$, then we call a set $\mathcal{P} \subset \mathcal{X}$ p.i. for the system (1), if, for all $t \geq 0$,

$$x(0) \in \mathcal{P} \text{ implies } \varphi(t, x(0)) \in \mathcal{P}, \tag{19}$$

i.e. any trajectory starting in $\mathcal{P}$ remains in $\mathcal{P}$ for all $t \geq 0$. The following theorem according to [6] states conditions for detecting p.i. sets using Lyapunov functions.

Theorem 2: *(see e.g. [6], Chap. 3.1) Let $\breve{x} = 0$ be an equilibrium point for (1) and $\mathcal{T} \subset \mathcal{X} \subset \mathbb{R}^n$ be a domain that contains $\breve{x}$. Let $v : \mathcal{T} \to \mathbb{R}$ be a continuously differentiable function. If $v(0) = 0$, $v(x) > 0$ for all $x \in \mathcal{T} \setminus \{0\}$, and*

$$\dot{v}(x) := \frac{d}{dt}v(\varphi(t,x))|_{t=0} = f^T(x)\,\nabla v(x) \quad < \quad 0$$

for all $x \in \mathcal{T} \setminus \{0\}$, then $\breve{x}$ is asymptotically stable. Moreover if $\mathcal{V}_c = \{x \in \mathcal{X} \mid v(x) \leq c\}$ is bounded and contained in $\mathcal{T}$, then any trajectory that starts in $\mathcal{V}_c$ remains in $\mathcal{V}_c$ (and tends to $\breve{x}$ for $t \to \infty$), i.e. $\mathcal{V}_c$ represents a p.i. set.

Just as in the previous sections we set $\breve{x} = 0$ without restriction. Quadratic forms are common candidate Lyapunov functions. We consider quadratic forms $v(x) = x^T P x$, $P = P^T \in \mathbb{R}^{n\times n}$, $P \succ 0$ on ellipsoids $\mathcal{E}_c = \{x | x^T P x \leq c\}$, $c \in \mathbb{R}$, $c > 0$ throughout the remainder of the paper. Since $v(0) = 0$ and $v(x) > 0$ for all $x \in \mathcal{E}_c \setminus \{0\}$ hold by construction, only the negative definiteness of $\dot{v}(x)$ remains to be established.

For linear systems of the form (3), $\dot{v}(x(t))$ evaluates to

$$\dot{v}(x(t)) = x(t)^T \left(A^T P + P A\right) x(t).$$

If the matrix

$$Q = -\left(A^T P + P A\right) \tag{20}$$

is p.d., then the linear system (3) is globally asymptotically stable (see e.g. [3] or [6]). This implies that every Lyapunov surface $\mathcal{V}_c \subset \mathcal{X}$ is a p.i. set.

3.1 Nonlinear systems

For nonlinear systems we use the affine inclusions from Sect. 2 to bound the nonlinear part $g : \mathcal{X} \to \mathbb{R}^n$ introduced in Eq. (2). This yields linear interval systems of the form (21)

$$\begin{aligned} f(x) &= A x + g(x) \\ &\subseteq A x + [W] \cdot x = [A] \cdot x \text{ for all } x \in \mathcal{B}, \end{aligned} \tag{21}$$

where $[W]$ and $[A]$ are the matrices with elements $[\underline{w}_{ij}, \overline{w}_{ij}]$ and $[\underline{w}_{ij} + a_{ij}, \overline{w}_{ij} + a_{ij}]$, respectively. Extending (20) from a real matrix A to an interval matrix $[A]$ results in

$$[Q] = -\left([A]^T P + P [A]\right), \tag{22}$$

and stability properties of the nonlinear system can be investigated by investigating the definiteness of $[Q]$. This is made more precise in the following proposition. We call a symmetric interval matrix $[Q]$ p.d., if every symmetric matrix contained in $[Q]$ is p.d..

