Manuscript submitted to doi:10.3934/xx.xx.xx.xx
Volume **X**, Number **0X**, XX **200X** pp. **X–XX**

# STEPLENGTH THRESHOLDS FOR INVARIANCE PRESERVING OF DISCRETIZATION METHODS OF DYNAMICAL SYSTEMS ON A POLYHEDRON

Zoltán Horváth

Department of Mathematics and Computational Sciences
Széchenyi István University
9026 Győr, Egyetem tér 1, Hungary

Yunfei Song and Tamás Terlaky

Department of Industrial and Systems Engineering
Lehigh University
200 West Packer Avenue, Bethlehem, PA, 18015-1582, USA

(Communicated by Kok Lay Teo)

Abstract. Steplength thresholds for invariance preserving of three types of discretization methods on a polyhedron are considered. For Taylor approximation type discretization methods we prove that a valid steplength threshold can be obtained by finding the first positive zeros of a finite number of polynomial functions. Further, a simple and efficient algorithm is proposed to numerically compute the steplength threshold. For rational function type discretization methods we derive a valid steplength threshold for invariance preserving, which can be computed by using an analogous algorithm as in the first case. The relationship between the previous two types of discretization methods and the forward Euler method is studied. Finally, we show that, for the forward Euler method, the largest steplength threshold for invariance preserving can be computed by solving a finite number of linear optimization problems.

1. **Introduction.** Invariant set is an important concept in the theory of dynamical systems and it has a wide range of applications in control. One of the reasons of the interest is due to the fact that invariant sets enable us to estimate the attraction region of a dynamical system. We consider linear continuous dynamical systems in the form

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

and discrete dynamical systems in the form

$$x_{k+1} = A_d x_k, \tag{2}$$

where $x_k, x(t) \in \mathbb{R}^n$ are the state variables, $A_c, A_d \in \mathbb{R}^{n\times n}$ are the coefficient matrices, and $t \in \mathbb{R}$ and $k \in \mathbb{N}$ indicate continuous and discrete time steps, respectively. Note that equations (1) and (2) can be treated as autonomous systems or as controlled systems. In the latter case, the coefficient matrix $A_c$ (or $A_d$) can be represented in the form of $A + BF$, where $A$ is the open-loop state matrix, $B$ is the

2010 *Mathematics Subject Classification.* Primary: 34A30, 37M25, 65K05.
*Key words and phrases.* Dynamical System, Invariant Set, Polyhedron, Discretization method, Invariance Preserving.

.

control matrix, and $F$ is the gain matrix. For simplicity, we use the term system to indicate dynamical system.

Intuitively, a set $\mathcal{S}$ is called an invariant set for a system, if all the trajectories of the system, which are starting in $\mathcal{S}$, remain in $\mathcal{S}$. Numerous surveys on the theory and applications of invariant sets are published in the recent decades, see e.g., Blanchini [3]. Recently, several sufficient and necessary conditions, which are simply refereed to as invariance conditions, are derived to verify if a set is an invariant set for a continuous or discrete system. Various convex sets with different characteristics are considered as candidates for invariant sets. Invariance conditions for polyhedra are given in [4, 5, 7, 8, 9]. Ellipsoidal sets as invariant sets are analyzed in [6]. Cones as invariant sets are studied in [18, 21, 23]. A novel unified approach to derive invariance conditions for polyhedra, ellipsoids, and cones is presented in [16].

Although many mathematical techniques are developed to directly solve continuous systems, in practice, one usually solves a continuous system by applying certain discretization methods. Assume that a set is an invariant set for a continuous system, then it should be also an invariant set for the discrete system, which is obtained by the discretization method, i.e., discretization should preserve the invariance. However, this is not always true for every steplength used in the discretization method, thus it will be convenient if there exists a predictable threshold for valid invariance preserving steplength. The existence of such steplength thresholds of invariance preserving on various sets is thoroughly studied in [15]. In this paper, we consider three types of discretization methods on polyhedra and we aim to derive valid thresholds of the steplength in terms of explicit form or obtained by using efficiently computable algorithms. The popularity of polyhedra as invariant sets is due to the fact that the state and control variables are usually represented in terms of linear inequalities. For Taylor approximation type discretization methods, i.e., the coefficient matrix of the discrete system is derived from the Taylor expansion of $e^{A_c\Delta t}$, we present an algorithm to derive a valid steplength threshold for invariance preserving. In particular, the algorithm aims to find the first positive zeros of some polynomial functions related to the system and the polyhedron. For general rational function type discretization methods, i.e., the coefficient matrix of the discrete system is a rational function with respect to $A_c$ and $\Delta t$, we derive a valid steplength threshold for invariance preserving that can be computed by using analogous methods as for the case of Taylor approximation type methods. This steplength threshold is related to the steplength threshold for the forward Euler method and the radius of absolute monotonicity of the discretization method. We note that this result is similar to the one presented in [13, 14], where Runge-Kutta methods are considered. Finally, we propose an optimization model to find the largest steplength threshold for the forward Euler method. We note that some results on the use of the forward Euler method to analyze invariance for continuous dynamical systems can be found in [4, 5].

*Notation*: For the sake of simplicity, the following notational conventions are introduced. A *nonnegative* matrix, denoted by $H \geq 0$, means that all entries of $H$ are nonnegative. An *off-diagonal nonnegative* matrix, denoted by $H \geq_o 0$, means that all entries, except the diagonal entries, of $H$ are nonnegative.

The paper is organized as follows. In Section 2, some fundamental concepts, theorems, and the key problems in this paper are introduced. In Section 3, we present our main results, i.e., deriving valid steplength thresholds for invariance

preserving, for the three types of discretization methods. Finally, conclusions are provided in Section 4.

## 2. Background.

We now introduce the definitions of invariant sets for continuous and discrete systems.

**Definition 2.1.** A set $\mathcal{S}$ in $\mathbb{R}^n$ is an invariant set for

- the continuous system (1) if $x(0) \in \mathcal{S}$ implies $x(t) \in \mathcal{S}$, for all $t \geq 0$.
- the discrete system (2) if $x_k \in \mathcal{S}$ implies $x_{k+1} \in \mathcal{S}$, for all $k \in \mathbb{N}$.

According to the definitions of invariant sets, we have that an invariant set means that the continuous (or discrete) trajectory of the system remains in the same set. In fact, there is an alternative perspective, see e.g., [16]. In that interpretation $\mathcal{S}$ is an invariant set for (1) if and only if $e^{A_c t}\mathcal{S} \subseteq \mathcal{S}$ for any $t \geq 0$, and $\mathcal{S}$ is an invariant set for (2) if and only if $A_d\mathcal{S} \subseteq \mathcal{S}$.

In this paper, candidate invariant sets are restricted to convex polyhedron in $\mathbb{R}^n$. A polyhedron $\mathcal{P}$ in $\mathbb{R}^n$ can be characterized as the intersection of a finite number of half spaces.

**Definition 2.2.** A polyhedron $\mathcal{P}$ in $\mathbb{R}^n$ is defined as

$$\mathcal{P} = \{x \in \mathbb{R}^n \,|\, g_1^T x \leq b_1, g_2^T x \leq b_2, ..., g_m^T x \leq b_m\} := \{x \in \mathbb{R}^n \,|\, Gx \leq b\}, \quad (3)$$

where $g_1, g_2, ..., g_m \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and $G^T = [g_1, g_2, ..., g_m] \in \mathbb{R}^{n\times m}$.

Two classical subsets of polyhedra are extensively studied in many applications. One is called polytope, which is a bounded polyhedron. The other one is called polyhedral cone, a polyhedron with $b = 0$ in (3), and the origin is its only vertex.

