Optimization Theory and Dynamical Systems: Invariant Sets and Invariance Preserving Discretization Methods

by

Yunfei Song

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

Industrial Engineering

Lehigh University

August 2015

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Date

_____
Dr. Tamás Terlaky, Dissertation Advisor

Committee Members:

_____
Dr. Tamás Terlaky, Committee Chair

_____
Dr. Frank E. Curtis

_____
Dr. Zoltán Horváth

_____
Dr. Mayuresh Kothare

_____
Dr. Sven Leyffer

_____
Dr. Katya Scheinberg

Acknowledgements

This dissertation would have been impossible to complete without the guidance of my advisor Dr. Tamás Terlaky. I would like to give my deep thank to him for his consistently insightful guidance through my overall Ph.D. process. Dr. Terlaky is not only an advisor for me, but also a mentor for many aspects in my life. I would also thank my Ph.D. committee members, Dr. Frank E. Curtis, Dr. Zoltán Horváth, Dr. Mayuresh Kothare, Dr. Sven Leyffer, and Dr. Katya Scheinberg for their valuable guidance and stimulating feedback. My fellow friends at Lehigh University deserve a special thanks for their discussions, supports, and kindness. I would like to thank Fang Chen, Julio Góez, Zheng Han, Dan Li, Murat Mut, Hao Wang, and many others. All of the ISE friends made my life at Lehigh so wonderful. I would like to thank Dr. József Bokor and Dr. Zoltán Horváth for their invitation for a short term visit to Széchenyi István University in Hungary. I also thank Tihamér A. Kocsis and Adrian Németh for the valuable discussions about my research. The comments and suggestions of Dr. Immanuel Bomze, Dr. Stefano Di Cairano, Dr. Piyush Grover, and Dr. David I. Ketcheson are greatly appreciated. Finally, I am deeply indebted to my dear parents for their love and for consistently supporting me and all of my decisions. They have both always been with me through the whole Ph.D. period, with its ups and downs. Most of the work presented in this dissertation was supported in part 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. Any opinions, findings, and conclusions or recommendations expressed in this dissertation are those of the author and do not necessarily reflect the views of the institutions above.

Contents

List of Figures

Notations and Symbols

Basic objects:

$A, B, ...$	matrices
A_i^T	the i-th row of the matrix A
A_{ij}	the i-th element of the j-th row of A
$v, u, ...$	vectors
$\alpha, \beta, ...$	parameters
$i, j, ...$	indices
I_n	the $n \times n$ identity matrix
$\tilde{I}$	the matrix diag$\{1, 1, ..., 1, -1\}$
e_i	the vector $(0, ..., 0, 1, 0, ..., 0)^T$ where the i-th element is 1
δ_{ij}	the Kronecker delta function, i.e, $\delta_{ij} = 1$ for $i = j$, and $\delta_{ij} = 0$ for $i \neq j$
$\lambda_i(A)$	the i-th (ordered from largest to smallest) eigenvalue of matrix A
λ_i	the i-th (ordered from largest to smallest) eigenvalue of a matrix for simplicity
$\lambda(A)$	the spectral radius of matrix A
inertia$\{A\}$	the inertia of matrix A
Δt	steplength of a discretization method
$\tau(x)$	local invariance preserving steplength threshold at the point x
$\tau_{\mathcal{S}}$	uniform invariance preserving steplength threshold on the set $\mathcal{S}$

Dynamical System:

t	time variable in a continuous dynamical system
$x(t)$	state variable in a continuous dynamical system
x_k	state variable in a discrete dynamical system

A_c coefficient matrix of a linear continuous dynamical system

A_d coefficient matrix of a linear discrete dynamical system

$\dot{x}(t)$ the derivative of $x(t)$

Sets:

$\mathbb{R}^n$ the real n-dimensional vector space

$\mathbb{R}^n_+$ the set of nonnegative vectors of $\mathbb{R}^n$

$\mathbb{R}^n_-$ the set of nonpositive vectors of $\mathbb{R}^n$, note that $\mathbb{R}^n_- = -\mathbb{R}^n_+$

$\mathcal{P}$ polyhedron

$\mathcal{S}$ convex set

$\mathcal{C}$ convex cone

$\mathcal{C}_\mathcal{P}$ polyhedral cone

$\mathcal{E}$ ellipsoid

$\mathcal{C}_\mathcal{L}$ Lorenz cone

$\mathcal{C}^*_\mathcal{L}$ standard Lorenz cone

$\mathcal{H}$ hyperplane

$\mathcal{C}^+$ base of a cone $\mathcal{C}$, note that $\mathcal{C}^+$ is bounded and $\mathcal{C}^+ = \mathcal{C} \cap \mathcal{H}$ for some hyperplane $\mathcal{H}$

x^i vertex of a polyhedron or polyhedral cone

$\hat{x}^i$ extreme ray of a polyhedral cone

$\mathcal{I}(n)$ index set $\{1, 2, ..., n\}$

$\mathcal{T}_\mathcal{S}(x)$ tangent cone of the set $\mathcal{S}$ at the point x

$\mathcal{I}_x$ index set of all constraints which are active at the point x

Relations:

$Q \succeq 0$ Q is a symmetric positive semidefinite matrix

$Q \succ 0$ Q is a symmetric positive definite matrix

$Q \preceq 0$ Q is a symmetric negative semidefinite matrix

$Q \prec 0$ Q is a symmetric negative definite matrix

$H \geq 0$ nonnegative matrix, i.e., $H_{ij} \geq 0$ for all i, j

$H \geq_o 0$ off-diagonal nonnegative matrix, i.e., $H_{ij} \geq 0$ for all $i \neq j$

Operators, functions:

$x^T y$ inner product of x and y

$\|x\|$	L_2-norm of the vector x
$\|M\|$	2-norm of the matrix M
$\text{int}(\mathcal{S})$	the interior of the set $\mathcal{S}$
$\partial\mathcal{S}$	the boundary of the set $\mathcal{S}$
$\text{aff}(\mathcal{S})$	the affine ball, i.e., the smallest affine subspace containing the set $\mathcal{S}$
$\text{cl}(\mathcal{S})$	the closure of the set $\mathcal{S}$
$\text{ri}(\mathcal{S})$	the set of all relative interior points of the set $\mathcal{S}$
$\text{rb}(\mathcal{S})$	the relative boundary of the set $\mathcal{S}$, i.e., $\text{rb}(\mathcal{S}) = \text{cl}(\mathcal{S})\backslash\text{ri}(\mathcal{S})$
$\text{dist}(x, \mathcal{C})$	the distance between the point x and the set $\mathcal{C}$

Abstract

Invariant set is an important concept in the theory of dynamical systems and it has a wide range of applications in control and engineering. This thesis has four parts, each of which studies a fundamental problem arising in this field. In the first part (Chapter 2), we propose a novel, simple, and unified approach to derive sufficient and necessary conditions, which are referred to as invariance conditions for simplicity, under which four classic families of convex sets, namely, polyhedra, polyhedral cones, ellipsoids, and Lorenz cones, are invariant sets for linear discrete or continuous dynamical systems. This novel method establishes a solid connection between optimization theory and dynamical systems. In the second part (Chapter 3), we propose novel methods to compute valid or largest uniform steplength thresholds for invariance preserving of three classic types of discretization methods, i.e., forward Euler method, Taylor type approximation, and rational function type discretization methods. These methods enable us to find a pre-specified steplength threshold which preserves invariance of a set. The identification of such steplength threshold has a significant impact in practice. In the third part (Chapter 4), we present a novel approach to ensure positive local and uniform steplength threshold for invariance preserving on a set when a discretization method is applied to a linear or nonlinear dynamical system. Our methodology not only applies to classic sets, discretization methods, and dynamical systems, but also extends to more general sets, discretization methods, and dynamical systems. In the fourth part (Chapter 5), we derive invariance conditions for some classic sets for nonlinear dynamical systems. This part can be considered as an extension of the first part to a more general case.

Chapter 1

Introduction

1.1 Background

Positively invariant sets play a key role in the theory and applications of dynamical systems. Stability, control and preservation of constraints of dynamical systems can be formulated, somehow in a geometrical way, with the help of positively invariant sets. For a given dynamical system, both of continuous or discrete time, a subset of the state space is called positively invariant set for the dynamical system if the system state at a certain time is in the invariant set, then forward in time all the states remain within the positively invariant set. Geometrically, the solution trajectories cannot escape from a positively invariant set if the initial state belongs to the set. The dynamical system is often a controlled system for which the maximal (or minimal) positively invariant set is to be constructed.

It is well known, see e.g., Blanchini [12], Blanchini and Miani [14], and Polanski [55], that Lyapunov stability theory is a powerful tool in obtaining many important results in control theory. The basic framework of Lyapunov stability theory synthesizes the identification and computation of a Lyapunov function of a dynamical system. Usually positive definite quadratic functions serve as candidate Lyapunov functions. Sufficient and necessary conditions for positive invariance of a polyhedral set with respect to discrete dynamical systems were first proposed by Bitsoris [8, 9]. A novel positively invariant polyhedral cone was constructed by Horváth [37]. The Riccati equation was proved to be connected with ellipsoidal sets as invariant sets of linear dynamical systems, see e.g., Lin et al. [48] and Zhou

et al. [86]. Birkhoff [7] proposed a necessary condition for positive invariance on a convex cone for linear discrete systems. A sufficient and necessary condition for positive invariance on a nontrivial convex set for linear discrete systems was derived by Elsner [24]. Stern [67] studied the properties of positive invariance on a proper cone for linear continuous systems. For a more general case, the mapping from a polyhedral cone to another polyhedral cone was studied by Haynsworth, Fiedler and Pták [30], and the mapping from a convex cone to another convex cone in finite-dimensional spaces was studied by Tam [71, 72]. Here we note that when the two cones are the same, then this is equivalent to positive invariance for discrete systems. The concept of cross positive matrices, which was introduced by Schneider and Vidyasagar [62], was used as tool to prove positive invariance of a Lorenz cone by Loewy and Schneider [49]. The existence and construction of common invariant cones for families of real matrices was studied by Rodman, Seyalioglu, and Spitkovsky [59]. Positively invariant sets with cone properties with respect to continuous systems were studied by Tarbouriech and Burgat [73]. According to Nagumo's theorem [52] and the theory of cross positive matrices, Stern and Wolkowicz [68] presented sufficient and necessary conditions for a Lorenz cone to be positively invariant with respect to a linear continuous system. A novel proof of the spectral characterization of real matrices that leave a polyhedral cone invariant was proposed by Valcher and Farina [76]. The spectral properties of the matrices, e.g., theorems of Perron-Frobenius type, were connected to set positive invariance by Vandergraft [62]. Approximating the minimal robust positively invariant set of an asymptotically stable discrete system was studied by Rakovic et al. [57]. For hyperchaotic Lorenz-Haken systems, Li et al. [47] investigated the estimation of ultimate bound and positively invariant set. For finding certain invariant sets of a given system, one may refer to Kouramas et al. [45], Pluymers et al. [54], Yu and Liao [80], Zhang et al. [81], etc. Zhao [85] derives a sufficient criteria for invariant sets and periodic attractors of non-autonomous systems.