Proposition 2: *Let P be a p.d. matrix. Assume there exist an interval matrix $[A]$ and a hyperrectangle $\mathcal{B} \subset \mathcal{X}$ such that $f(x) \in [A]\, x$ for all $x \in \mathcal{B}$. If $[Q]$ as defined in Eq. (22) is p.d., then every $\mathcal{V}_c = \{x | x^T P x \leq c\}$ that is contained in $\mathcal{B}$ is a p.i. set of the system (1).*

Proof. The function $v(x) = x^T P x$ is p.d., since P is p.d. by assumption. According to Theorem 2, $\mathcal{V}_c \subset \mathcal{B}$ is p.i., if

$$\dot{v}(x) = \dot{x}^T P x + x^T P \dot{x} = f(x)^T P x + x^T P f(x)$$

is negative definite on $\mathcal{B}$. Since $f(x) \in [A]\, x$ for all $x \in \mathcal{B}$, we find

$$\dot{v}(x) \in x^T \left([A]^T P + P [A]\right) x \tag{23}$$

for all $x \in \mathcal{B}$. Since $[Q]$ is p.d. by assumption, $[A]^T P + P[A] = -[Q]$ is negative definite, therefore $x^T (A^T P + P A)\, x < 0$ for all $A \in [A]$ and all $x \in \mathcal{B} \setminus \{0\}$. Together with Eq. (23) this implies $\dot{v}(x) < 0$ for all $x \in \mathcal{B} \setminus \{0\}$, which proves the claim. ■

Based on Prop. 2 we can search for p.i. sets as follows. We choose a candidate set $\mathcal{V}_c = \{x \,|\, x^T P x \leq c\}$ by selecting a symmetric positive definite matrix $P \in \mathbb{R}^{n\times n}$ and a

$c > 0$. Subsequently, we bound $\mathcal{V}_c$ by a hyperrectangle $\mathcal{B} \supset \mathcal{V}_c$. Following [8] we choose $\mathcal{B} = [-\Delta x_1, \Delta x_1] \times \cdots \times [-\Delta x_n, \Delta x_n]$, where

$$\Delta x_j = \overline{x}_j = \sqrt{c \cdot \sum_{k=j}^{n} u_{jk}^2}, \tag{24}$$

and u_{ij} are the elements of $U = (L^{-1})^T$, where L results from the Cholesky factorization $P = L\,L^T$. After $\mathcal{B}$ has been determined, $[A]$ and $[Q]$ can be calculated, and the p.d. of $[Q]$ can be checked.

3.2 Detecting positive definiteness of interval matrices $[Q]$

A symmetric interval matrix $[Q]$ is p.d., if and only if the smallest eigenvalue

$$\underline{\lambda}^* = \min_{S \in [Q]} \min_{\|x\|=1} x^T S\, x \tag{25}$$

is positive (see e.g. [4]). A lower bound $\underline{\lambda} \leq \underline{\lambda}^*$ can be calculated with an interval variant of Gershgorin's circle criterion [2]. Specifically, $\underline{\lambda} = \min_k \lambda_k$, where

$$\underline{\lambda}_k = \underline{q}_{kk} - \sum_{\substack{i=1 \\ i \neq k}}^{n} \max\left(|\underline{q}_{ik}|, |\overline{q}_{ik}|\right) \tag{26}$$

for $k = 1, \ldots, n$. Since $\underline{\lambda}$ is a lower bound for $\underline{\lambda}^*$, $\underline{\lambda} > 0$ implies that $[Q]$ is p.d.. Clearly, the converse is not true. In fact, Gershgorin's circle is known to be quite conservative.

An exact but numerically more expensive procedure for checking the definiteness of a symmetric interval matrix $[Q]$ was introduced by [4]. According to Hertz the p.d. of $[Q]$ can be established by checking the p.d. of 2^n real matrices $S_l \in [Q]$, where each S_l is associated with one of the 2^n orthants of $\mathbb{R}^n$. By exploiting a certain symmetry, the number of real matrices that need to be analyzed can be reduced to 2^{n-1} [4]. Since checking the p.d. of a real matrix requires $\mathcal{O}(n^3)$ operations, an overall computational complexity of $\mathcal{O}(2^{n-1}\, n^3)$ results.

Hertz's method can be improved upon in the particular application treated here by evaluating $[Q]_l$ on each of the 2^n orthants of $\mathcal{B} \subset \mathbb{R}^n$ separately. Note this does not change the computational complexity significantly, since it is dominated by the exponential growth in the number of orthants. Without giving details, we claim that the computational effort of the modified Hertz method is $\mathcal{O}(2^n\, n^3)$. While twice as expensive as Hertz's method, the modified method provides a less conservative test of p.d. for a nonlinear system.

4 Implementation and complexity

We pointed out in the introduction that we are interested in automated methods for checking positive invariance. In Sect. 4.1 we summarize how the two main steps of the proposed approach, the computation of $[Q]$ and the p.d. test for $[Q]$, can be implemented. In Sect. 4.2 we discuss the computational complexity of several variants of these implementations.