Given a system and a polyhedron, the invariance condition indicates sufficient and necessary condition such that the polyhedron is an invariant set for the system. There are many such equivalent invariance conditions, e.g., [2, 7]. The most common ones are presented in Theorem 2.3. A novel and unified approach to derive these invariance conditions is proposed in [16]. The invariance conditions in Theorem 2.3 provide powerful and practical tools to verify whether a polyhedron is an invariant set for a given system.

**Theorem 2.3.** [2, 7, 16] *A polyhedron $\mathcal{P}$ given in the form of (3) is an invariant set for*

- *the continuous system (1) if and only if there exists an $H \in \mathbb{R}^{m\times m}$, such that*
$$H \geq_o 0, \;\; HG = GA_c, \;\; \text{and } Hb \leq 0. \quad (4)$$
- *the discrete system (2) if and only if there exists an $\tilde{H} \in \mathbb{R}^{m\times m}$, such that*
$$\tilde{H} \geq 0, \;\; \tilde{H}G = GA_d, \;\; \text{and } \tilde{H}b \leq b. \quad (5)$$

From the theoretical perspective, when a discretization method is applied to a continuous system, the invariant polyhedron for the continuous system should also be an invariant set for the discrete system. This means that conditions (4) and (5) are satisfied simultaneously, when the system, polyhedron, and discretization method are given. However, this is not always true. Intuitively, the smaller steplength used in the discretization method has larger possibility to yield that the polyhedron is also an invariant set for the discrete system. For the sake of self-contained presentation, the formal definitions of invariance preserving and steplength threshold are introduced as follows.

**Definition 2.4.** Assume a polyhedron $\mathcal{P}$ is an invariant set for the continuous system (1), and a discretization method is applied to the continuous system to yield a discrete system. If there exists a $\tau > 0$, such that $\mathcal{P}$ is also an invariant set for the discrete system for any steplength $\Delta t \in [0, \tau]$, then the discretization method is **invariance preserving** for $\Delta t \in [0, \tau]$ on $\mathcal{P}$, and $\tau$ is a **steplength threshold** for invariance preserving of this discretization method on $\mathcal{P}$.

The steplength threshold in Definition 2.4 implies that any value smaller than this threshold is also a valid steplength threshold[1]. This is an important property. In certain cases, a discretization method may be invariance preserving on a set in the form of $[0, \tau_1] \cup [\tau_2, \tau_3]$, where $\tau_1 < \tau_2$. Here we are only interested in finding $\tau_1$. We also note that the steplength threshold in Definition 2.4 is uniform[2] on $\mathcal{P}$, i.e., $\tau$ needs to be a valid steplenth threshold for every initial point in $\mathcal{P}$.

Since a continuous system is usually solved by using various discretization methods in practice, invariance preserving property of the chosen discretization method plays an important role. Further, a larger steplength threshold has many advantages in practice. For example, for larger steplength, the size of the discretized system is smaller, which yields that the computation is less expensive. Thus, we introduce the key problem in the paper:

> *Find a valid (if possible the largest) steplength threshold $\tau > 0$, such that a discretization method is invariance preserving for every $\Delta t \in [0, \tau]$ on $\mathcal{P}$.*

## 3. Main Results.

In this section, we present the approaches for computing a valid (or largest) steplength threshold such that three classes of discretization methods are invariance preserving on a polyhedron. These three classes of discretization methods are considered in the following order: Taylor approximation type discretization methods, rational function type discretizatin methods, and the forward Euler method. The Taylor approximation type represents a family of explicit methods. The rational function type is an extended family of the Taylor approximation type, which also includes some implicit methods. The relationship between these discretization methods and the forward Euler method is also studied. Finally, for the forward Euler method, we derive the largest steplength threshold for invariance preserving.

### 3.1. Taylor Approximation Type Discretization Methods.

We first consider the Taylor approximation type discretization methods. Note that the solution of the continuous system (1) is explicitly represented as $x(t) = e^{A_c t} x_0$, thus one can use the Taylor approximation to numerically solve the continuous system. The $p$-order Taylor approximation of $e^{A_c \Delta t}$ is given as follows:

$$e^{A_c \Delta t} \approx I + A_c \Delta t + \frac{1}{2!} A_c^2 \Delta t^2 + \cdots + \frac{1}{p!} A_c^p \Delta t^p = \sum_{i=0}^{p} \frac{1}{i!} A_c^i \Delta t^i := A_d. \tag{6}$$

The discrete system obtained by applying the Taylor approximation type discretization methods is given as $x_{k+1} = A_d x_k$, where $A_d$ is defined by (6). In fact, the Taylor

[1] This is a key reason why the problem of finding a valid steplength threshold is not an easy problem. In the interval $[0, \tau]$, one needs to check every $\Delta t$ in this interval, which means that there are infinitely many values to be considered.

[2] This is another key reason why the problem of finding a valid steplength threshold is not an easy problem.

approximation type methods form a family of discretization methods. For example, $p = 1$ corresponds to the forward Euler method, $p = 2$ corresponds to the general Runge-Kutta 2nd order methods.

3.1.1. *Existence of Steplength Threshold.* Our approach to derive steplength threshold is based on the invariance conditions presented in Theorem 2.3. The basic ideas is that we build the relationship between these two invariance conditions of the continuous and discrete systems. In fact, conditions (4) and (5) are essentially linear feasibility problems [19]. The unknowns in the two invariance conditions are the matrix $H$ and $\tilde{H}$ given by (4) and (5), respectively. Thus, the key is to find relationship between those matrices.

**Lemma 3.1.** [12] *Assume $H$ satisfies (4), then there exists $\gamma > 0$, such that $\hat{H} = H + \gamma I \geq 0$.*

*Proof.* Since $H \geq_o 0$, we can choose $\gamma > \max\{0, -\min\{h_{ii},\ 1 \leq i \leq n\}\}$, which yields $H + \gamma I \geq 0$. The result is immediate by taking $\hat{H} = H + \gamma I$, □

We note that $\gamma$ in Lemma 3.1 is not unique, e.g., any value greater than a valid $\gamma$ is also valid. We will show more about the effect of $\gamma$ to the steplength threshold in Section 3.2, and the way to derive a larger steplength threshold based on $\gamma$ is also presented.

**Lemma 3.2.** *Assume $H$ satisfies (4), and define*

$$\tilde{H}(\Delta t) = I + H\Delta t + \frac{1}{2!}H^2\Delta t^2 + \cdots + \frac{1}{p!}H^p\Delta t^p = \sum_{i=0}^{p}\frac{1}{i!}H^i\Delta t^i. \tag{7}$$

*a). For the $\gamma$ and $\hat{H}$ given in Lemma 3.1, we have*

$$\tilde{H}(\Delta t) = f_0(\Delta t)I + f_1(\Delta t)\hat{H} + ... + f_p(\Delta t)\hat{H}^p, \tag{8}$$

*where*

$$f_i(\Delta t) = \sum_{k=i}^{p}\frac{(-1)^{k-i}}{k!}\binom{k}{i}\gamma^{k-i}\Delta t^k,\ \text{for } i = 0, 1, ..., p, \tag{9}$$

*and*

$$\sum_{i=0}^{p}\gamma^i f_i(\Delta t) = 1. \tag{10}$$

*b). Let $\tau = \min_{i=0,...,p}\{\tau_i\}$, where $\tau_i$ is the first positive zero of $f_i(\Delta t)$. Then for all $\Delta t \in [0, \tau]$, the matrix $\tilde{H}(\Delta t)$ satisfies (5), where $A_d$ is defined by (6).*