Now we introduce the basic concepts related to invariant sets for dynamical systems.

Dynamical Systems: Discrete and continuous linear dynamical systems are respectively described by the following equations:

$$x_{k+1} = A_d x_k, \tag{1.1}$$

$$\dot{x}(t) = A_c x(t), \tag{1.2}$$

where $A_d, A_c \in \mathbb{R}^{n\times n}$ are constant real matrices, $x_k, x(t) \in \mathbb{R}^n$ are referred to as *state variables* for $k \in \mathbb{N}$, and $t \in \mathbb{R}$, respectively. We may assume, without loss of generality, that neither A_d nor A_c is the zero matrix.

Similarly, discrete and continuous autonomous dynamical systems in general forms are respectively described by the following equations:

$$x_{k+1} = f_d(x_k), \tag{1.3}$$

$$\dot{x}(t) = f_c(x(t)), \tag{1.4}$$

where $f_d, f_c : \mathbb{R}^n \to \mathbb{R}^n$ are continuous functions, and $x_k, x \in \mathbb{R}^n$ are referred to as *state variables*.

Invariant Sets: The study of invariant sets is one of the main subjects of in this thesis, thus now we introduce invariant sets, see, e.g., [13], for both discrete and continuous systems.

Definition 1.1.1. *A set $\mathcal{S} \subseteq \mathbb{R}^n$ is an invariant set for the discrete system (1.1) (or 1.3) if $x_k \in \mathcal{S}$ implies $x_{k+1} \in \mathcal{S}$, for all $k \in \mathbb{N}$.*

Definition 1.1.2. *A set $\mathcal{S} \subseteq \mathbb{R}^n$ is an invariant set for the continuous system (1.2) (or (1.4)) if $x(0) \in \mathcal{S}$ implies $x(t) \in \mathcal{S}$, for all $t \geq 0$.*

In fact, the sets given in Definition 1.1.1 and 1.1.2 are conventionally referred to as positively invariant sets. Since only positively invariant sets are considered in this thesis, we simply call them invariant sets. One can prove the following properties: the operators A (or[1] for all $t \geq 0$, e^{At}) leave $\mathcal{S}$ invariant if $\mathcal{S}$ is an invariant set for the discrete (or continuous) systems.

Proposition 1.1.3. [4, 17] *The set $\mathcal{S}$ is an invariant set for the discrete system (1.1) if and only if $A\mathcal{S} \subseteq \mathcal{S}$. Similarly, the set S is an invariant set for the continuous system (1.2) if and only if for all $t \geq 0$, $e^{At}\mathcal{S} \subseteq \mathcal{S}$.*

[1]The exponential function with respect to a matrix is defined as $e^{At} = \sum_{k=0}^{\infty} \frac{1}{k!} A^k t^k$.

Some well known examples of invariant sets are equilibrium point, limit cycle [53], etc. Here we point out that the concept of *stability* [44] is one of the most important properties of invariant sets. Intuitively, we say an invariant set to be stable for a dynamical system if any trajectories of the system starting close to the set remain close to it as time moves forward, and unstable if they do not. We say an invariant set is asymptotically stable for a dynamical system if it is stable and in addition any trajectories of the system starting close to the set converge to the set as $t \to \infty$.

1.2 Motivations

The motivations of this thesis are presented in this section. Although the definition of invariant sets may be used as the tool of verification if a set is indeed an invariant set of a given dynamical system, it is usually not efficient, or even impossible to directly use the definition in many cases. For example, the invariance of a set of a discrete system means that any point x_k in the set implies x_{k+1} is also in the set, then one has to verify this property for all points in the set. This would require the verification of infinitely many points, which is usually not an easy task. Therefore, to derive sufficient, or sufficient and necessary conditions for sets to be invariant sets of a dynamical system is important both from the theoretical and practical perspectives. In particular, we are interested in sufficient and necessary conditions under which a set is an invariant set for a dynamical system. Here we consider both continuous and discrete dynamical systems. A good verification condition has the following characteristics: simple and efficient to verify, i.e., one can easily and efficiently prove or disprove the invariance of a set for a dynamical system.

Numerous mathematical methods are developed to directly solve continuous systems, but, in practice, one usually needs to solve a continuous system by applying certain discretization methods. Assume that a set is an invariant set for a continuous system. Then it should also be an invariant set for the discrete system which is obtained by a discretization method, i.e., discretization should preserve invariance. However, this is not always true for every steplength used in the discretization method. Thus it is desirable to prove the existence of a steplength threshold which only depends on the set and the discretization

method, and all steplengths smaller than or equal to the threshold preserves the invariance of the set. This threshold is referred to as *invariance preserving steplength threshold*. Besides investigating the existence of such threshold, we are also interested in finding large invariance preserving steplength thresholds. This is because a large invariance preserving steplength has several advantages in practice, e.g., larger steplength implies that the discretized system is smaller in size. Finding a predictable invariance preserving steplength depends on properties of the dynamical system and the set. We are choosing some classic sets and linear dynamical systems, as well as using special discretization methods, to find the largest possible invariance preserving steplength threshold via establishing optimization models.

We are going to establish the theoretical foundation of invariance preserving, in which we focus on the existence of invariance preserving steplength threshold. We are interested in identifying general classes of sets, dynamical systems, and discretization methods with invariance preserving property. This is an important open question. In particular, we are going to investigate and find sufficient conditions, which ensure that when sets, dynamical systems, and discretization methods satisfy those conditions, then an invariance preserving steplength threshold exists. This research has significant theoretical impact, since we are considering the existence of uniform, i.e., global, rather than local invariance preserving steplength thresholds.

Set Invariance: Let us give an example of an invariant set: consider the normalized state space model of a double-integrator [13]:

$$x'(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t), \tag{1.5}$$

where $x_1(t), x_2(t)$ are the state variables and $u(t)$ is the control variable. Further, we assume that the state variables $x_1(t)$ and $x_2(t)$ satisfy the constraints $|x_1(t)| \leq 25$ and $|x_2(t)| \leq 5$ for all $t \geq 0$. Let $\mathcal{S} = \{(x_1, x_2) \mid |x_1| \leq 25, |x_2| \leq 5\}$. Now if we choose the linear state

feedback control law, see, e.g., [66], as follows: $u(t) = -\frac{1}{25}(x_1(t) + x_2(t))$, then we have

$$x'(t) = \begin{bmatrix} 0 & 1 \\ -\frac{1}{25} & -\frac{1}{25} \end{bmatrix} x(t). \tag{1.6}$$

The solution of (1.6) is given as

$$\begin{aligned} x_1(t) &= e^{-\frac{t}{50}} \left(\alpha \cos(\tfrac{3\sqrt{11}}{50} t) + \tfrac{\sqrt{11}}{33}(\alpha + 50\beta) \sin(\tfrac{3\sqrt{11}}{50} t) \right), \\ x_2(t) &= e^{-\frac{t}{50}} \left(\beta \cos(\tfrac{3\sqrt{11}}{50} t) - \tfrac{\sqrt{11}}{33}(2\alpha + \beta) \sin(\tfrac{3\sqrt{11}}{50} t) \right), \end{aligned} \tag{1.7}$$

where α and β depend on the initial state. If we set $x_1(0) = 25, x_2(0) = 5$, then we can show that $(x_1(t), x_2(t))$ will move outside of $\mathcal{S}$. Now we define an ellipsoid as follows:

$$\mathcal{E} = \{(x_1, x_2) \mid x_1^2 + 25x_2^2 \leq 1\}.$$

One can show that for any $(x_1(0), x_2(0)) \in \mathcal{E}$, we have $(x_1(t), x_2(t)) \in \mathcal{E}$ for all $t \geq 0$, i.e., $\mathcal{E}$ is an invariant set for system (1.6). Clearly, we have $\mathcal{E} \subset \mathcal{S}$, so the trajectory of system (1.6) cannot escape from $\mathcal{S}$ if $(x_1(t), x_2(t)) \in \mathcal{E}$. This means that to ensure that the state variable $x(t)$ is always in the feasible region $\mathcal{S}$, we can choose the initial state $x(0)$ from $\mathcal{E}$. The way of verifying that the ellipsoid is an invariant set requires to derive the solution of the system. However, if we can find an equivalent condition which only depends on the system and the set to verify whether the set is an invariant set for the system, then we do not need to solve the system. Such an equivalent condition, which is referred to as invariance condition, will significantly reduce the difficulty of verifying if a set is indeed an invariant set.

Invariance Preserving: In this thesis, we also study invariance preserving discretization methods, i.e., numerical methods which ensure that both the continuous and its discretized systems share the same invariant set. Let us consider the example presented at the beginning of this section, i.e., the double-integrator given as in (1.5). If we use a discretization method to solve the continuous system, then we hope that the ellipsoid is also an invariant set for the discrete system. This means that not only the continuous trajectories,

but also the discrete state variables stay in $\mathcal{S}$. If the discretization method cannot preserve invariance, then the ellipsoid is not an invariant set for the discrete system.

Let us give another example about invariance preserving, see, e.g., [38]. Consider a heat transformation with a fixed temperature T_b on the boundary of the heated body. Let us denote $T(t,x)$ the temperature at time t and position x. Then $\phi(t,x) = T(t,x) - T_b$ satisfies the following heat equation:

$$\phi_t(t,x) = \sigma\phi_{xx}(t,x), \tag{1.8}$$

where σ is the diffusion coefficient. A basic rule in thermodynamics is that the heat moves only from warmer position to colder position, and reverse direction move cannot occur. Let $T_{\min} = \min_x T(0,x)$ and $T_{\max} = \max_x T(0,x)$, then, according to the thermodynamics rule, we have $T_{\min} \leq T(t,x) \leq T_{\max}$ for all t. This yields that $\phi(t,x) \in [\phi_{\min}, \phi_{\max}]$, where $\phi_{\min} = T_{\min} - T_b$ and $\phi_{\max} = T_{\max} - T_b$, for all t. In practice, numerical methods that solve the heat equation require the discretization for both the spatial variable, x_k, and the time variable, t_n, i.e., $\phi(t_n, x_k) := \phi_{n,k}$. In the discretization for the spatial variable, we have $\phi(t, x_k) := \phi_k(t), k = 1, 2, ..., N$, which yields a dynamical system

$$h'(t) = Dh(t), \tag{1.9}$$

where $h(t) = (\phi_1(t), \phi_2(t), ..., \phi_N(t))^T$ and D is usually a tridiagonal matrix, e.g., by using finite difference method. To ensure that the aforementioned thermodynamics rule is satisfied, we need that $h(t) \in \mathcal{P}$ for all $h(0) \in \mathcal{P}$, where $\mathcal{P} = [\phi_{\min}, \phi_{\max}]^N$, i.e., $\mathcal{P}$ is an invariant set for the dynamical system (1.9). Then for the discretization of the time variable, one needs discretization methods which are invariance preserving while the aforementioned thermodynamics rule is satisfied, i.e., $\phi_{n,k} \in [\phi_{\min}, \phi_{\max}]$ for any $\phi(0,x) \in [\phi_{\min}, \phi_{\max}]$.