4.1 Sketch of the implementation

The interval matrix $[Q]$ depends on $[W]$ as specified in Eqs. (21) and (22). As detailed in Sect. 2 the elements $[w_{ij}]$ of $[W]$ can either be calculated based on bounds $[\nabla g_i]$, or

based on $[g_i]$, $[\nabla g_i]$ and trait information according to Prop. 1. We introduce the variable $\texttt{DT} \in \{\mathrm{D}, \mathrm{T}\}$, where D ("derivatives") and T ("traits") refer to the methods without and with trait variables. The interval matrix $[Q]$ can be calculated as follows.

```
function intMatQ(A, P, g, B, DT)
  switch(DT)
    case(D) calculate [g_i] and [(∇g_i)_j] on B;
      set [w_ij] = [(∇g_i)_j];
    case(T) calculate [g_i], [(∇g_i)_j] and [(∇²g_i)_jj] on B;
      compute [w_ij] acc. to Tab. 2 depending on B;
  calculate [A] = A + [W];
  calculate [Q] according to (22);
return [Q];
```

Once $[Q]$ has been computed, its p.d. can be checked with one of the methods summarized in Sect. 3.2. We introduce the notation $\texttt{GHO} \in \{\mathrm{G}, \mathrm{H}, \mathrm{O}\}$ to refer to the method based on Gershgorin's circle criterion (G), Hertz's approach (H), and the modified Hertz method that checks the p.d. on each orthant (O). The following function returns $\texttt{pd} = 1$ if $[Q]$ is p.d. and $\texttt{pd} = 0$ otherwise.

```
function checkDef(A, P, g, B, GHO, DT)
  switch(GHO)
    case(G) calculate [Q]=intMatQ(A, P, g, B, DT);
      check p.d. of [Q] using (26) and set pd;
    case(H) calculate [Q]=intMatQ(A, P, g, B, DT);
      for l = 1, ..., 2^(n-1)
        check p.d. of orthant matrix S_l and set pd;
    case(O) for l = 1, ..., 2^n
        calculate [Q]_l =intMatQ(A, P, g, X_l, DT)
        for the l-th orthant X_l of B;
        check p.d. of orthant matrix S_l and set pd;
return pd;
```

Since any of the choices $\texttt{DT} \in \{\mathrm{D}, \mathrm{T}\}$ and $\texttt{GHO} \in \{\mathrm{G}, \mathrm{H}, \mathrm{O}\}$ can be combined with one another, a total of six variants of the proposed method results. Without giving details we state some relations between the sizes of the ellipsoids for which p.i. can be established with these six methods. By $\overline{c}_{\texttt{DT},\texttt{GHO}}$ denote the largest $c \geq 0$ for which p.d. of $\mathcal{V}_c$ can be established for a combination of $\texttt{DT} \in \{\mathrm{D}, \mathrm{T}\}$ and $\texttt{GHO} \in \{\mathrm{G}, \mathrm{H}, \mathrm{O}\}$. Without proof we claim

$$\overline{c}_{\mathrm{T},\texttt{GHO}} \geq \overline{c}_{\mathrm{D},\texttt{GHO}}, \tag{27}$$

for all $\texttt{GHO} \in \{\mathrm{G}, \mathrm{H}, \mathrm{O}\}$ and

$$\overline{c}_{\texttt{DT},\mathrm{O}} \geq \overline{c}_{\texttt{DT},\mathrm{H}} \geq \overline{c}_{\texttt{DT},\mathrm{G}} \tag{28}$$

for all $\texttt{DT} \in \{\mathrm{D}, \mathrm{T}\}$.

4.2 Computational complexity of the algorithms

A precise analysis of the computational effort of the functions `intMatQ` and `checkDef` introduced in Sect. 4.1 is straight forward but beyond the scope of the paper. We summarize some results in Tab. 3, where $N(\cdot)$ denotes the number of operations necessary to compute the respective information. $N([g_i], [(\nabla g_i)_j], [(\nabla^2 g_i)_{jj}])$, for example, represents

the number of operations necessary to calculate the intervals $[g_i]$, $[(\nabla g_i)_j]$, and $[(\nabla^2 g_i)_{jj}]$ for all $i, j \in \mathcal{N}$. Note that the complexity given for $N([g_i], [(\nabla g_i)_j])$ results if backward mode automatic differentiation is used (see e.g. [5]).