*Proof.* a). According to Lemma 3.1, there exists $\gamma > 0$, such that $\hat{H} = H + \gamma I \geq 0$. The matrix $\tilde{H}(\Delta t)$ given by (7) is represented in terms of $\Delta t$. By substituting $H = \hat{H} - \gamma I$ into (7), we now reformulate $\tilde{H}(\Delta t)$ in terms of $\hat{H}$, i.e.,

$$\begin{aligned}\tilde{H}(\Delta t) = I + (\hat{H} - \gamma I)\Delta t + \frac{1}{2!}(\hat{H}^2 - 2\gamma\hat{H} + \gamma^2 I)\Delta t^2 + \cdots \\ + \frac{1}{p!}(\hat{H}^p - p\gamma\hat{H}^{p-1} + \cdots + (-1)^p\gamma^p I)\Delta t^p.\end{aligned} \tag{11}$$

According to (11), the coefficients of $\hat{H}^i$, for $i = 0, 1, ..., p$, is given as

$$\frac{1}{i!}\Delta t^i + \frac{-1}{(i+1)!}\binom{i+1}{i}\gamma\Delta t^{i+1} + \frac{(-1)^2}{(i+2)!}\binom{i+2}{i}\gamma^2\Delta t^{i+2} + \cdots + \frac{(-1)^{p-i}}{p!}\binom{p}{i}\gamma^{p-i}\Delta t^p,$$

which is the same as (9).

We note that $\sum_{i=0}^{p}\gamma^i f_i(\Delta t)$ is equivalent to replacing $I$ and $\hat{H}$ in (8) by 1 and $\gamma$, respectively. Then, according to (11), we have

$$\sum_{i=0}^{p}\gamma^i f_i(\Delta t)=\sum_{i=0}^{p}\frac{1}{i!}(\gamma\Delta t)^i\sum_{k=0}^{i}(-1)^k\binom{i}{k}. \tag{12}$$

For $i>0$, we have $\sum_{k=0}^{i}(-1)^k\binom{i}{k}=(x-1)^i|_{x=1}=0$, implying that the right hand side of (12) equals to 1, thus (10) follows immediately.

b). We note that for every $i$ the first term of $f_i(\Delta t)$ given as in (9) is $\frac{1}{i!}\Delta t^i$. Then we can write

$$f_i(\Delta t)=\frac{\Delta t^i}{i!}\big(1+\mathcal{O}(\Delta t)\big). \tag{13}$$

Thus, we have that there exists a $\tau_i>0$, i.e., the first positive zero of $f_i(\Delta t)$, where $\tau_i$ may be infinity, such that $f_i(\Delta t)\geq 0$ for all $\Delta t\in[0,\tau_i]$. Then we let

$$\tau=\min_{i=0,1,\ldots,p}\{\tau_i\}, \tag{14}$$

thus we have $f_i(\Delta t)\geq 0$ for all $\Delta t\in[0,\tau]$ and $i=0,1,...,p$. According to (8), and by noting that $\hat{H}^i\geq 0$ for any $i=1,2,...,p$, we have that $\tilde{H}(\Delta t)\geq 0$ for all $\Delta t\in[0,\tau]$, where $\tau$ is defined by (14). Thus, we have proved that the first condition in (5) is satisfied.

By recursively using $HG=GA_c$, for any $i$, we have

$$H^iG=H^{i-1}(HG)=H^{i-1}GA_c=H^{i-2}(HG)A_c=H^{i-2}GA_c^2=...=GA_c^i. \tag{15}$$

Then, according to (15), and substituting (7) and (6), we have

$$\tilde{H}(\Delta t)G=\sum_{i=0}^{p}\frac{1}{i!}H^iG\Delta t^i=\sum_{i=0}^{p}\frac{1}{i!}GA^i\Delta t^i=G\sum_{i=0}^{p}\frac{1}{i!}A^i\Delta t^i=GA_d.$$

Thus, we have proved that the second condition in (5) is satisfied.

Since $H$ satisfies (4), we have $Hb\leq 0$. Also, note that $H=\hat{H}-\gamma I$, thus we have $(\hat{H}-\gamma I)b\leq 0$, i.e., $\frac{\hat{H}}{\gamma}b\leq b$. Since $\frac{\hat{H}}{\gamma}\geq 0$, we have

$$\Big(\frac{\hat{H}}{\gamma}\Big)^i b\leq b,\ \text{i.e., } \hat{H}^ib\leq\gamma^ib,\ \text{for any } i=1,2,...,p. \tag{16}$$

Then, according to (16) and (10), we have

$$\begin{aligned}\tilde{H}_{\Delta t}b&=(f_0(\Delta t)I+f_1(\Delta t)\hat{H}+\cdots+f_p(\Delta t)\hat{H}^p)b\\&\leq(f_0(\Delta t)+\gamma f_1(\Delta t)+\cdots+\gamma^p f_p(\Delta t))b\\&\leq b.\end{aligned}$$

Thus, we have proved that the third condition in (5) is satisfied. The proof is complete. □

Lemma 3.2 presents an important relationship between the two matrices $H$ and $\tilde{H}$ corresponding to the continuous and discrete systems invariance conditions. This relationship is explicitly represented in (7), which is derived from the Taylor approximation (6). According to Lemma 3.2 and Theorem 2.3, we have the following theorem.

**Theorem 3.3.** *Assume a polyhedron $\mathcal{P}$ be given as in (3) is an invariant set for the continuous system (1), and a Taylor approximation type discretization method (6) is applied to the continuous system (1). Then, the steplength threshold $\tau > 0$ as given in Lemma 3.2 is a valid steplength threshold for invariance preserving for the given Taylor approximation type discretization method (6) on $\mathcal{P}$.*

According to the proof of Lemma 3.2, we have that a valid $\tau$ requires $f_i(\Delta t) \geq 0$ for all $\Delta t \in [0, \tau]$ and all $i = 0, 1, ..., p$, where $f_i(\Delta t)$ given as (9). Since each $f_i(\Delta t)$ can be represented in the form of (13), the following corollary is immediate.

**Corollary 1.** *The value of $\tau$ given in Theorem 3.3 (or Lemma 3.2) is a valid steplength threshold for invariance preserving on $\mathcal{P}$ for the Taylor approximation type discretization methods (6). To compute $\tau$, one needs to find the first positive zeros of finitely many polynomial functions in the form*

$$f(\Delta t) = 1 + \alpha_1 \Delta t + \alpha_2 \Delta t^2 + ... + \alpha_q \Delta t^q, \ \alpha_q \neq 0, \tag{17}$$

*where $\alpha_1, \alpha_2, ..., \alpha_q \in \mathbb{R}$ and $q \in \mathbb{N}$.*

In fact, Lemma 3.2 can be extended to a more general case for polynomial approximation rather than Taylor type discretization methods.

**Theorem 3.4.** *Assume $H$ satisfies (4), and define*

$$\tilde{H}(\Delta t) = I + \sigma_1 H \Delta t + \sigma_2 H^2 \Delta t^2 + \cdots + \sigma_p H^p \Delta t^p = \sum_{i=0}^{p} \sigma_i H^i \Delta t^i. \tag{18}$$

*a). For the $\gamma$ and $\hat{H}$ given in Lemma 3.1, we have*

$$\tilde{H}(\Delta t) = f_0(\Delta t) I + f_1(\Delta t) \hat{H} + ... + f_p(\Delta t) \hat{H}^p, \tag{19}$$

*where*

$$f_i(\Delta t) = \sum_{k=i}^{p} (-1)^{k-i} \sigma_k \binom{k}{i} \gamma^{k-i} \Delta t^k, \ \text{for } i = 0, 1, ..., p, \tag{20}$$

*and*

$$\sum_{i=0}^{p} \gamma^i f_i(\Delta t) = 1. \tag{21}$$

*b). Let $\tau = \min_{i=0,...,p}\{\tau_i\}$, where $\tau_i$ is the first positive zero of $f_i(\Delta t)$. Then for all $\Delta t \in [0, \tau]$, the matrix $\tilde{H}(\Delta t)$ satisfies (5), where $A_d$ is defined by (6).*