Invariance Preserving Numerical Methods: Now we consider the effects of Euler methods on the continuous system, i.e., given a vector x_k in $\mathcal{S}$, we investigate conditions that ensure that x_{k+1} obtained by using the forward or the backward Euler methods is also in $\mathcal{S}$. A geometric interpretation of the forward Euler method is that x_{k+1} is on the tangent line of $x(t)$. For a convex set $\mathcal{S}$, it is well known that the tangent space at any x_k on the

boundary of $\mathcal{S}$ is a supporting hyperplane to $\mathcal{S}$, see e.g., [58]. Figure 1.1 illustrates the effects of the Euler methods on two classes of trajectories. In these two cases, the convex sets include the trajectory on its boundary, and include the region above the curves. The left side subfigure of Figure 1.1 shows that the forward and backward Euler methods lead the discrete steps direct outside and inside the convex set, respectively. The right side subfigure of Figure 1.1 shows that the discrete steps for both Euler methods are on the boundary.

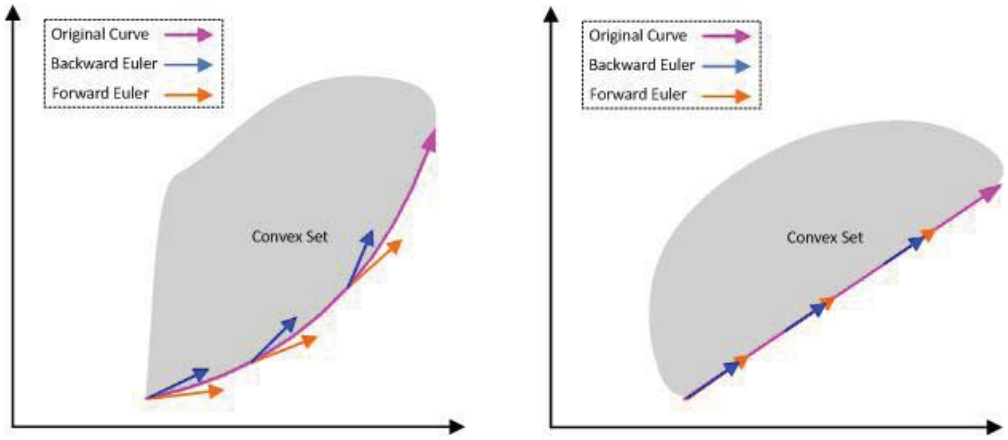

Figure 1.1: The left figure illustrates when $x(t)$ is a curve, the right figure illustrates when $x(t)$ is a line.

Characteristics Preserving: Mathematical modeling of many problems from the real world often leads to differential equations in continuous form. When we solve these differential equations numerically, we not only need to obtain a good approximation of the differential equations, but also hope to preserve the basic characteristics of these mathematical variables and models. Invariance preserving is one of the latter type requirements. In fact, there are various characteristics preserving topics, e.g., positivity preserving, strong stability preserving, area preserving, etc, which are widely studied in recent decades.

1. *Positivity Preserving:* Positivity preserving is an important topic in the numerical analysis community, see, e.g., [37, 38, 39, 82, 83, 84]. Positivity preserving is equivalent to invariance preserving in the positive orthant, i.e., consider the positive orthant, which is a polyhedral cone. Let us assume that the positive orthant is an invariant set for a continuous system, and assume that it is also an invariant set for the discrete system which is obtained

by using a discretization method with a certain steplength. In practice, many variables, e.g., energy, density, mass, etc, are nonnegative. When these variables are used in some mathematical models in a continuous form, e.g., in the heat equation, one should choose appropriate discretization method with appropriate steplength such that solution of the the discretized systems are also nonnegative.

2. *Strong Stability Preserving (SSP):* Strong stability preserving (SSP) numerical methods are developed to solve ordinary differential equations, see, e.g., [26, 27], etc. Particularly, SSP numerical method are used for the time integration of semi-discretizations of hyperbolic conservation laws. It is well known that the exact solutions of scalar conservation laws holds the property that total variation does not increase in time, see, e.g., [27]. SSP methods are also referred to as total variation diminishing methods. These are higher order numerical methods that also preserve this property.

3. *Area Preserving-Symplectic Methods:* Intuitively, a map from the phase-plane to itself is said to be symplectic if it preserves areas. In mathematics, a matrix $M \in \mathbb{R}^{2n\times 2n}$ is called symplectic if it satisfies the condition $M^T\Omega M = \Omega$, where $\Omega = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}$. A symplectic map is a real-linear map T that preserves a symplectic form f, i.e., $f(Tx, Ty) = f(x, y)$ for all x, y, see, e.g., [51]. A numerical one-step method $x_{n+1} = D_{\Delta t}(x_n)$ is called symplectic if, when applied to a Hamiltonian system, the discrete flow $x \to D_{\Delta t}(x)$ is a symplectic map for all sufficiently small step sizes, see, e.g., [25, 50], etc. There is one compelling example that shows symplectic methods are the right way to solve planetary trajectories. If we solve the trajectory of the earth using forward Euler method, then the discrete trajectory will spiral away from the sun. If we use backward Euler method, then the discrete trajectory will sink into the sun. If we use symplectic methods, then the discrete trajectory will stay on the original continuous trajectory.

1.3 Main Results

In this section, the main results and novelties of this thesis are presented. The main intellectual challenge of this research is to establish a strong connection among optimization theory,

dynamical systems, and numerical analysis. Our fundamental studies will bring a unified novel optimization theory based deep understanding of invariant sets of dynamical systems and invariance preserving discretization methods. Due to the introduction of optimization theory and algorithms, the verification of an invariant set becomes more efficient; we are deriving novel characterizations of invariance preserving discretization methods; providing optimization models to compute optimal invariance preserving discretization step sizes; and open avenues to construct optimal invariant sets for dynamical systems.

In Chapter 2, we deal with dynamical systems in finite dimensional spaces and introduce a novel and unified method for the determination of whether a set is a positively invariant set for a linear dynamical system. Here the sets are polyhedral sets, ellipsoids, and Lorenz cones. In addition, we formulate optimization methods to check the resulting equivalent conditions.

The main tool in the continuous time case is the explicit computation of the tangent cones of the positively invariant sets and their application along the lines of the Nagumo Theorem 7.2.5. This theorem says that a set is positively invariant, under some conditions on solvability of the underlying differential equation, if and only if at each point of the set, the vector field of the differential equation points toward the tangent cone at that point. The resulting conditions are constructive in the sense that they can be checked by well established optimization methods. Our unified approach is based on optimization methodology. The analysis in the discrete case is based on the theorems of alternatives of optimization, namely on the Farkas Lemma 7.2.1 and the S-lemma 7.2.2. Let us mention that the technique with the tangent cones in the continuous time case and the theorem of alternatives of optimization in the discrete case show common features.

First, in Chapter 2, we consider various sets as candidates for positively invariant sets with respect to a discrete system. Sufficient and necessary conditions for the four types of sets are derived using the Farkas Lemma 7.2.1 and the S-lemma 7.2.2, respectively. The Farkas Lemma and the S-lemma are frequently referred to as Theorems of the Alternatives in the optimization literature. Note that the approach based on the Farkas Lemma is originally due to Hennet [31]. Our approach, based on the S-lemma for ellipsoids and Lorenz cones, is not only simpler compared to the traditional Lyapunov theory based approach, but also

highlights the strong relationship between dynamical system and optimization theories. It also enables us to extend invariance conditions to any set represented by a quadratic inequality. Such sets include nonconvex and unbounded sets. Positively invariant sets for continuous systems are linked to the ones for discrete systems by applying Euler method. The forward Euler method or backward Euler method is used to discretize a continuous system to a discrete system. Then, sufficient and necessary conditions under which the four types of convex sets are positively invariant sets for the continuous systems are derived by using Euler methods and the corresponding sufficient and necessary conditions for the discrete systems.

The main novelty of Chapter 2 is that we propose a simple, novel, unified approach, different from the traditional Lyapunov stability theory approach, to derive invariance conditions for the four types of sets to be positively invariant sets with respect to discrete systems. Our approach is based on the so-called Theorems of Alternatives, i.e., Farkas Lemma and S-lemma. For discrete systems, the Farkas lemma is used for polyhedral sets, while the S-lemma is used for ellipsoids and Lorenz cones. We also establish a framework according to Euler methods to derive invariance conditions for the four types of sets with respect to the continuous systems to be positively invariant. Although some theorems presented in Chapter 2 are known, there is no existing chapter considering invariance conditions for the four types of sets, and both for discrete and continuous dynamical systems together in a unified framework. We also strengthen the power of Euler methods as a tool to study invariance conditions to build connection between continuous and discrete dynamical systems.

In Chapter 3, 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. First, 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 [14, 15]. 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 Δt, 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 [38, 39], where Runge-Kutta methods are considered.

In Chapter 4, our focus is to find conditions, in particular steplength thresholds for the discretization methods, such that the considered discretization method is invariance preserving for the given linear or nonlinear dynamical system. This topic is of great interest in the fields of dynamical systems, partial differential equations, and control theory. A basic result is presented in [16], which considers linear problems and invariance preserving on the positive orthant from a perspective of numerical methods. For invariance preserving on the positive orthant or polyhedron for Runge-Kutta methods, the reader is refereed to [37]. A similar concept named strong stability preserving (SSP) used in numerical methods is studied in [26, 63]. These papers deal with invariance preserving of general sets and they usually use the assumption that the Euler methods are invariance preserving with a steplength threshold τ_0. Then the uniform invariance preserving steplength threshold for other advanced numerical methods, e.g., Runge-Kutta methods, is derived in terms of τ_0. Therefore, to make the results applicable to solve real world problems, this approach requires to check whether such a positive τ_0 exists for Euler methods.