Table 3: Computational complexity.

quantity	complexity order
$N(L)$ Cholesky dec.	$\mathcal{O}(n^3)$
$N([g_i], [(\nabla g_i)_j])$	$\mathcal{O}(n)$
$N([g_i], [(\nabla g_i)_j], [(\nabla^2 g_i)_{jj}])$	$\mathcal{O}(n^2)$
$N([W]$ acc. to Tab. 2	$\mathcal{O}(n^2)$
$N([A])$ acc. (21)	$\mathcal{O}(n^2)$
$N([Q])$ acc. (22)	$\mathcal{O}(n^3)$
$N(\underline{\lambda})$ Gershgorin acc. (26)	$\mathcal{O}(n^2)$
$N($`intMatQ`$)$ for all `DT`	$\mathcal{O}(n^3)$
$N($`checkDef`$)$, `GHO` $=$ G, for all `DT`	$\mathcal{O}(n^3)$
$N($`checkDef`$)$, `GHO` $=$ H, for all `DT`	$\mathcal{O}(2^{n-1}\, n^3)$
$N($`checkDef`$)$, `GHO` $=$ O, for all `DT`	$\mathcal{O}(2^n\, n^3)$

We briefly note that the complexities for `intMatQ` and `checkDef` in Tab. 3 result from the entries above the horizontal line. For example, $N($`intMatQ`$) = N([g_i], [(\nabla g_i)_j]) + N([Q]) = \mathcal{O}(n) + \mathcal{O}(n^3) = \mathcal{O}(n^3)$. Note that the choice of the method `DT` has no effect on the order of $N($`intMatQ`$)$. The exact number of operations does depend on `DT`, however.

Finally, we claim without proof that closer inspection of the functions `intMatQ` and `checkDef` results $N($`checkDef`$($`GHO`$, $`DT`$ = T)) > N($`checkDef`$($`GHO`$, $`DT`$ = D))$ for all `GHO` $= \{\mathrm{G}, \mathrm{H}, \mathrm{O}\}$ and $N($`checkDef`$($`GHO`$ = O, $`DT`$)) > N($`checkDef`$($`GHO`$ = H, $`DT`$)) > N($`checkDef`$($`GHO`$ = G, $`DT`$))$ for all `DT` $= \{\mathrm{D}, \mathrm{T}\}$.

5 Examples

We consider a continuous time system in Sect. 5.1. Section 5.2 briefly introduces the extension of the approach to discrete time systems and presents another example.

5.1 Continuous time example

Consider the following model of a nonlinear oscillator

$$\begin{aligned} \dot{x}_1(t) &= f_1(x(t)) = x_2(t) \\ \dot{x}_2(t) &= f_2(x(t)) = -0.5\,(x_1(t) + x_1^2(t)) - 0.2\,x_2(t) \end{aligned} \tag{29}$$

with equilibrium $\breve{x} = 0$. Separating the functions f_i into their linear and nonlinear parts according to Eq. (2) results in

$$A = \begin{pmatrix} 0 & 1 \\ -0.5 & -0.2 \end{pmatrix} \quad \text{and} \quad g(x) = \begin{pmatrix} 0 \\ -0.5\,x_1^2 \end{pmatrix}.$$

Since the eigenvalues of A are $\lambda_{1,2} = -0.1 \pm 0.7i$, the equilibrium $\breve{x} = 0$ is locally asymptotically stable. From $g_1(x) = 0$ we infer $[g_1] = [w_{11}] = [w_{12}] = [0, 0]$. Since $g_2(x)$

is independent of x_2, $[w_{22}] = [0, 0]$. For $[g_2]$ and the remaining element $[w_{21}]$ of $[W]$, interval arithmetics results in $[g_2] = [-0.5 \max(\underline{x}_1^2, \overline{x}_1^2), 0]$, $[(\nabla g_2)_1] = [-\overline{x}_1, -\underline{x}_1]$, and $[(\nabla^2 g_2)_{11}] = [-1, -1]$.

Table 4: Results for $[w_{21}]$.