*Proof.* a). According to Lemma 3.1, there exists a $\gamma > 0$, such that $\hat{H} = H + \gamma I \geq 0$. The matrix $\tilde{H}(\Delta t)$ given by (18) is represented in terms of $\Delta t$. By substituting $H = \hat{H} - \gamma I$ into (18), we now reformulate $\tilde{H}(\Delta t)$ in terms of $\hat{H}$, i.e.,

$$\begin{aligned} \tilde{H}(\Delta t) = I &+ \sigma_1 (\hat{H} - \gamma I) \Delta t + \sigma_2 (\hat{H}^2 - 2\gamma \hat{H} + \gamma^2 I) \Delta t^2 + \cdots \\ &+ \sigma_p (\hat{H}^p - p\gamma \hat{H}^{p-1} + \cdots + (-1)^p \gamma^p I) \Delta t^p. \end{aligned} \tag{22}$$

According to (22), the coefficient of $\hat{H}^i$, for $i = 0, 1, ..., p$, is given as

$$\sigma_i \Delta t^i - \sigma_{i+1} \binom{i+1}{i} \gamma \Delta t^{i+1} + \sigma_{i+2} \binom{i+2}{i} \gamma^2 \Delta t^{i+2} + \cdots + (-1)^{p-i} \sigma_p \binom{p}{i} \gamma^{p-i} \Delta t^p,$$

which is the same as (20).

We note that $\sum_{i=0}^{p}\gamma^i f_i(\Delta t)$ is equivalent to replacing $I$ and $\hat{H}$ by 1 and $\gamma$, respectively, in (19). Then, according to (22), we have

$$\sum_{i=0}^{p}\gamma^i f_i(\Delta t)=\sum_{i=0}^{p}\alpha_i(\gamma\Delta t)^i\sum_{k=0}^{i}(-1)^k\binom{i}{k}. \tag{23}$$

For $i>0$, we have $\sum_{k=0}^{i}(-1)^k\binom{i}{k}=(x-1)^i|_{x=1}=0$, implying that the right hand side of (23) equals to 1, thus (21) follows immediately.

The proof for Part b) is the same as the one presented for Part b) in Lemma 3.2, thus we are not presenting here. □

3.1.2. *Compute Steplength Threshold.* We now consider the value of $\tau$, i.e., the steplength threshold. In this section, we present an algorithm to numerically compute $\tau$. In particular, this algorithm aims to find the first positive zero of a polynomial function in the form of (17).

**Lemma 3.5.** *Let $f(\Delta t)$ be given as in (17). There exists a $\tau^*>0$, such that $f(\Delta t)\geq 0$ for all $\Delta t\in[0,\tau^*]$.*

*Proof.* Since $f(0)=1>0$, and $f(\Delta t)$ is a continuous function, the lemma is immediate. □

Let $f(\Delta t)$ be given as in (17). If $\alpha_1,\alpha_2,...,\alpha_q\geq 0$, then $f(\Delta t)\geq 0$ for all $\Delta t\geq 0$, which implies $\tau^*=\infty$ in Lemma 3.5. Also, since $f(\Delta t)$ is dominated by $\alpha_q\Delta t^q$ for $\Delta t\gg 1$, we have that $\tau^*=\infty$ implies $\alpha_q>0$. Therefore, the largest $\tau^*$ that satisfies Lemma 3.5 is the first positive zero of $f(\Delta t)$, otherwise, we have $\tau^*=\infty$. In fact, we can find a predicted large $t^*>0$, such that if there is no zeros of $f(\Delta t)$ in $[0,t^*]$, then we have $\tau^*=\infty$. Note that this case only occurs when $\alpha_q\Delta t^q$ dominates $f(\Delta t)$. This is presented in the following lemma.

**Lemma 3.6.** *Let $f(\Delta t)$ be given as in (17) and $\alpha_q>0$. Let $\alpha^*=\max\{1,|\alpha_1|,|\alpha_2|,...,|\alpha_{q-1}|\}$ and $t^*=\frac{\alpha^*}{\alpha_q}+1$. Then if $f(\Delta t)$ has no real zero in $[0,t^*]$, then $f(\Delta t)>0$ for all $\Delta t>0$.*

*Proof.* Since $f(\Delta t)$ has no real zero in $[0,t^*]$, we have $f(\Delta t)>0$ on $[0,t^*]$. Thus, we only need to prove the following holds:

$$\alpha_q\Delta t^q>|1+\alpha_1\Delta t+\alpha_2\Delta t^2+...+\alpha_{q-1}\Delta t^{q-1}|,\ \text{for all } \Delta t\in(t^*,\infty].$$

Note that $t^*=\frac{\alpha^*}{\alpha_q}+1$ implies $\alpha_q=\frac{\alpha^*}{t^*-1}>\frac{\alpha^*}{\Delta t-1}$ for all $\Delta t\in(t^*,\infty]$. Then we have

$$\begin{aligned}|1+\alpha_1\Delta t+\alpha_2\Delta t^2+...+\alpha_{q-1}\Delta t^{q-1}|&\leq\alpha^*(1+\Delta t+\Delta t^2+...+\Delta t^{q-1})\\&=\alpha^*\frac{\Delta t^q-1}{\Delta t-1}<\alpha_q(\Delta t^q-1)<\alpha_q\Delta t^q.\end{aligned}$$

The proof is complete. □

In fact, the value $t^*$ given in Lemma 3.6 can be considered as one of the termination criteria of the algorithm to find the first positive zero of $f(\Delta t)$, where $f(\Delta t)$ is defined as (17).

The Sturm sequence $\{s_i(t)\}$ of $f(t)$ and the Sturm Theorem presented in the following definition play a key role in our algorithm. The Sturm Theorem aims to give the number of real zeros of a univariate polynomial function in an interval by using the property of Sturm sequence on the end points of the interval.

**Definition 3.7.** [22] Let $f(t)$ be a univariate polynomial function. The **Sturm sequence** $\{s_i(t)\}, i = 1, 2, ...,$ of $f(t)$ is defined as

$$s_0(t) = f(t),\ s_1(t) = s'(t),\ s_i(t) = -rem(s_{i-2}(t), s_{i-1}(t)),\ i \geq 2,$$

where $s'(t)$ is the derivative of $s(t)$ with respect to $t$, and $s_i(t)$ is the negative of the remainder on division of $s_{i-2}(t)$ by $s_{i-1}(t)$.

For the sake of simplicity, we introduce the following definition and notation, which are used in the statement of the Sturm Theorem.

**Definition 3.8.** For a sequence $\{\nu_i\}$, $i = 1, 2, ..., q$, the **number of sign changes**, denoted by $\#\{\nu_i\}$, is the number of the times of the signs change (zeros are ignored) from $\nu_1$ to $\nu_q$.

For example, if a sequence is given as $\{\nu_i\} = \{1, 0, 3, -2, 0, 2, -1, 0, -3\}$, then the signs of the sequence are $\{+, 0, +, -, 0, +, -, 0, -\}$. By eliminating all zeros, we have $\{+, +, -, +, -, -\}$, which has 3 sign changes, i.e., $\#\{\nu_i\} = 3$.