In Section 4.2 first we prove that for the forward Euler method, a local invariance preserving steplength threshold exists for a given polyhedron when a linear dynamical system is considered. For the backward Euler method we prove that a local steplength threshold exists for polyhedron, ellipsoid, and Lorenz cone. These proofs are using elementary concepts. We also quantify a valid local steplength threshold for the backward Euler method. Second, we prove that a uniform invariance preserving steplength threshold exists for poly-

hedra when the forward or backward Euler method is applied to linear dynamical systems. For the backward Euler method, we also quantify the optimal uniform steplength threshold. In Section 4.3 we first prove that a uniform steplength threshold exists, and also quantify the optimal uniform steplength threshold for ellipsoids or Lorenz cones when the backward Euler method is applied to linear dynamical systems. Moreover, we extend the results about the invariance preserving steplength threshold for the backward Euler method to general proper cones. Finally, we present our main results about uniform steplength thresholds. These results are natural extensions from the proofs used to analyze Euler methods. We quantify the optimal uniform steplength threshold of the backward Euler method for convex sets. We also extend the results of steplength thresholds to general compact sets and proper cones when a general discretization method is applied to linear or nonlinear dynamical systems. In particular, the existence of steplength thresholds depends on a condition that is stronger than the existence of local steplength threshold[2]. It also depends on a Lipschitz condition when the set is a compact set, and on a homogeneity condition, when the set is a proper cone.

The main novelty of Chapter 4 is establishing the foundation of characterizing invariance preserving discretization methods in dynamical systems and differential equations. As mentioned before, several existing results on invariance preserving of advanced numerical methods, e.g., Runge-Kutta methods, require the existence of a positive steplength threshold for Euler methods. In Chapter 4, we present the results for special classical sets. Our general results about steplength threshold for general discretization methods for linear and nonlinear dynamical systems on convex sets, compact sets, and proper cones not only play an important role in theoretical, but also show the potential of significant impacts in practice. These general results provide theoretical criteria for the verification of the existence of invariance preserving steplength threshold for discretization methods. Once the existence is ensured by our results, this also motivates one to further investigate the possibility to find the optimal steplength threshold, which has several advantages in practice. Such advantages include computational efficiency and smaller size of discrete systems.

In Chapter 5, we derive invariance condition of some classical convex sets for discrete

[2]In particular, this condition requires that if x_k is in the set, then x_{k+1} is in the interior of the set.

and continuous nonlinear systems. This chapter is an extension of Chapter 2 from linear systems to nonlinear systems. The main tools we used are the nonlinear Farkas Lemma and the S-lemma. We also propose optimization methods to verify if these invariance conditions hold.

1.4 The Structure of the Thesis

In Chapter 2 we propose a novel, simple, and unified approach to explore sufficient and necessary conditions under which four classic families of convex sets, i.e., polyhedra, polyhedral cones, ellipsoids, and Lorenz cones, are invariant sets for a linear discrete or continuous dynamical system.

In Chapter 3 we propose novel methods to compute valid or largest uniform steplength thresholds for invariance preserving for three classic types of discretization methods, i.e., Taylor approximation type, rational function type discretization methods, and Euler methods.

In Chapter 4 we propose a theory of studying the existence of local and uniform steplength thresholds for invariance preserving on a set when a discretization method is applied to a linear or nonlinear dynamical system. We not only consider classic sets, discretization methods, and dynamical systems, but also extend to more general sets, i.e., convex sets, compact sets, proper cones, discretization methods, and dynamical systems.

In Chapter 5, we derive invariance conditions of some classic sets for nonlinear dynamical systems. One can consider this chapter as an extension of the first part to a general case.

Finally, Chapter 6 summarizes the results of this thesis and presents the future research.

Chapter 2 is based on the paper [43]. Chapter 3 is based on paper [41]. Chapter 4 is based on paper [40]. Chapter 5 is based on paper [42].

1.5 Notations and Conventions

To avoid unnecessary repetitions, the following notations and conventions are used in this thesis. A dynamical system, positively invariant, and sufficient and necessary condition for positive invariance are called a *system*, *invariant*, and *invariance condition*, respectively.

The sets considered in this chapter are non-empty, closed, and convex sets if not specified otherwise. The interior and the boundary of a set $\mathcal{S}$ is denoted by $\text{int}(\mathcal{S})$ and $\partial\mathcal{S}$, respectively. When a matrix Q is positive definite, positive semidefinite, negative definite, or negative semidefinite matrix, then it is denoted by $Q \succ 0$, $Q \succeq 0, Q \prec 0$, or $Q \preceq 0$, respectively. The i-th row of a matrix G is denoted by G_i^T. The eigenvalues of a real symmetric matrix Q, whose eigenvalues are always real, are ordered as $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$, and the corresponding eigenvectors are denoted by $u_1, u_2, ..., u_n$. The spectrum of Q is represented by $\lambda(Q) = \max\{|\lambda_i(Q)|\}$, and $\text{inertia}\{Q\} = \{\alpha, \beta, \gamma\}$ indicates that the number of positive, zero, and negative eigenvalues of Q are α, β, and γ, respectively. The index set $\{1, 2, ..., n\}$ is denoted by $\mathcal{I}(n)$. The inner product of vectors $x, y \in \mathbb{R}^n$ is denoted by $x^T y$.

Chapter 2

Invariance Conditions for Classical Convex Sets

2.1 Introduction

In this chapter, we propose a novel, simple, and unified approach to explore sufficient and necessary conditions, i.e., invariance conditions, under which four classic families of convex sets, namely, polyhedra, polyhedral cones, ellipsoids, and Lorenz cones, are invariant sets for a linear discrete or continuous dynamical system. For discrete dynamical systems, we use the Theorems of Alternatives, i.e., the Farkas Lemma and S-lemma, to obtain simple and general proofs to derive invariance conditions. This novel method establishes a solid connection between optimization theory and dynamical system. Also, using the S-lemma allows us to extend invariance conditions to any set represented by a quadratic inequality. Such sets include nonconvex and unbounded sets. For continuous dynamical systems, we use the forward or backward Euler method to obtain the corresponding discrete dynamical systems while discretization preserves invariance. This enables us to develop a novel and elementary method to derive invariance conditions for continuous dynamical systems by using the ones for the corresponding discrete systems.

2.2 Invariance Conditions

In this section, we present invariance conditions, i.e., sufficient and necessary conditions under which polyhedral sets, ellipsoids, and Lorenz cones are invariant sets for discrete and continuous systems. For each convex set, the invariance conditions for discrete systems are first derived by using the Theorems of Alternatives, i.e., the Farkas Lemma 7.2.1 or the S-lemma 7.2.2. Then the invariance conditions for continuous systems are derived by using a discretization method to discretize the continuous system and applying the invariance conditions for the obtained discrete systems.

We use Euler methods to discretize the continuous system (1.2) to derive a discrete system, because for sufficiently small step size they preserve the invariance of a set, i.e., a set, which is an invariant set for a continuous system, is also an invariant set for the corresponding derived discrete system. Here we formally present these results as follows. The first statement can be found in [10, 13], and the second statement can be found in [71].

Theorem 2.2.1. *[10, 13, 71] Assume a polyhedron $\mathcal{P}$, polyhedral cone $\mathcal{C}_\mathcal{P}$, ellipsoid $\mathcal{E}$ or Lorenz cone $\mathcal{C}_\mathcal{L}$ is an invariant set for the continuous system (1.2). Then*

- *there exists a $\hat{\tau} > 0$, such that $\mathcal{P}$ (or $\mathcal{C}_\mathcal{P}$) is also an invariant set for the discrete system $x_{k+1} = (I + A_c\Delta t)x_k$ for all $0 \leq \Delta t \leq \hat{\tau}$, and*
- *there exists a $\tilde{\tau} > 0$, such that $\mathcal{P}$ ($\mathcal{C}_\mathcal{P}, \mathcal{E}$ or $\mathcal{C}_\mathcal{L}$) is also an invariant set for the discrete system $x_{k+1} = (I - A_c\Delta t)^{-1}x_k$ for all $0 \leq \Delta t \leq \tilde{\tau}$.*

Remark 2.2.2. *The first statement in Theorem 2.2.1 means that the forward Euler method preserves the invariance of a polyhedral set, while the second statement means that the backward Euler method preserves the invariance of polyhedral set, ellipsoid, and Lorenz cone.*

Proposition 1.1.3 allows us to use the Theorems of Alternatives 7.2.1 and 7.2.2 to derive invariance conditions for discrete systems. According to Proposition 1.1.3, to prove that a set $\mathcal{S}$ is an invariant set for a discrete system, we need to prove $A\mathcal{S} \subseteq \mathcal{S}$, which is equivalent to $(\mathbb{R}^n \setminus \mathcal{S}) \cap (A\mathcal{S}) = \emptyset$. Since we assume that $\mathcal{S}$ is a closed set, we have that $\mathbb{R}^n \setminus \mathcal{S}$ is an open set. Open sets are usually represented by strict inequalities. As the Theorems

of Alternatives include strict inequalities, they provide the proper tools to characterize invariance conditions for continuous and discrete systems. This is one of the statements in the Theorems of Alternatives 7.2.1 or 7.2.2.

2.2.1 Polyhedral Sets

Since every polyhedral set has two different representations as shown in Section 7.1, we present the invariance conditions for both forms, respectively.

2.2.1.1 Invariance Conditions for Discrete Systems

Let a polyhedral set $\mathcal{P}$ be given as in (7.1). The invariance condition of a polyhedral set $\mathcal{P}$ for a discrete system is presented in Theorem 2.2.4. The study of invariance condition of polyhedral sets for discrete system can be traced back to Bitsoris in [8, 9], who consider a special class of polyhedral sets, normally those polyhedral sets which are symmetric with respect to the origin. We give a more straightforward proof here by using the Farkas Lemma for polyhedral sets in the form of (7.1). It was brought to our attention recently that the result is the same as the one presented by Hennet [31], which also uses the Farkas Lemma. To keep the chapter self-contained, we also present the proof of this important result.

Definition 2.2.3. *A matrix H is called a **nonnegative matrix**, denoted by $H \geq 0$, if $H_{ij} \geq 0$ for all i, j. A matrix L is called an **essentially nonnegative matrix**[1], denoted by $L \geq_o 0$, if $L_{ij} \geq 0$ for $i \neq j$.*

Theorem 2.2.4. (Hennet [31]) *A polyhedron $\mathcal{P}$ given as in (7.1) is an invariant set for the discrete system (1.1) if and only if [2] there exists a matrix $H \in \mathbb{R}^{m \times m}$, such that $H \geq 0, HG = GA_d$ and $Hb \leq b$.*

Proof. We have that $\mathcal{P}$ is an invariant set for the discrete system (1.1) if and only if $A_d\mathcal{P} \subseteq \mathcal{P}$, which is the same as $\mathcal{P} \subseteq \mathcal{P}' = \{x \mid GA_dx \leq b\}$. Note that $\mathcal{P} \subseteq \mathcal{P}'$ if and only if for

[1] An essentially nonnegative matrix is also called Metzler matrix, see e.g., [19], or quasipositive matrix, see, e.g., [5].