DT	GHO	$\mathtt{xv}_{21}$	$\mathtt{lr}_1$	orthant k	$[w_{21}]$
D	G, H, O	–	–	–	$[-\Delta x_1, \Delta x_1]$
T	G, H	-1	0	–	$[-0.5\,\Delta x_1, 0.5\,\Delta x_1]$
T	O	-1	1	1, 3	$[-0.5\,\Delta x_1, 0]$
		-1	-1	2, 4	$[0, 0.5\,\Delta x_1]$

Table 4 lists the bounds $[w_{21}]$ that result on $\mathcal{B} = [-\Delta x_1, \Delta x_1] \times [-\Delta x_2, \Delta x_2]$ with the methods introduced here. The results in rows 3-5 of Tab. 4 are calculated according to Tab. (2) using trait information.

In order to calculate $[Q]$ according to Eq. (22) we need to choose a p.d. matrix P. For the sake of a concise comparison of all methods, we choose the same P in all cases. Specifically, we calculate P by solving

$$\max_P \operatorname{vol}\left(\mathcal{V}_1 = \{x | v(x) = x^T P\, x \leq 1\}\right) \text{ s.t. } \dot{v}(x) \text{ n.d.} \forall\, x \in \mathcal{V}_1$$

by local optimization on a grid of starting values. This results in

$$P = \begin{pmatrix} 6.641 & 1.956 \\ 1.956 & 12.877 \end{pmatrix}.$$

Figure 1 illustrates the resulting $\mathcal{V}_1$ along with the regions $\mathcal{V}_{\overline{c}} = \{x | x^T P\, x \leq \overline{c}\}$, for the values $\overline{c}$ that result with the variants of the method presented here. Table 5 lists the numerical results for $\overline{c}$. Note that in general bigger p.i. sets can be found if other geometries than ellipsoids are admitted. Figure 1 shows the set $\mathcal{P} = \{x | v(x) = \frac{1}{6}x_1^3 + \frac{1}{4}x_1^2 + \frac{1}{2}\,x_2^2 \leq \frac{1}{12}\}$, which can be found here, because a potential is known for the nonlinear oscillator.

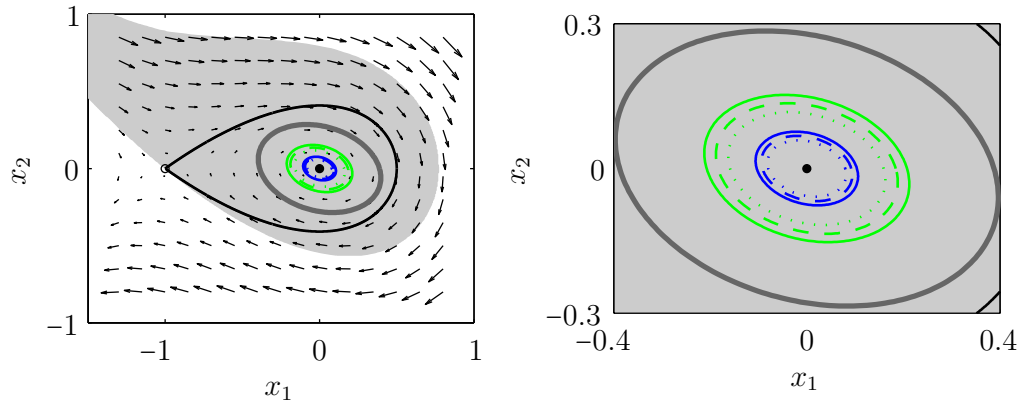

Figure 1: Plot of the domain of attraction of (29) (light gray). Optimal ellipsoidal Lyapunov surfaces (gray). Lyapunov surfaces obtained for $\mathtt{DT} = \mathrm{T}$ (green) and $\mathtt{DT} = \mathrm{D}$ (blue). Solid, dashed and dotted lines refer to $\mathtt{GHO} = \mathrm{O}, \mathrm{H}, \mathrm{G}$, respectively.

5.2 Discrete time example

The proposed approach can be extended to discrete time systems of the form

$$x(t_{k+1}) = f(x(t_k)) \tag{30}$$

in a straight forward fashion. Without restriction we assume $\breve{x} = 0$ is a fixed point. Essentially, the derivative along trajectories $\dot{v}(x(t))$ introduced in Prop. 2 has to be replaced by the forward difference $\Delta v(x(t_k)) = v(x(t_{k+1})) - v(x(t_k))$. This gives rise to the interval matrix

$$[Q] = P - [A]^T P\,[A] \tag{31}$$

in the second Lyapunov equation. As an example for a discrete time system, we analyze a variant of a Lotka-Volterra model given by

$$\begin{aligned} x_1(t_{k+1}) &= 0.9\,x_1(t_k) + 0.1\,x_1^2(t_k) - 0.1\,x_1(t_k)\,x_2(t_k) \\ x_2(t_{k+1}) &= 0.8\,x_2(t_k) + 0.2\,x_2^2(t_k) - 0.1\,x_2(t_k)\,x_3(t_k) \\ x_3(t_{k+1}) &= 0.7\,x_3(t_k) + 0.3\,x_3^2(t_k) - 0.1\,x_3(t_k)\,x_1(t_k). \end{aligned} \tag{32}$$