**Theorem 3.9.** [22] ***(Sturm Theorem)*** *Let $f(t)$ be a univariate polynomial function. If $\alpha < \beta$ and $f(\alpha), f(\beta) \neq 0$. Then the number of distinct real zeros of $f(t)$ in the interval $[\alpha, \beta]$ is equal to $|\#\{s_i(\alpha)\} - \#\{s_i(\beta)\}|$, where $\{s_i(t)\}$ is the Sturm sequence of $f(t)$.*

According to Lemma 3.6 and Theorem 3.9, we now propose our algorithm to numerically find the first positive zero of $f(\Delta t)$ where $f(\Delta t)$ is defined as (17). Let us denote $\#f[\delta]$ the number of positive zeros of $f(\Delta t)$ at interval $[0, \delta]$. The value of $\#f[\delta]$ can be computed by Sturm Theorem 3.9. The basic idea in our algorithm is by using the bisection method to shrink the interval, which contains the first positive zero of $f(t)$, by 2 in each iteration. Our algorithm is presented as follows.

**Step 0: [Initial Inputs]:** Set $t^\circ = 1$. Iterate $t^\circ = \frac{t^\circ}{2}$ until $\#f[t^\circ] = 0$. Let $t^*$ be given as in Lemma 3.6.
**Step 1: [Initial Setting]:** Set $t_l = t^\circ$, $t_r = t^*$, and $\epsilon$ be the precision.
**Step 2: [Termination 1]:** If $\#f[t_r] = 0$, then $\tau = \infty$.
**Step 3: [Termination 2]:** If $\#f[t_r] = 1$ and $f(t_r) = 0$, then $\tau = t^*$.
**Step 4: [Bisection Method]:** Set $t_m = \frac{t_l + t_r}{2}$.
Repeat until $|t_l - t_r| < \epsilon$:
- **[Termination 3]** If $\#f[t_m] = 1$ and $f(t_m) = 0$, then $\tau = t_m$.
- **[Update $t_r$]** If $\#f[t_m] = 1$ and $f(t_m) \neq 0$, or $\#f[t_m] > 1$, then set $t_r = t_m$.
- **[Update $t_l$]** If $\#f[t_m] = 0$, then set $t_l = t_m$.

End
**Step 5: [Termination 4]:** If Step 4 is terminated at $|t_l - t_r| < \epsilon$, then $\tau = t_l$.

The correctness of the termination condition in Step 2 is ensured by Lemma 3.6. If neither of the termination conditions in Step 2 and 3 are satisfied, then it means that the first positive zero of $f(t)$ exists and is located in the interval $(t_l, t_r)$. The second case in Step 4 means that the first positive zero of $f(t)$ is located in the interval $(t_l, t_m)$. Analogously, the third case in Step 4 means that the first positive zero of $f(t)$ is located in the interval $(t_m, t_r)$. In Step 5, we conclude that the first positive zero of $f(t)$ is located in the interval $(t_l, t_r)$. Recall that we are interested to find a value $\tau$, such that $f(t) \geq 0$ for all $[0, \tau]$, thus we return $t_l$, i.e., the left end of the interval.

**Remark 1.** If all coefficients $\sigma_i \geq 0$ for $i = 1, 2, ..., p$ in (18), then the algorithm is also applicable to compute a valid steplength threshold for invariance preserving for the polynomial approximation (18).

3.2. **Rational Function Type Discretization Methods.** The previous discussion is mainly about a steplength threshold for invariance preserving for a Taylor approximation type discretization methods as specified in (6). In this section, we consider more general discretization methods, which are refereed to as the rational function type discretization methods. To be specific, these discretization methods applying to the continuous system yield the discrete system

$$x_{k+1} = r(A_c\Delta t)x_k, \tag{24}$$

where $r(t) : \mathbb{R} \to \mathbb{R}$ is a rational function defined as

$$r(t) = \frac{g(t)}{h(t)} = \frac{\lambda_0 + \lambda_1 t + \cdots + \lambda_p t^p}{\mu_0 + \mu_1 t + \cdots + \mu_q t^q}, \tag{25}$$

where $\lambda_0, \lambda_1, ..., \lambda_p \in \mathbb{R}$, $\mu_0, \mu_1, ..., \mu_q \in \mathbb{R}$, and $p, q \in \mathbb{N}$. It is clear that Taylor approximation type discretization methods belong to this type. Some implicit methods are also in this type, e.g., the backward Euler method, Lobatto methods [10], etc.

**Definition 3.10.** [11] Let $r(t)$ be given as in (25), and let $M$ be a matrix. Assume $h(M)$ is nonsingular, then

$$r(M) := (h(M))^{-1}g(M) = g(M)(h(M))^{-1}. \tag{26}$$

3.2.1. *Existence of Steplength Threshold.* In this subsection, our analysis uses the so called *radius of absolute monotonicity* of a function.

**Definition 3.11.** [20] Let $r(t) : \mathbb{R} \to \mathbb{R}$. If $\rho = \max\{\kappa \,|\, r^{(i)}(t) \geq 0 \text{ for all } i = 1, 2, ..., \text{ and } t \in [-\kappa, 0]\}$, where $r^{(i)}(t)$ is the $i^{th}$ derivative of $r(t)$, then $\rho$ is called the **radius of absolute monotonicity** of $r(t)$.

The radius of absolute monotonicity of a function is extensively used in the analysis of positivity, monotonicity, and contractivity of discretization methods for ordinary differential equations, see e.g., [13, 17, 20].

**Theorem 3.12.** *Assume $r(t)$ is a rational function with $r(0) = 1$. Let $\rho$ be the radius of absolute monotonicity of $r(t)$. Assume a polyhedron $\mathcal{P}$ be given as in (3) is an invariant set for the continuous system (1), and the rational function type discretization method given as in (24) is applied to the continuous system (1). Then $\tau = \frac{\rho}{\gamma}$, where $\gamma$ is given in Lemma 3.1, is a valid steplength threshold for invariance preserving of the rational function type discretization method given as in (24) on $\mathcal{P}$.*

*Proof.* The framework of this proof is similar to the one presented for Lemma 3.2. Since $\mathcal{P}$ is an invariant set for the continuous system, according to Theorem 2.3 and Lemma 3.1, there exists an $H$, and $\gamma > 0$, such that

$$H + \gamma I \geq 0, \; HG = GA_c, \text{ and } Hb \leq 0. \tag{27}$$

Then, according to Theorem 2.3, to ensure $\mathcal{P}$ is also an invariant set for the discrete system, we need to prove that there exists an $\tilde{H}(\Delta t) \in \mathbb{R}^{m\times m}$, such that

$$\tilde{H}(\Delta t) \geq 0, \; \tilde{H}(\Delta t)G = Gr(A_c\Delta t), \text{ and } \tilde{H}(\Delta t)b \leq b. \tag{28}$$

Let $\tilde{H}(\Delta t) = r(H\Delta t)$. Now we prove that $\tilde{H}(\Delta t)$ satisfies (28).

For the first condition in (28), we use the Taylor expansion of $r(t)$ at the value $-\rho$ as

$$r(t) = \sum_{i=0}^{\infty} \frac{r^{(i)}(-\rho)}{i!}(t+\rho)^i. \tag{29}$$

By substituting $t = H\Delta t$ into (29) we have

$$\tilde{H}(\Delta t) = r(H\Delta t) = \sum_{i=0}^{\infty} \frac{r^{(i)}(-\rho)}{i!}(H\Delta t + \rho I)^i = \sum_{i=0}^{\infty} \frac{r^{(i)}(-\rho)}{i!}(\Delta t)^i \left(H + \frac{\rho}{\Delta t} I\right)^i. \tag{30}$$

Since $\rho$ is the radius of absolute monotonicity of $r(t)$, we have $\frac{r^{(i)}(-\rho)}{i!} \geq 0$ for all $i$. Also, according to (27), and $\Delta t \leq \frac{\rho}{\gamma}$, i.e., $\frac{\rho}{\Delta t} \geq \gamma$, so we have $H + \frac{\rho}{\Delta t} I \geq H + \gamma I \geq 0$. Then we have $(H + \frac{\rho}{\Delta t} I)^i \geq 0$ for all $i$. According to (30), we have $\tilde{H}(\Delta t) \geq 0$ for $\Delta t \leq \frac{\rho}{\gamma}$, thus the first condition in (28) is satisfied.