[2] One can also show the "if" part as follows: let $x \in \mathcal{P}$, i.e., $Gx \leq b$. Since $H \geq 0, HG = GA_d$ and $Hb \leq b$, we have $GA_dx = HGx \leq Hb \leq b$, i.e., $A_dx \in \mathcal{P}$.

every $i \in \mathcal{I}(m)$, we have

$$\{x \,|\, Gx \leq b\} \cap \{x \,|\, (GA_d)_i^T x > b_i\} = \emptyset,$$

i.e., the inequality system $Gx \leq b$ and $(GA_d)_i^T x > b_i$ has no solution. According to the Farkas Lemma 7.2.1, this is equivalent to that there exists a vector $h_i \geq 0$, such that $h_i^T G = (GA_d)_i$, and $h_i^T b \leq b_i$. We let $H = [h_1^T; h_2^T; ...; h_m^T]$, then we have $H \geq 0, HG = GA_d$ and $Hb \leq b$. The proof is complete. □

We highlight that Castelan and Hennet [19] present an algebraic characterization of the matrix G satisfying the conditions in Theorem 2.2.4. They prove that given A and G, there exists a matrix H satisfying $HG = GA$ if and only if the kernel of G is an A-invariant subspace. Observe that the existence of matrix H such that $HG = GA, Hb \leq b$, and $x \geq 0$ can be verified by solving the linear inequality system or interior point methods (IPMs). If IPMs are used, then the verification can be done in polynomial time.

The invariance condition of a polyhedral set given as in (7.2) for discrete systems is provided in Theorem 2.2.5. Note that a similar result is presented in [13], which considers only the case when the set is a polytope. Invariance condition of a polytope is presented in [13], while invariance condition of a polyhedral cone is presented in [74]. Here we integrate these two results in one theorem.

Theorem 2.2.5. *A polyhedron $\mathcal{P}$ given as in (7.2) is an invariant set for the discrete system (1.1) if and only if there exists a matrix $L \in \mathbb{R}^{(\ell_1+\ell_2)\times(\ell_1+\ell_2)}$, such that $L \geq 0, XL = A_d X$ and $\tilde{1}L = \bar{1}$, where $X = [x^1, ..., x^{\ell_1}, \hat{x}^1, ..., \hat{x}^{\ell_2}]$, $\tilde{1} = (1_{\ell_1}, 0_{\ell_2}), \bar{1} = 1_{\ell_1+\ell_2}$.*

Proof. Note that $\mathcal{P}$ given as in (7.2) is an invariant set for the discrete system if and only if $A_d x^i \in \mathcal{P}$, for all $i \in \mathcal{I}(\ell_1)$, and $A(O^+\mathcal{P}) \subseteq O^+\mathcal{P}$, where $O^+\mathcal{P}$ denotes the recession cone of $\mathcal{P}$. Since $O^+\mathcal{P}$ is generated by the vectors $\hat{x}^j$, where $j \in \mathcal{I}(\ell_2)$, we have $A_d(O^+\mathcal{P}) \subseteq O^+\mathcal{P}$, which can be rewritten as $A_d \hat{x}^j \in \mathcal{P}$, for all $j \in \mathcal{I}(\ell_2)$. Then for all $p_1 \in \mathcal{I}(\ell_1)$, there exist $\theta_{p_1}^i, \theta_{p_1}^j \geq 0$, such that $\sum_{p_1=1}^{\ell_1} \theta_{p_1}^i = 1, \sum_{p_1=1}^{\ell_1} \theta_{p_1}^j = 1$, and for all $p_2 \in \mathcal{I}(\ell_2)$, there exist

$\hat{\theta}^i_{p_2}, \hat{\theta}^j_{p_2} \geq 0$, such that

$$A_d x^i = \sum_{p_1=1}^{\ell_1} \theta^i_{p_1} x^{p_1} + \sum_{p_2=1}^{\ell_2} \hat{\theta}^i_{p_2} x^{p_2}, \quad A_d \hat{x}^j = \sum_{p_2=1}^{\ell_2} \hat{\theta}^j_{p_2} x^{p_2}. \tag{2.1}$$

Let $L = [\theta^1, .., .\theta^{\ell_1}, \hat{\theta}^1, ..., \hat{\theta}^{\ell_2}]$, then the theorem is immediate by (2.1). □

Observe again that the condition in Theorem 2.2.5 is also a linear inequality system, thus this condition can be verified in polynomial time by using IPMs.

A polyhedral cone is a special polyhedral set, thus we have the following invariance condition of a polyhedral cone for discrete systems.

Corollary 2.2.6. *1). A polyhedral cone $\mathcal{C}_\mathcal{P}$ given as in (7.3) is an invariant set for the discrete system (1.1) if and only if there exists a matrix $H \in \mathbb{R}^{m\times m}$, such that $H \geq 0$ and $HG = GA_d$.*

2). A polyhedral cone $\mathcal{C}_\mathcal{P}$ given as in (7.4) is an invariant set for the discrete system (1.1) if and only if there exists a matrix $L \in \mathbb{R}^{\ell\times\ell}$, such that $L \geq 0$ and $XL = A_d X$, where $X = [\hat{x}^1, ..., \hat{x}^\ell]$.

Verifying if a polyhedral set is an invariant set: For a given polyhedral set and a discrete system, according to Theorem 2.2.4 (Theorem 2.2.5, or Corollary 2.2.6), to determine whether the set is an invariant set for the system is equivalent to verify the existence of a nonnegative matrix H (or L), which is actually a linear optimization problem. Rather than computing H (or L) directly, it is more efficient to independently solve all the small subproblems. Let us choose polyhedron $\mathcal{P}$ as given in (7.1) and Theorem 2.2.4 as an example to illustrate this idea. We can independently examine the feasibility of the subproblems. Find $h_i \in \mathbb{R}^n$, such that $h_i^T G = G_i^T A_d$, $h_i \geq 0$, and $h_i^T b \leq b_i$, for all $i \in \mathcal{I}(m)$. Clearly, these are linear feasibility problems which can be considered as special cases of linear optimization problems, see, e.g., [34]. A linear optimization problem can be solved in polynomial time, e.g., by using interior point methods [60]. In fact, we note that for each i, we can write $h_i^T G = G_i^T A_d, h_i \geq 0$, and $h_i^T b \leq b_i$ in the form of $\bar{A}x = b, x \geq 0$, where $A \in \mathbb{R}^{2\times(m+1)}$ and $b \in \mathbb{R}^2$. So each of this problem has exactly 2 equality constraints. For such problem, the

number of constraint is at most $m(m+1)/2$, then all finite pivot algorithms are also polynomial time. Actually, this is strongly polynomial time. If all of these linear optimization problems are feasible, then their solutions form a nonnegative matrix H that satisfies the condition in Theorem 2.2.4. Otherwise, we can conclude that the set is not an invariant set for the system, and computation is terminated at the first infeasible subproblem.

2.2.1.2 Invariance Conditions for Continuous Systems

According to the results presented in Chapter 3, we have that both the forward and backward Euler methods are invariance preserving for a polyhedral set. Blanchini [10, 13] presents the connection between invariant sets for continuous and discrete systems by using the forward Euler method. The discrete system obtained by using the forward Euler method is refered to as Euler Approximating System [10, 13]. We first present the following invariance condition which is obtained by using Nagumo's Theorem 7.2.5. For $x \in \mathcal{P}$, let $\mathcal{I}_x$ denote the set of indices of the constraints which are active at x, i.e., the corresponding linear inequality holds as equality at x. Clearly, we have $x \in \partial\mathcal{P}$ if and only if $\mathcal{I}_x \neq \emptyset$.

Lemma 2.2.7. *Let a polyhedron $\mathcal{P}$ be given as in (7.1), and $\mathcal{I}_x \neq \emptyset$. Then $\mathcal{P}$ is an invariant set for the continuous system (1.2) if and only if for every $x \in \partial\mathcal{P}$, i.e., $G_i^T x = b_i$, for $i \in \mathcal{I}_x$, we have*

$$G_i^T A_c x \leq 0,\ i \in \mathcal{I}_x. \tag{2.2}$$

Proof. The tangent cone at x, where $G_i^T x = b_i$ for $i \in \mathcal{I}_x$, is $\mathcal{T}_{\mathcal{P}}(x) = \{y \,|\, G_i^T y \leq 0, i \in \mathcal{I}_x\}$ (see [35, p.138]). Then the lemma immediately follows from Nagumo's Theorem 7.2.5. □

We now present another invariance condition of a polyhedron in the form of (7.1) for the continuous system (1.2). The following theorem can also be found in Castelan and Hennet [19, Proposition 1].

Theorem 2.2.8. *A polyhedron $\mathcal{P}$ given as in (7.1) is an invariant set for the continuous system (1.2) if and only if there exists a matrix $\tilde{H} \in \mathbb{R}^{m\times m}$, such that $\tilde{H} \geq_o 0, \tilde{H}G = GA_c$ and $\tilde{H}b \leq 0$.*

Proof. We first consider the "if" part. Note that if $\tilde{H}G = GA_c$, then we have $\tilde{H}_i^T Gx = G_i^T A_c x$, for every $i \in \mathcal{I}(n)$. Since $\tilde{H} \geq_o 0$ and $x \in \partial\mathcal{P}$,

$$\begin{aligned} &\text{when } j = i, \quad \text{we have } \tilde{h}_{ii} \in \mathbb{R} \text{ and } G_i^T x = b_i, \\ &\text{when } j \neq i, \quad \text{we have } \tilde{h}_{ij} \geq 0 \text{ and } G_j^T x \leq b_j, \end{aligned} \tag{2.3}$$

where $\tilde{h}_{ij}$ is the (i,j)-th entry of $\tilde{H}$. According to (2.3), we have $\sum_{j=1}^m \tilde{h}_{ij}(G_j^T x - b_j) \leq 0$, i.e., $\tilde{H}_i^T Gx \leq \tilde{H}_i^T b$. Since $\tilde{H}b \leq 0$, we have $\tilde{H}_i^T b \leq 0$. Then, we have $G_i^T A_c x = \tilde{H}_i^T Gx \leq \tilde{H}_i^T b \leq 0$. According to Lemma 2.2.7, we have that $\mathcal{P}$ is an invariant set for the continuous system.