Separating linear and nonlinear contributions according to Eq. (2) yields

$$A = \begin{pmatrix} 0.9 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.7 \end{pmatrix} \quad \text{and} \quad g(x) = \begin{pmatrix} 0.1\,x_1^2 - 0.1\,x_1\,x_2 \\ 0.2\,x_2^2 - 0.1\,x_2\,x_3 \\ 0.3\,x_3^2 - 0.1\,x_3\,x_1 \end{pmatrix}.$$

A matrix P can be found just as in the continuous time example presented in Sect. 5.1. The resulting matrix reads

$$P = \begin{pmatrix} 1.209 & 0 & -0.003 \\ 0 & 1.065 & 0 \\ -0.003 & 0 & 1.054 \end{pmatrix}.$$

Positive invariance can be established for the sets $\mathcal{V}_c = \{x | x^T P\,x \leq \overline{c}\}$ for the values of $\overline{c}$ given in Tab. 5.

Table 5: Results for $\overline{c}_{\mathtt{DT},\mathtt{GHO}}$ for the different variants $\mathtt{DT} \in \{\mathrm{D},\mathrm{T}\}$, $\mathtt{GHO} = \{\mathrm{G},\mathrm{H},\mathrm{O}\}$ of the method.

Sys.	D,G	D,H	D,O	T,G	T,H	T,O
(29)	0.042	0.056	0.071	0.168	0.225	0.286
(32)	0.074	0.101	0.122	0.075	0.111	0.253

The results in Tab. 5 corroborate the relations between the various $\overline{c}_{\mathtt{DT},\mathtt{GHO}}$ stated in Eqs. (27) and (28).

6 Conclusion

We investigated approaches to an automatic detection of p.i. sets of nonlinear autonomous continuous and discrete time systems. Six variants of a method for the detection of p.i. ellipsoidal sets were introduced and analyzed. We showed that the order of the computational complexity of the introduced approaches varies between $\mathcal{O}(n^3)$ and $\mathcal{O}(2^n\,n^3)$. The

considered examples suggest that p.i. detection can be improved upon at a reasonable additional computational cost by exploiting traits of the model equations such as certain convexity properties. Furthermore, is apparent that the proposed function trait method is an interesting candidate in particular if combined with an orthant based analysis of the state space (case $\texttt{DT} = \mathrm{T}$, $\texttt{GHO} = \mathrm{O}$). This is reasonable, since tighter affine inclusions can be obtained with trait information in general. The improvement turns out to be considerable, however, if subdomains are considered that do not contain 0 in their interior (cases $\texttt{lr} \neq 0$).

Future work has to address the application of the method for the calculation of terminal sets (or regions) occurring within the framework of nonlinear model predictive control with guaranteed stability.

Acknowledgment

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, MO 1086/9).

References

[1] F. Blanchini. Set invariance in control. *Automatica*, 35:1747–1767, 1999.

[2] S. Gershgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk SSSR, Ser. fizmat.*, 6:749–754, 1931.

[3] W. Hahn. *Stability of Motion*. Springer Verlag, Berlin, 1967.

[4] D. Hertz. The extreme eigenvalues and stability of real symmetric interval matrices. *IEEE Transactions on automatic control*, 37:532–535, 1992.

[5] R. B. Kearfott. *Rigorous Global Search: Continuous Problems*. Kluwer Academic Publishers, 1996.

[6] H. K. Khalil. *Nonlinear Systems*. Prentice Hall, 1996.

[7] M. Mönnigmann. Positive invariance tests with efficient Hessian matrix eigenvalue bounds. In *Proc. of 17th IFAC World Congress*, pp. 1117–1122, 2008.

[8] A. Neumaier. The wrapping effect, ellipsoid arithmetic, stability and confidence regions. *Computing Supplementum*, 9:175–190, 1993.