For the second condition in (28), according to Definition 3.10, the second condition in (28) can be rewritten as $(h(H\Delta t))^{-1} g(H\Delta t) G = G g(A_c \Delta t)(h(A_c \Delta t))^{-1}$, i.e.,

$$g(H\Delta t) G h(A_c \Delta t) = h(H\Delta t) G g(A_c \Delta t). \tag{31}$$

According to (25), we have

$$\begin{aligned} h(H\Delta t) G g(A_c \Delta t) &= \sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_i \mu_j H^i G H^j \Delta t^{i+j}, \\ g(H\Delta t) G h(A_c \Delta t) &= \sum_{j=1}^{q} \sum_{i=1}^{p} \lambda_i \mu_j H^j G H^i \Delta t^{i+j}. \end{aligned} \tag{32}$$

By recursively using $HG = GA_c$, for any $i, j$, we have

$$H^i G A_c^j = G A_c^{i+j} = H^{i+j} G = H^j G A_c^i. \tag{33}$$

According to (32) and (33), we have that (31) is true, i.e., the second condition (28) is satisfied.

For the third condition in (28) we have

$$\begin{aligned} \tilde{H}(\Delta t) b = r(H\Delta t) b &= \sum_{i=0}^{\infty} \frac{r^{(i)}(-\rho)}{i!}(H\Delta t + \rho I)^i b \\ &= \sum_{i=0}^{\infty} \frac{r^{(i)}(-\rho)}{i!}(H\Delta t + \rho I)^{i-1}(H\Delta t + \rho I) b \\ &\leq \sum_{i=0}^{\infty} \frac{r^{(i)}(-\rho)}{i!}(H\Delta t + \rho I)^{i-1} \rho b \leq \sum_{i=0}^{\infty} \frac{r^{(i)}(-\rho)}{i!} \rho^i b = r(0) b = b. \end{aligned}$$

Thus, the third condition in (28) is also satisfied. The proof is complete. □

The assumption $r(0) = 1$ in Theorem 3.12 is a fundamental condition for most discretization methods. This is since the steplength $\Delta t = 0$, yielding that the coefficient matrix of the discrete system is the identity matrix.

3.2.2. *Compute Steplength Threshold.* The steplength threshold given in Theorem 3.12 is related to $\rho$ and $\gamma$. Recall that $\gamma$ is given in Lemma 3.1, thus we only consider the computation of $\rho$.

Since $r(t)$ is a rational function, all of its derivatives $r^{(i)}(t)$ have the same format, i.e., they are represented as quotients of two polynomial functions. Now recall that the radius of absolute monotonicity $\rho$ is defined as $r^{(i)}(t) \geq 0$ for $t \in [-\rho, 0]$. This requires that the polynomial function in the numerator of $r^{(i)}(t)$ is nonnegative for $t \in [-\rho, 0]$. Thus, a valid $\rho$ is the negative of the first negative real zero of this polynomial function. Then an algorithm similar to the one presented in Section 3.1.2 can be proposed to numerically compute $\rho$. We are not repressing the algorithm here due to the space consideration.

3.3. **Parameter of Steplength Threshold.** According to Theorem 3.3 and Theorem 3.12, we have that the parameter $\gamma$ plays an important role to derive a large valid steplength threshold. In this section, we consider the effect of $\gamma$ to the steplength threshold.

3.3.1. *Best Parameter.* Let us first consider the case for Taylor approximation type discretization methods. By simple modification, we have that $f_i(\Delta t)$ defined in (9) can be written as

$$f_i(\Delta t) = \Delta t^i \sum_{k=i}^{p} \frac{(-1)^{k-i}}{k!} \binom{k}{i} (\gamma \Delta t)^{k-i}, \text{ for } i = 0, 1, ..., p, \tag{34}$$

which means that smaller $\gamma$ will yield larger steplength threshold for Taylor type discretization method given as in (6). Similarly, according to Theorem 3.12, we also have that smaller $\gamma$ will yield larger steplength threshold for the rational function type discretization methods (24). Thus we prefer the smallest possible $\gamma$, which in fact can be computed by solving the following optimization problem

$$\min\{\gamma \,|\, H + \gamma I \geq 0, \ HG = GA_c, \text{ and } Hb \leq 0\}. \tag{35}$$

In optimization problem (35), the variables are $H$ and $\gamma$, while $G, A_c$ and $b$ are known, thus problem (35) is a linear optimization problem, which can be easily solved by existing optimization algorithms, e.g., simplex methods [1] or interior point methods [19]. In particular, if there exists an $H \geq 0$ such that $HG = GA_c$ and $Hb \leq 0$, then the optimal solution, denoted by $\gamma^*$, of (35) is nonpositive. In this case, according to (34), we have $f_i(\Delta t) \geq 0$ for all $\Delta t \geq 0$. Then according to the proof of Lemma 3.2, we have that the steplength threshold for invariance preserving for Taylor approximation type discretization methods (6) on polyhedron $\mathcal{P}$ is infinity. Similarly, if $\gamma^* \leq 0$, according to Theorem 3.12, we have that the steplength threshold for invariance preserving for rational function type discretization methods (24) on polyhedron $\mathcal{P}$ is also infinity. Thus, we have the following theorem.

**Theorem 3.13.** *If the optimal solution of (35) is nonpositive, then the steplength threshold for invariance preserving on the polyhedron $\mathcal{P}$ is infinity for Taylor approximation type discretization methods (6) and rational function type discretization methods (24).*

One should note that the steplength thresholds given in Theorem 3.3 and Theorem 3.12 may not be the largest steplength thresholds. For example, for the

Taylor approximation type discretization methods, we aim to find the first positive zeros of finitely many polynomial functions. In fact, the first positive zeros may not be the best in some cases. For example, if the function is given as $f(\Delta t) = (\Delta t - 1)^2(\Delta t - 2)^2$, then its first positive zero is 1. Then, by our methods, we have $\tau = 1$. However, it is clear that $f(\Delta t) \geq 0$ for any $\Delta t \geq 0$. Thus, in this case, we have $\tau = \infty$.

If the first zero, $\Delta t^*$, of a function is a local minimum of this function, i.e., $f'(\Delta t^*) = 0$, then the first zero should not be used for computing the steplength threshold. This is since the function is tangent to the $x$ axis at the first zero. To verify if a zero is a local minimum, one can check the first order and second order directives $f'(\Delta t^*)$ and $f''(\Delta t^*)$. If $f(\Delta t^*) = 0$ and $f'(\Delta t^*) < 0$, then we can say that $\Delta t^*$ is not a local minimum, and thus it is a valid positive zero. If $f(\Delta t^*) = 0, f'(\Delta t^*) = 0$, and $f''(\Delta t^*) > 0$, we can say that $\Delta t^*$ is a local minimum. Then we have to make $\Delta t$ to be larger, and use an algorithm similar to the one presented in Section 3.1.2 to find the next zero of $f(\Delta t)$.

3.3.2. *Relation to the Forward Euler Method.* The following lemma presents the relationship between $\gamma$ that satisfies the constraints in (35) and the operator $I + \gamma^{-1}A_c$ on $\mathcal{P}$. Recall that $I + \Delta t A_c$ is the coefficient matrix of the discrete system by using the forward Euler method.