Now we consider the "only if" part. According to Theorem 2.2.1, we have that there exists a $\hat{\tau} > 0$, such that $\mathcal{P}$ is also an invariant set for the discrete system $x_{k+1} = (I + A_c\Delta t)x_k$, for every $0 \leq \Delta t \leq \hat{\tau}$. Then, according to Theorem 2.2.4, there exists a matrix $H(\Delta t) \geq 0$, such that $H(\Delta t)G = G(I + A_c\Delta t),$ and $H(\Delta t)b \leq b$, i.e.,

$$\frac{H(\Delta t) - I}{\Delta t} G = GA_c, \text{ and } \frac{H(\Delta t) - I}{\Delta t} b \leq 0. \tag{2.4}$$

Clearly $\tilde{H} = \frac{H(\Delta t)-I}{\Delta t}$ for $\Delta t > 0$ satisfies this theorem. □

We consider the invariance condition of the polyhedron in the form of (7.2) for the continuous system (1.2). For an arbitrary convex set in $\mathbb{R}^n$, we have the following conclusion

Lemma 2.2.9. *Let $\mathcal{S}$ be a convex set in $\mathbb{R}^n$. For any $\ell \in \mathbb{N}$ and $x, y^1, y^2, ..., y^\ell \in \mathcal{S}$ satisfying $x = \sum_{i=1}^\ell \beta_i y^i$, where $\sum_{i=1}^\ell \beta_i = 1$ and $\beta_i > 0$ for every $i \in \mathcal{I}(\ell)$, we have $\mathcal{T}_\mathcal{S}(y^i) \subseteq \mathcal{T}_\mathcal{S}(x)$ for every $i \in \mathcal{I}(\ell)$.*

Proof. We denote $\text{cone}(x, \mathcal{S}) = \{\alpha(y - x) \,|\, y \in \mathcal{S}, \alpha \geq 0\}$, then we have that $\mathcal{T}_\mathcal{S}(x)$ is the same as the topological closure of $\text{cone}(x, S)$. Let $\Phi(x)$ denote the face of $\mathcal{S}$ generated by x, i.e., the set $\{y \in \mathcal{S} \,|\, \mu x + (1-\mu)y \in \mathcal{S} \text{ for some } \mu > 1\}$. We first show that for any $x, u \in \mathcal{S}$, if $u \in \Phi(x)$, then $\mathcal{T}_\mathcal{S}(u) \subseteq \mathcal{T}_\mathcal{S}(x)$. In fact, by definition of $\Phi(x)$ there exists $\mu > 1$, such that $v := \mu x + (1-\mu)u \in \mathcal{S}$. Then we have $x = (1-\alpha)u + \alpha v$ for some $\alpha, 0 < \alpha < 1$. Note that for any $y \in \mathcal{S}$, we have $(1-\alpha)y + \alpha v \in \mathcal{S}$ and $[(1-\alpha)y + \alpha v] - x = (1-\alpha)(y-u)$. It follows that $\text{cone}(u, \mathcal{S}) \subseteq \text{cone}(x, \mathcal{S})$. By taking the closure of both sides, we have $\mathcal{T}_\mathcal{S}(u) \subseteq \mathcal{T}_\mathcal{S}(x)$.

Since $\sum_{i=1}^{\ell} \beta_i = 1$ and $\beta_i > 0$ for every $i \in \mathcal{I}(\ell)$, $y^i \in \Phi(x)$, for every $i \in \mathcal{I}(\ell)$ we have $y^i \in \Phi(x)$, the lemma follows immediately. □

For the polyhedron $\mathcal{P}$ given as in (7.2), a vertex of $\mathcal{P}$ is given as x^i, for some $i \in \mathcal{I}(\ell_1)$, and an extreme ray of $\mathcal{P}$ is represented as $x^i + \alpha\hat{x}^j, \alpha > 0$, for some $i \in \mathcal{I}(\ell_1)$ and $j \in \mathcal{I}(\ell_2)$. Applying Lemma 2.2.9 to $\mathcal{P}$, we have the following Corollary 2.2.10 about the relationship between tangent cones at a vector and the vertices and extreme rays of $\mathcal{P}$. Note that $\mathcal{T}_{\mathcal{P}}(x) = \mathbb{R}^n$ for every $x \in \text{int}(\mathcal{S})$, thus Corollary 2.2.10 is only nontrivial for $x \in \partial\mathcal{P}$.

Corollary 2.2.10. *Let a polyhedron $\mathcal{P}$ be given as in (7.2), and $x \in \mathcal{P}$ be a point in $\mathcal{P}$ given as in formula (7.2). Let $\mathcal{I}_1 = \{i \in \mathcal{I}(\ell_1) \,|\, \theta_i > 0\}$ and $\mathcal{I}_2 = \{j \in \mathcal{I}(\ell_2) \,|\, \hat{\theta}_j > 0\}$. Then $\mathcal{T}_{\mathcal{P}}(x^i) \subseteq \mathcal{T}_{\mathcal{P}}(x)$ and $\mathcal{T}_{\mathcal{P}}(x^i + \alpha\hat{x}^j) = \mathcal{T}_{\mathcal{P}}(x^i + \hat{x}^j) \subseteq \mathcal{T}_{\mathcal{P}}(x)$ for $i \in \mathcal{I}_1, j \in \mathcal{I}_2$, and $\alpha > 0$, where $x^i + \alpha\hat{x}^j$ is an extreme ray of $\mathcal{P}$.*

Let us consider a polytope $\tilde{\mathcal{P}}$ generated by $\{x^1, x^2, ..., x^{\ell_1}\}$ as its vertices. Then, according to [10], we have that $\mathcal{T}_{\tilde{\mathcal{P}}}(x^i)$ can be generated as a conic combination of $x^p - x^i$ for all $p \in \mathcal{I}(\ell_1)$, i.e., $\mathcal{T}_{\tilde{\mathcal{P}}}(x^i) = \{y | y = \sum_{p=1, p\neq i}^{\ell_1} \alpha_p(x^p - x^i), \alpha_p \geq 0\}$. Let $\alpha_i = \sum_{p=1, p\neq i}^{\ell_1} \alpha_p$. Then we have

$$\mathcal{T}_{\tilde{\mathcal{P}}}(x^i) = \Big\{y | y = \sum_{p=1}^{\ell_1} \alpha_p x^p, \alpha_p \geq 0, p \neq i, \sum_{p=1}^{\ell_1} \alpha_p = 0\Big\}.$$

By a similar argument, we have that the exact representations of the tangent cones at vertices or extreme rays of $\mathcal{P}$ given as in (7.2) are presented in Lemma 2.2.11.

Lemma 2.2.11. *Let a polyhedron $\mathcal{P}$ be given as in (7.2), and $\mathcal{I}_1' = \{i \in \mathcal{I}(\ell_1) \,|\,$ for any $j \in \mathcal{I}(\ell_2), x^i + \hat{x}^j$ is not an extreme ray.$\}$, $\mathcal{I}_1'' = \mathcal{I}(\ell_1) \backslash \mathcal{I}_1'$, then*

1). For every $i \in \mathcal{I}_1'$, we have

$$\mathcal{T}_{\mathcal{P}}(x^i) = \{y \in \mathbb{R}^n \,|\, y = \sum_{p=1}^{\ell_1} \alpha_p x^p, \alpha_p \geq 0, p \neq i, \sum_{p=1}^{\ell_1} \alpha_p = 0\}.$$

2). For every $i \in \mathcal{I}_1''$, we have

$$\mathcal{T}_{\mathcal{P}}(x^i) = \{y \in \mathbb{R}^n \,|\, y = \sum_{p=1}^{\ell_1} \alpha_p x^p + \sum_{q=1}^{\ell_2} \hat{\alpha}_q \hat{x}^q, \alpha_p, \hat{\alpha}_q \geq 0, p \neq i, \sum_{p=1}^{\ell_1} \alpha_p = 0\}.$$

3). For every $i \in \mathcal{I}_1''$ and $j \in \mathcal{I}(\ell_2)$ such that $x^i + \hat{x}^j$ is an extreme ray, we have

$$\mathcal{T}_{\mathcal{P}}(x^i + \hat{x}^j) = \{y \in \mathbb{R}^n \,|\, y = \sum_{q=1}^{\ell_2} \hat{\alpha}_q \hat{x}^q, \hat{\alpha}_q \geq 0, j \neq q\}.$$

Lemma 2.2.12. *Let $\mathcal{C}$ be a closed convex cone. If $x + \alpha y \in \mathcal{C}$ for all $\alpha > 0$, then $x, y \in \mathcal{C}$.*

The following lemma presents an invariance condition for a polyhedron in the form of (7.2) for the continuous system (1.2).

Lemma 2.2.13. *Let a polyhedron $\mathcal{P}$ be given as in (7.2). Then $\mathcal{P}$ is an invariant set for the continuous system (1.2) if and only if $A_c x^i \in \mathcal{T}_{\mathcal{P}}(x^i)$ and $A_c \hat{x}^j \in \mathcal{T}_{\mathcal{P}}(x^i + \hat{x}^j)$ for $i \in \mathcal{I}(\ell_1)$ and $j \in \mathcal{I}(\ell_2)$, respectively, where $x^i + \alpha \hat{x}^j$ for $\alpha \geq 0$ is an extreme ray of $\mathcal{P}$.*

Proof. We first consider the "only if" part. According to Nagumo's Theorem 7.2.5, for any $i \in \mathcal{I}(\ell_1)$ and $j \in \mathcal{I}(\ell_2)$ when $x^i + \alpha \hat{x}^j$ for $\alpha \geq 0$ is an extreme ray, we have $A_c x^i \in \mathcal{T}_{\mathcal{P}}(x^i)$ and $A_c(x^i + \alpha \hat{x}^j) \in \mathcal{T}_{\mathcal{P}}(x^i + \hat{x}^j)$. By Lemma 2.2.12, this implies that $A_c \hat{x}^j \in \mathcal{T}_{\mathcal{P}}(x^i + \hat{x}^j)$.

For the "if" part, we choose $x \in \mathcal{P}$. We represent x as $x = \sum_{i \in \mathcal{I}_1} \theta_i x^i + \sum_{j \in \mathcal{I}_2} \hat{\theta}_j \hat{x}^j$, where $\mathcal{I}_1 = \{i \in \mathcal{I}(\ell_1) \,|\, \theta_i > 0\}$ and $\mathcal{I}_2 = \{j \in \mathcal{I}(\ell_2) \,|\, \hat{\theta}_j > 0\}$. Then according to Corollary 2.2.10, we have $A_c x = \sum_{i \in \mathcal{I}_1} \theta_i A_c x^i + \sum_{j \in \mathcal{I}_2} \hat{\theta}_j A_c \hat{x}^j \in (\cup_{i \in \mathcal{I}_1} \mathcal{T}_{\mathcal{P}}(x^i)) \cup (\cup_{j \in \mathcal{I}_2} \mathcal{T}_{\mathcal{P}}(x^i + \hat{x}^j)) \subseteq \mathcal{T}_{\mathcal{P}}(x)$. Finally, the "if" part follows by Nagumo's Theorem 7.2.5. □

By Lemma 2.2.11 and Lemma 2.2.13, the following corollary is immediate.

Corollary 2.2.14. *Let a polyhedron $\mathcal{P}$ be given as in (7.2). Then $\mathcal{P}$ is an invariant set for the continuous system (1.2) if and only if for x^i, $i \in \mathcal{I}(\ell_1)$, there exist $\alpha_p^i, \hat{\alpha}_q^i \geq 0$ for $p \neq i$, $\alpha_i^i \leq 0$, and $\hat{\alpha}_i^i \in \mathbb{R}$, such that*

$$A_c x^i = \sum_{p=1}^{\ell_1} \alpha_p^i x^p + \sum_{q=1}^{\ell_2} \hat{\alpha}_q^i \hat{x}^q, \text{ and } \sum_{p=1}^{\ell_1} \alpha_p^i = 0, \tag{2.5}$$

for $\hat{x}^j$, $j \in \mathcal{I}(\ell_2)$, there exist $\hat{\alpha}_q^j \geq 0$ for $q \neq j$, and $\hat{\alpha}_j^j \in \mathbb{R}$, such that $A_c \hat{x}^j = \sum_{q=1}^{\ell_2} \hat{\alpha}_q^j \hat{x}^q$.