**Lemma 3.14.** *The conditions $H + \gamma I \geq 0, HG = GA_c$, and $Hb \leq 0$ are satisfied if and only if $(I + \gamma^{-1}A_c)\mathcal{P} \subseteq \mathcal{P}$.*

*Proof.* " $\Rightarrow$ " For $x \in \mathcal{P}$, i.e., $Gx \leq b$, we have

$$\begin{aligned} G(I + \gamma^{-1}A_c)x &= Gx + \gamma^{-1}GA_cx \\ &= Gx + \gamma^{-1}HGx \ \leftarrow \text{since } HG = GA_c \\ &= \gamma^{-1}(H + \gamma I)Gx \\ &\leq \gamma^{-1}(H + \gamma I)b \ \leftarrow \text{since } Gx \leq b \text{ and } H + \gamma I \geq 0 \\ &= b + \gamma^{-1}Hb \leq b \ \leftarrow \text{since } Hb \leq 0. \end{aligned}$$

Thus we have $(I + \gamma^{-1}A_c)x \in \mathcal{P}$, i.e., $(I + \gamma^{-1}A_c)\mathcal{P} \subseteq \mathcal{P}$.

" $\Leftarrow$ " We note that $(I + \gamma^{-1}A_c)\mathcal{P} \subseteq \mathcal{P}$ means that $\mathcal{P}$ is an invariant set for the following discrete system:

$$x_{k+1} = (I + \gamma^{-1}A_c)x_k.$$

Then according to Theorem 2.3, we have that there exists an $\tilde{H} \in \mathbb{R}^{m\times m}$, such that $\tilde{H} \geq 0, \tilde{H}G = G(I + \gamma^{-1}A_c)$, and $\tilde{H}b \leq b$. Let $\hat{H} = \gamma\tilde{H}$, and then we have

$$\hat{H} \geq 0, \ \hat{H}G = G(\gamma I + A_c), \text{ and } \hat{H}b \leq \gamma b,$$

i.e.,

$$(\hat{H} - \gamma I) + \gamma I \geq 0, \ (\hat{H} - \gamma I)G = GA_c, \text{ and } (\hat{H} - \gamma I)b \leq 0.$$

Thus replacing $\hat{H} - \gamma I$ by $H$, the proof is complete. □

We highlight that the forward Euler method is used to analyze invariance in continuous dynamical systems in [4, 5]. In [4], the largest domain of attraction of a continuous dynamical system is approximated with arbitrarily precision by using a polyhedral domain of attraction of a discrete dynamical system. This discrete dynamical system is obtained by the forward Euler method and referred to as Euler approximating system in [4]. The value of $\gamma^{-1}$ in Lemma 3.14 can be considered as

the step size of the forward Euler method for preserving the invariance of polyhedral $\mathcal{P}$, and the value of $\gamma$ is easily quantified. The existence of a step size for preserving the contractivity of a set is also presented in [4] for the forward Euler method. A similar result to Lemma 3.14 is presented in [5], which is an extension of [8], for (A,B)-invariance condition. The forward Euler method is also applied to build the connection between continuous and discrete dynamical systems. The value of the step size of the forward Euler method in [5] for (A,B)-invariance condition is computed in a similar way to the one given as in Lemma 3.14.

3.4. **Forward Euler Method.** As illustration, we consider the simplest discretization method, the forward Euler method, in this section. For simplicity, a polytope, i.e., a bounded polyhedron, is chosen as the invariant set for the forward Euler method. A polytope can be defined in terms of convex combination of its vetices, i.e.,

$$\mathcal{P} = \text{conv}\{x^1, x^2, ..., x^\ell\} = \Big\{x \,|\, x = \sum_{i=1}^{\ell} \lambda_i x^i,\ \sum_{i=1}^{\ell} \lambda_i = 1,\ \lambda_i \geq 0\Big\}, \tag{36}$$

where $\{x^i\}$ are the vertices of $\mathcal{P}$. A sufficient and necessary condition under which a polytope is an invariant set for the continuous system is presented below.

**Lemma 3.15.** [16] *The polytope $\mathcal{P}$ defined as in (36) is an invariant set for the continuous system (1) if and only if $A_c x^i \in \mathcal{T}_{\mathcal{P}}(x^i)$, for $i = 1, 2, ..., \ell$, where $\mathcal{T}_{\mathcal{P}}(x^i)$ is the tangent cone[3] at $x^i$, which can be given*

$$\mathcal{T}_{\mathcal{P}}(x^i) = \{y \,|\, y = \sum_{j \neq i} \gamma_j (x^j - x^i),\ \gamma_j \geq 0\}. \tag{37}$$

**Corollary 2.** *The polyhedron $\mathcal{P}$ defined as in (36) is an invariant set for the continuous system (1) if and only if there exist $\gamma_j^{(i)} \geq 0, j = 1, 2, ..., \ell$, such that*

$$A_c x^i = \sum_{j \neq i} \gamma_j^{(i)} (x^j - x^i),\ \textit{for all } i = 1, 2, ..., \ell. \tag{38}$$

*Let $\epsilon^i = \left(\sum_{j \neq i} \gamma_j^{(i)}\right)^{-1}$ for $i = 1, 2, ..., \ell$, then*

$$x^i + \Delta t A_c x^i \in \mathcal{P} \textit{ for any } \Delta t \in [0, \epsilon^i]. \tag{39}$$

*Proof.* According to Lemma 3.15 and equation (37), equation (38) is immediate. According to (38) and $\epsilon^i \sum_{j \neq i} w_j^{(i)} = 1$, we have

$$\epsilon^i A_c x^i = \sum_{j \neq i} \epsilon^i \gamma_j^{(i)} (x^j - x^i) = \sum_{j \neq i} \epsilon^i \gamma_j^{(i)} x^j - \sum_{j \neq i} \epsilon^i \gamma_j^{(i)} x^i = \sum_{j \neq i} \epsilon^i \gamma_j^{(i)} x^j - x^i. \tag{40}$$

According to (40), we have $x^i + \epsilon^i A_c x^i = \sum_{j \neq i} \epsilon^i \gamma_j^{(i)} x^j$, which is a convex combination of $\{x^j\}$, thus $x^i + \epsilon^i A_c x^i \in \mathcal{P}$. For any $\Delta t \in [0, \epsilon^i]$, by the convexity of $\mathcal{P}$, we have

$$x^i + \Delta t A_c x^i = \frac{\Delta t}{\epsilon^i}(x_i + \epsilon^i A_c x^i) + \frac{\epsilon^i - \Delta t}{\epsilon^i} x^i \in \mathcal{P},$$

which completes the proof. □

[3] The tangent cone of a set $\mathcal{S}$ at $x$, denoted by $\mathcal{T}_{\mathcal{S}}(x)$, is given as $\mathcal{T}_{\mathcal{S}}(x) = \{y \in \mathbb{R}^n \mid \liminf_{t \to 0_+} \frac{\text{dist}(x+ty, \mathcal{S})}{t} = 0\}$, where $\text{dist}(x, \mathcal{S}) = \inf_{s \in \mathcal{S}} \|x - s\|$.

We now consider the calculation of $\epsilon^i$, where $\epsilon^i$ is defined as in Corollary 2. By the formula of $\epsilon^i$, we need to compute $\gamma_j^{(i)}, j = 1, 2, ..., \ell$, such that (38) is satisfied. In fact, this can be achieved by solving the following optimization problem:

$$\min \Big\{ \sum_{j\neq i} \gamma_j^{(i)} \mid \sum_{j\neq i} \gamma_j^{(i)} (x^j - x^i) = A_c x^i,\ \gamma_j^{(i)} \geq 0. \Big\} \tag{41}$$

Since $x^1, x^2, ..., x^k$, and $A_c$ are known, optimization problem (41) is a linear optimization problem. One may obtain different values of $\hat{\gamma}_j^{(i)}, j = 1, 2, ..., \ell$, by choosing other objective functions in (41). The advantage by using the current objective function in (41) is that this optimization problem yields the largest $\epsilon^i$ that satisfies (38). This is since the objective function in (41) is $(\epsilon^i)^{-1}$. Thus, the value of $\epsilon^i$ obtained by solving the optimization problem (41) is the largest possible value of $\epsilon^i$.

An alternative is presented by the following discussion. Equation (37) implies that $Ax^i$ is a feasible direction, i.e., $x^i + \tau^i A_c x^i \in \mathcal{P}$, for sufficiently small $\tau^i > 0$. Then we can formulate the following linear optimization problem:

$$\max \Big\{ \tau^i \mid \sum_{j=1}^{\ell} u_j^{(i)} x^j = x^i + \tau^i A_c x^i,\ \sum_{j=1}^{\ell} u_j^{(i)} = 1,\ u_j^{(i)} \geq 0 \Big\}. \tag{42}$$

Optimization problems (41) and (42) are equivalent problems, i.e., we claim that $\tau^i$ is equal to $\epsilon^i$. Observing that $\sum_{j=1}^{n} \beta_j^{(i)} = 1$ for the first constraint in (42), we have

$$\tau^i A_c x^i = \sum_{j=1}^{\ell} u_j^{(i)} x^j - \sum_{j=1}^{\ell} u_j^{(i)} x^i = \sum_{j=1}^{\ell} \tau^i \frac{u_j^{(i)}}{\tau^i} x^j - \sum_{j=1}^{\ell} \tau^i \frac{u_j^{(i)}}{\tau^i} x^i = \tau^i \sum_{j=1}^{\ell} \frac{u_j^{(i)}}{\tau^i} (x^j - x^i), \tag{43}$$

i.e., $A_c x^i = \sum_{j\neq i} \frac{u_j^{(i)}}{\tau^i} (x^j - x^i)$. This, by letting $\frac{u_j^{(i)}}{\tau^i} = \gamma_j^{(i)}$ gives the first constraint in (41).

According to the argument for $\epsilon^i$ above, we have the following theorem.

**Theorem 3.16.** *Assume that the polytope $\mathcal{P}$ defined as in (36) is an invariant set for the continuous system (1), and the forward Euler method is applied to (1). Then, $\tau = \min_{i=1,2,...,\ell}\{\epsilon^i\}$, where $\epsilon^i$ is defined as in Corollary 2, is the largest steplength threshold $\tau > 0$ for invariance preserving of the forward Euler method on $\mathcal{P}$.*

*Proof.* For any $x \in \mathcal{P}$, and $\Delta t \in [0, \tau]$, we have $x + \Delta t A_c x = \sum_{i=1}^{\ell} \lambda_i (x^i + \Delta t A_c x^i)$. According to Corollary 2 and $0 \leq \Delta t \leq \tau \leq \epsilon^i$, we have $x^i + \Delta t A_c x^i \in \mathcal{P}$. Thus we have $x + \Delta t A_c x \in \mathcal{P}$. The proof is complete. □

4. **Conclusions.** Many real world problems are studied by developing dynamical system models. In practice, continuous systems are usually solved by using discretization methods. In this paper, we consider invariance preserving steplength thresholds on polyhedron, when the discrete system is obtained by using special classes of discretization methods. We particularly study three classes of discretization methods, which are: Taylor approximation type, rational function type, and the forward Euler method.

For the first class of discretization methods, we show that a valid steplength threshold can be obtained by finding the first positive zeros of a finite number of

polynomial functions. We also present a simple and efficient algorithm to numerically compute these positive zeros. For the second class of discretization methods, a valid steplength threshold for invariance preserving is presented. This steplength threshold depends on the radius of absolute monotonicity, and can be computed by analogous method as in the first case. For the forward Euler method we prove that the largest steplength threshold can be obtained by solving a finite number of linear optimization problems.

**Acknowledgments.** This research is supported by a Start-up grant of Lehigh University and by TAMOP-4.2.2.A-11/1KONV-2012-0012: Basic research for the development of hybrid and electric vehicles. The TAMOP Project is supported by the European Union and co-financed by the European Regional Development Fund.

## REFERENCES

[1] D. Bertsimas and J. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Nashua, 1998.
[2] G. Bitsoris, On the positive invariance of polyhedral sets for discrete-time systems, *System and Control Letters*, **11** (1998), 243-248.
[3] F. Blanchini, Set invariance in control, *Automatica*, **35** (1999), 1747-1767.
[4] F. Blanchini and S. Miani, Constrained stabilization of continuous-time linear systems, *Systems and Control Letters*, **29** (1996), 95-102.
[5] F. Blanchini, S. Miani, C.E.T. Dórea and J.C. Hennet, Discussion on: '(A, B)- invariance conditions of polyhedral domains for continuous-time systems by C.E.T. Dórea and J.-C. Hennet', *European Journal of Control*, **5** (1999), 82-86.
[6] S. Boyd, L. El Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Studies in Applied Mathematics, Philadelphia, 1994.
[7] E.B. Castelan and J.C. Hennet, On invariant polyhedra of continuous-time linear systems, *IEEE Transactions on Automatic Control*, **38** (1993), 1680-1685.
[8] C.E.T. Dórea and J.C. Hennet, (A, B)-invariance conditions of polyhedral domains for continuous-time systems, *European Journal of Control*, **5** (1999), 70-81.
[9] C.E.T. Dórea and J.C. Hennet, (A,B)-invariant polyhedral sets of linear discrete time systems, *Journal of Optimization Theory and Applications*, **103** (1999), 521-542.
[10] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, Springer-Verlag, New York, 1993.
[11] N.J. Higham, *Functions of Matrices: Theory and Computation*, Society for Industrial and Applied Mathematics, Philadelphia, 2008.
[12] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1990.
[13] Z. Horváth, Invariant cones and polyhedra for dynamical systems, *Proceeding of the International Conference in Memoriam Gyula Farkas*, 2005, 65-74.
[14] Z. Horváth, On the positivity step size threshold of Runge-Kutta methods, *Applied Numerical Mathematics*, **33** (2005), 341-356.
[15] Z. Horváth, Y. Song and T. Terlaky, Invariance preserving discretization methods of dynamical systems, *Lehigh University, Department of Industrial and Systems Engineering, Technical Report 14T-009*, 2014.
[16] Z. Horváth, Y. Song and T. Terlaky, A novel unified approach to invariance in control, *Lehigh University, Department of Industrial and Systems Engineering, Technical Report 14T-003*, 2014.
[17] J.F.B.M. Kraaijevanger, Absolute monotonicity of polynomials occurring in the numerical solution of initial value problems, *Numerische Mathematik*, **48**, (1986), 303-322.
[18] R. Loewy and H. Schneider, Positive operators on the $n$-dimensional ice cream cone, *Journal of Mathematical Analysis and Applications*, **49** (1975), 375-392.
[19] C. Roos, T. Terlaky and J.-Ph. Vial, *Interior Point Methods for Linear Optimization*, Springer Science, Heidelberg, 2006.
[20] M.N. Spijker, Contractivity in the numerical solution of initial value problems, *Numerische Mathematik*, **42** (1983), 271-290.

[21] R. Stern and H. Wolkowicz, Exponential nonnegativity on the ice cream cone, *SIAM Journal on Matrix Analysis and Applications*, **12** (1991), 160-165.
[22] B. Sturmfels, *Solving Systems of Polynomial Equations*, CBMS Lectures Series, American Mathematical Society, 2002.
[23] J. Vandergraft, Spectral properties of matrices which have invariant cones, *SIAM Journal on Applied Mathematics*, **16** (1968), 1208-1222.

Received xxxx 20xx; revised xxxx 20xx.

*E-mail address*: horvathz@sze.hu
*E-mail address*: yus210@lehigh.edu
*E-mail address*: terlaky@lehigh.edu