Theorem 2.2.15. *A polyhedron $\mathcal{P}$ given as in (7.2) is an invariant set for the continuous system (1.2) if and only if there exists a matrix $\tilde{L} \in \mathbb{R}^{(\ell_1+\ell_2)\times(\ell_1+\ell_2)}$, such that $\tilde{L} \geq_o 0$, $X\tilde{L} = A_c X$, and $\bar{1}\tilde{L} = \bar{0}$, where $X = [x^1, ..., x^{\ell_1}, \hat{x}^1, ..., \hat{x}^{\ell_2}]$, $\bar{1} = [1_{\ell_1}, 0_{\ell_2}]$.*

Proof. This proof is similar to the one given in Theorem 2.2.8. We denote the i-th column of $\tilde{L}$ by $(l_{1,i}, ..., l_{\ell_1+\ell_2,i})^T$.

For the "if" part, we consider x^i with $i \in \mathcal{I}(\ell_1)$. Since $\tilde{L} \geq_o 0$, $X\tilde{L} = A_c X$, and $\bar{1}\tilde{L} = \bar{0}$, we have $A_c x^i = \sum_{p=1}^{\ell_1} l_{p,i} x^i + \sum_{q=1}^{\ell_2} l_{\ell_1+q,i} \hat{x}^q$, with $\sum_{p=1}^{\ell_1} l_{p,i} = 0$, and $l_{p,i} \geq 0$, for $p \neq i$. The argument for $\hat{x}^j$ with $j \in \mathcal{I}(\ell_2)$ is similar. Then, according to Corollary 2.2.14, we have that $\mathcal{P}$ is an invariant set for the continuous system.

For the "only if" part, the proof is similar to the one in Theorem 2.2.8. According to Theorem 2.2.1 and Theorem 2.2.5, we know that there exists a nonnegative matrix $L(\Delta t)$ and a scalar $\hat{\tau} > 0$, such that $XL(\Delta t) = (I + \Delta t A_c)X$, $\bar{1}L(\Delta t) = \bar{1}$, for $0 \leq \Delta t \leq \hat{\tau}$, i.e.,

$$X\frac{L(\Delta t) - I}{\Delta t} = AX, \quad \bar{1}\frac{L(\Delta t) - I}{\Delta t} = \bar{0}.$$

Let $\tilde{L} = \frac{L(\Delta t) - I}{\Delta t}$, the theorem is immediate. □

Since the invariance conditions for a polyhedral cone given in the two different forms can be obtained by similar discussions as above, we only present these invariance conditions without providing the proofs.

Corollary 2.2.16. *The following two statements hold:*

1. *A polyhedral cone $\mathcal{C}_\mathcal{P}$ given as in (7.3) is an invariant set for the continuous system (1.2) if and only if there exists a matrix $\tilde{H} \in \mathbb{R}^{m \times m}$, such that $\tilde{H} \geq_o 0$ and $\tilde{H}G = GA_c$.*

2. *A polyhedral cone $\mathcal{C}_\mathcal{P}$ given as in (7.4) is an invariant set for the continuous system (1.2) if and only if there exists a matrix $\tilde{L} \in \mathbb{R}^{\ell \times \ell}$, such that $\tilde{L} \geq_o 0$ and $X\tilde{L} = A_c X$, where $X = [\hat{x}^1, ..., \hat{x}^\ell]$.*

Verifying if a polyhedral set is an invariant set for a continuous system: Analogous to the discussion in Section 2.2.1.1, according to Theorem 2.2.15 and Corollary 2.2.16, verifying if a polyhedron given as in (7.2) or polyhedral cone given as in (7.4) is an invariant set for the continuous system (1.2) can be done by solving a series of linear optimization problems.

2.2.2 Ellipsoids

In this section, we consider invariance conditions for ellipsoids, which are represented by a quadratic inequality.

2.2.2.1 Invariance Conditions for Discrete Systems

The S-lemma 7.2.2 and Proposition 1.1.3 are our main tools to obtain the invariance condition of an ellipsoid for a discrete system. First, we present a technical lemma.

Lemma 2.2.17. *Let Q be an $n \times n$ real symmetric matrix and let α be a given real number. Then $x^T Q x \geq \alpha$ for all $x \in \mathbb{R}^n$ if and only if $Q \succeq 0$, and $\alpha \leq 0$.*

Theorem 2.2.18. *An ellipsoid $\mathcal{E}$ given as in (7.5) is an invariant set for the discrete system (1.1) if and only if*

$$\exists\, \mu \in [0,1], \text{ such that } A_d^T Q A_d - \mu Q \preceq 0. \tag{2.6}$$

Proof. According to Proposition 1.1.3, to prove this theorem is equivalent to prove $\mathcal{E} \subseteq \mathcal{E}'$, where $\mathcal{E} = \{x \,|\, x^T Q x \leq 1\}$ and $\mathcal{E}' = \{x \,|\, x^T A_d^T Q A_d x \leq 1\}$. Clearly, $\mathcal{E} \subseteq \mathcal{E}'$ holds if and only if the following inequality system has no solution:

$$-x^T A_d^T Q A_d x + 1 < 0, \quad x^T Q x - 1 \leq 0. \tag{2.7}$$

Note that the left sides of the two inequalities in (2.7) are both quadratic functions, thus, according to the S-lemma, we have that (2.7) has no solution is equivalent to that there exists $\mu \geq 0$, such that $-x^T A_d^T Q A_d x + 1 + \mu(x^T Q x - 1) \geq 0$, or equivalently,

$$x^T(\mu Q - A_d^T Q A_d)x \geq \mu - 1, \quad \text{for all } x \in \mathbb{R}^n. \tag{2.8}$$

The theorem follows by applying Lemma 2.2.17 to (2.8). □

We can also consider an ellipsoid as an invariant set for a system in the following perspective. Invariance of a bounded set for a system is possible only if the system is non-expansive, which means that for discrete system (1.1), all eigenvalues of A_d are in a closed

unit disc of the complex plane. Then it becomes clear that (2.6) has a solution only if (1.1) is non-expansive, i.e., the trajectory of (1.1) is non-expansive. One can conclude from this that there is an invariant ellipsoid for (1.1) if and only if (2.6) has a solution for a positive definite Q. Moreover, the smallest μ solving (2.6) is the largest eigenvalue of $WA_d^TQA_dW$, where W is the symmetric positive definite square root of Q^{-1}, i.e., $W^2 = Q^{-1}$.

We now present two examples such that condition (2.6) does not hold for $\mu \notin [0,1]$. First, let Q be positive definite and $\mu < 0$, then $A_d^TQA_d - \mu Q$ is always a positive definite matrix. Thus condition (2.6) does not hold. Second, let Q be positive definite and $\mu > 1$, consider the discrete system $x_{k+1} = -x_k$. One can prove that $\{x \mid x^TQx \leq 1\}$ is an invariant set for this discrete system. However, in this case, we have $A_d^TQA_d - \mu Q = (1-\mu)Q$, which is always a negative definite matrix. Thus condition (2.6) does not hold either.

Apart from its simplicity, another advantage of the approach given in the proof of Theorem 2.2.18 is that it obtains a sufficient and necessary condition. Also, this approach highlights the close relationship between the theory of invariant sets and the Theorem of Alternatives, which is a fundamental result in the theory of optimization.

Corollary 2.2.19. *Condition (2.6) holds if and only if*

$$\exists\, \nu \in [0,1], \text{ such that } \tilde{Q} = \begin{pmatrix} Q^{-1} & A_d \\ A_d^T & \nu Q \end{pmatrix} \succeq 0. \tag{2.9}$$

Proof. First, $Q \succ 0$ yields $Q^{-1} \succ 0$. By Schur's lemma [18], $\tilde{Q} \succeq 0$ if and only if its Schur complement $\nu Q - A_d^T(Q^{-1})^{-1}A_d = \nu Q - A_d^TQA_d \succeq 0$, i.e., if (2.6) holds. □

Corollary 2.2.20. *Condition (2.6) holds if and only if*

$$A_d^TQA_d - Q \preceq 0. \tag{2.10}$$

Proof. The "if" part is immediate by letting $\mu = 1$ in (2.6). For the "only if" part, we let $\nu = 1 - \mu$, which, by reformulating (2.6), yields $A_d^TQA_d - Q \preceq -\nu Q \preceq 0$, for $\nu \in [0,1]$, where the second " $\preceq$" holds due to the fact that $\nu \geq 0$ and $Q \succ 0$. □

The left side of (2.10) is called the Lyapunov operator [17] in discrete form or Stein

transformation [65] in dynamical systems. Corollary 2.2.20 is consistent with the invariance condition of an ellipsoid for discrete systems given as in [13, 17]. The invariance condition presented in [13] is the same as (2.10) without the equality. This is since contractivity rather than invariance of a set for a system is analyzed in [13]. Lyapunov method is used to derive condition (2.10) in [17]. Apparently, condition (2.10) is easier to apply than condition (2.6), since the former one involves only the ellipsoid and the system.

The attentive reader may observe that the positive definiteness assumption for matrix Q is never used in the proof of Theorem 2.2.18. That assumption was only needed to ensure that the set $\mathcal{S}$ is convex. Recall that the quadratic functions in the S-lemma are not necessarily convex, thus we can extend Theorem 3.16 to more general sets which are represented by a quadratic inequality.

Theorem 2.2.21. *A set $\mathcal{S} = \{x \in \mathbb{R}^n \,|\, x^T Q x \leq 1\}$, where $Q \in \mathbb{R}^{n\times n}$, is an invariant set for the discrete system (1.1) if and only if*

$$\exists\, \mu \in [0,1], \text{ such that } A_d^T Q A_d - \mu Q \preceq 0. \tag{2.11}$$

The proof of Theorem 2.2.21 is the same as that of Theorem 2.2.18, so we do not duplicate that proof here. A trivial example that satisfy the condition in is given by choosing Q to be any indefinite matrix, $A_d = I$, and we choose $\mu = 1$. It is easy to see that for this choice condition (2.11) holds. Further exploring the implications of possibly using nonconvex and unbounded invariant sets is far from the main focus of this chapter, so this topic remains the subject of further research.

Verifying if an ellipsoid for $\mathcal{S} = \{x \in \mathbb{R}^n | x^T Q x \leq 1\}$ is an invariant set: Conditions (2.9) and (2.11) are semidefinite optimization feasibility problems, so they can be solved in polynomial time, e.g., by using SeDuMi [69].

2.2.2.2 Invariance Conditions for Continuous Systems

We first present an interesting result about the solution of continuous system.

Proposition 2.2.22. *The solution of the continuous system (1.2) is on the boundary of the ellipsoid $\mathcal{E}$ given as in (7.5) (or the Lorenz cone $\mathcal{C}_{\mathcal{L}}$ given as in (7.6)) whenever $x_0 \in \partial\mathcal{E}$ (or*

$x_0 \in \partial\mathcal{C}_\mathcal{L}$) *if and only if*

$$\sum_{i=0}^{k-1} \frac{1}{(k-1)!}\binom{k-1}{i}(A_c^i)^T Q A_c^{k-i-1} = 0, \ \textit{for } k = 2, 3, ... \tag{2.12}$$

Proof. We consider only ellipsoids, and the proof for Lorenz cones is analogous. The solution of (1.2) is given as $x(t) = e^{A_c t}x_0$, thus $x(t) \in \partial\mathcal{E}$ if and only if $x_0^T(e^{A_c t})^T Q e^{A_c t}x_0 = 1$, which can be expanded, by substituting $e^{A_c t} = \sum_{i=0}^{\infty} \frac{1}{i!}A_c^i t^i$, as

$$\sum_{k=1}^{\infty} t^{k-1} x_0^T \tilde{Q}_{k-1} x_0 = 1, \ \text{where } \tilde{Q}_{k-1} = \sum_{i=0}^{k-1} \frac{1}{(i)!(k-i-1)!}(A_c^i)^T Q A_c^{k-i-1},$$

for any $x_0^T Q x_0 = 1$ and $t \geq 0$. Thus, $\tilde{Q}_{k-1} = 0$, for $k \geq 2$. Also, note that $\frac{1}{(k-1)!}\binom{k-1}{i} = \frac{1}{(i)!(k-i-1)!}$, thus condition (2.12) is immediate. □

In particular, when $k = 2$, condition (2.12) yields $A_c^T Q + Q A_c = 0$. The left hand side of this equation is called Lyapunov operator in continuous form. The following invariance conditions is first given by Stern and Wolkowicz [68], where they consider only Lorenz cones and their proof is using the concept of cross-positivity. Here we present a simple proof.

Lemma 2.2.23. *[68] An ellipsoid $\mathcal{E}$ given in the form of (7.5) (or a Lorenz cone $\mathcal{C}_\mathcal{L}$ given in the form of (7.6)) is an invariant set for the continuous system (1.2) if and only if*

$$(A_c x)^T Q x \leq 0, \ \textit{for all } x \in \partial\mathcal{E} \ (\ \textit{or } x \in \partial\mathcal{C}_\mathcal{L}). \tag{2.13}$$

Proof. We consider only ellipsoids, and the proof is analogous for Lorenz cones. Note that $\partial\mathcal{E} = \{x \,|\, x^T Q x = 1\}$, thus the outer normal vector of $\mathcal{E}$ at $x \in \partial\mathcal{E}$ is Qx. Then we have $\mathcal{T}_\mathcal{E}(x) = \{y \,|\, y^T Q x \leq 0\}$, thus this theorem follows by Theorem 7.2.5. □

We now present a sufficient and necessary condition that an ellipsoid is invariant for the continuous system.

Theorem 2.2.24. *An ellipsoid $\mathcal{E}$ given as in (7.5) is an invariant set for the continuous system (1.2) if and only if*

$$A_c^T Q + Q A_c \preceq 0. \tag{2.14}$$

Proof. According to Lemma 2.2.23, we have that condition (2.13) holds, i.e., $\mathcal{E}$ is an invariant set for the continuous system if and only if

$$x^T(A_c^TQ + QA_c)x \leq 0, \text{ for all } x \in \partial\mathcal{E}. \tag{2.15}$$

Clearly (2.14) implies (2.15). Now assume (2.15) holds, then for all nonzero $y \in \mathbb{R}^n$, there exists an $x \in \partial\mathcal{E}$ and $\gamma > 0$, such that $y = \gamma x$. Then $y^T(A_c^TQ + QA_c)y = \frac{1}{\gamma^2}x^T(A_c^TQ + QA_c)x \leq 0$, which yields condition (2.14). □

The presented method in the proof of Theorem 2.2.24 is simpler than the traditional Lyapunov method to derive the invariance condition. However, the approach in the proof cannot be used for Lorenz cones, since the origin is not in the interior of Lorenz cones.

2.2.3 Lorenz Cones

A Lorenz cone $\mathcal{C}_\mathcal{L}$ given as in (7.6) can also be represented by a quadratic form, but the way to obtain the invariance condition of a Lorenz cone for discrete systems is much more complicated than that for an ellipsoid. The difficulty is mainly due to the existence of the second constraint in (7.6).

2.2.3.1 Invariance Conditions for Discrete Systems

The representation of the nonconvex set $\mathcal{C}_\mathcal{L} \cup (-\mathcal{C}_\mathcal{L}) = \{x \,|\, x^TQx \leq 0\}$ involves only the quadratic form, which is almost the same as an ellipsoid. We can first derive the invariance condition of this set for discrete system. Recall that the S-lemma does not require that the quadratic functions have to be convex, thus the S-lemma is still valid for the nonconvex set.

Theorem 2.2.25. *The nonconvex set $\mathcal{C}_\mathcal{L} \cup (-\mathcal{C}_\mathcal{L})$ is an invariant set for the discrete system (1.1) if and only if*

$$\exists\, \mu \geq 0, \text{ such that } A_d^TQA_d - \mu Q \preceq 0. \tag{2.16}$$

Proof. The proof is closely following the ideas in the proof of Theorem 2.2.18. The only difference is that the right side in (2.8) is 0 rather than $1 - \mu$, which is why the condition

$\mu \leq 1$ is absent in this case. $\square$

The invariance condition for $\mathcal{C}_{\mathcal{L}} \cup (-\mathcal{C}_{\mathcal{L}})$ shown in (2.16) is similar to the one proposed by Loewy and Schneider in [49]. They proved by contradiction using the properties of copositive matrices that when the rank of A_d is greater than 1, $A_d\mathcal{C}_{\mathcal{L}} \subseteq \mathcal{C}_{\mathcal{L}}$ or $-A_d\mathcal{C}_{\mathcal{L}} \subseteq \mathcal{C}_{\mathcal{L}}$ if and only if (2.16) holds. They also concluded (see [49, Lemma 3.1]) that when the rank of A_d is 1, $A_d\mathcal{C}_{\mathcal{L}} \subseteq \mathcal{C}_{\mathcal{L}}$ if and only if there exist two vectors $x, y \in \mathcal{C}_{\mathcal{L}}$, such that $A_d = xy^T$.

The following example shows that for the given A_d and Q, only $\mu = 1$ satisfies condition (2.16). Let $A_d = Q = \text{diag}\{1, ..., 1, -1\}$. Then the Lorenz cone is an invariant set for the system, since such a Lorenz cone is a self-dual cone[3]. The left hand side in (2.16) is, however, now simplified to $(1 - \mu)Q$ which is negative semidefinite only for $\mu = 1$, because inertia$\{Q\} = \{n - 1, 0, 1\}$.

In the case of ellipsoids, we used Schur's lemma, see, e.g., [60], to simplify invariance condition (2.6) to (2.9), which was further simplified to the parameter free invariance condition (2.10). Although conditions (2.6) and (2.16) are similar, it seems to be impossible to develop a parameter free condition analogous to (2.10) for Lorenz cones. This is due to the fact that matrix Q for a Lorenz cone is neither positive nor negative semidefinite.

To find the scalar μ in (2.16) is essentially a semidefinite optimization (SDO) problem, and it is shown to be solve in polynomial time, see, e.g., [77]. Various celebrated SDO solvers, e.g., SeDuMi [69], CVX [28], and SDPT3 [75] have been shown robust performance in solving a SDO problems numerically.

Corollary 2.2.26. *If $\lambda_1(A_d^T Q A_d) \leq 0$, then the Lorenz cone $\mathcal{C}_{\mathcal{L}}$ given as in (7.6) is an invariant set for the discrete system (1.1).*

Corollary 2.2.26 gives a simple sufficient condition such that a Lorenz cone is an invariant set, but it is only valid when the A is a singular matrix. In fact, if A is nonsingular, by Sylvester's law of inertia [36], we have that $\lambda_1(A_d^T Q A_d) > 0$. When $\lambda_1(A_d^T Q A_d) \leq 0$, we have that the rank of A_d is no more than 1. This is because, if the rank is larger than 1, then range$(A_d) \cap \text{span}\{u_1, u_2, ..., u_{n-1}\}$ must be a nonzero subspace, because for example

[3] A self-dual cone is a cone that coincides with its dual cone, where the dual cone for a cone $\mathcal{C}$ is defined as $\{y \mid x^T y \geq 0, \forall x \in \mathcal{C}\}$.

$A_d x$ is a nonzero vector in the intersection. Then $x^T(A_d^T Q A_d)x > 0$, so we cannot have $\lambda_1(A_d^T Q A_d) < 0$.

The interval of the scalar μ in (2.16) can be tightened by incorporating the eigenvalues and eigenvectors of Q. Such a tighter condition is presented in Corollary 2.2.27.

Corollary 2.2.27. *If condition (2.16) holds, then*

$$\max\left\{0, \max_{1\le i\le n-1}\left\{\frac{u_i^T A_d^T Q A_d u_i}{\lambda_i}\right\}\right\} \le \mu \le \frac{u_n^T A_d^T Q A_d u_n}{\lambda_n}. \tag{2.17}$$

Proof. Multiplying condition (2.16) by u_i^T from the left and u_i from the right, we have $u_i^T A_d^T Q A_d u_i - \mu u_i^T Q u_i \le 0$. Since $u_i^T Q u_i = \lambda_i u_i^T u_i = \lambda_i > 0$, for $i \in \mathcal{I}(n-1)$, and $u_n^T Q u_n = \lambda_n < 0$, condition (2.17) follows immediately. □

Corollary 2.2.27 presents tighter bounds for the scalar μ in (2.17) in terms of an algebraic form. The existence of a scalar μ implies that the upper bound should be no less than the lower bound in (2.17). However, this is not always true. We now present a geometrical interpretation of the interval of the scalar μ, that can be directly derived from Corollary 2.2.27.

Corollary 2.2.28. *The relationship between the vector $A_d u_i$, and the scalars $u_i^T A_d^T Q A_d u_i$, and μ are as follows:*

- *If $A_d u_n \notin \mathcal{C}_\mathcal{L} \cup (-\mathcal{C}_\mathcal{L})$, then μ satisfying (2.17) does not exist.*
- *If $A_d u_i \in \mathcal{C}_\mathcal{L} \cup (-\mathcal{C}_\mathcal{L})$ for all $i \in \mathcal{I}(n-1)$, then*
  - *if $A_d u_n \in \partial\mathcal{C}_\mathcal{L} \cup (-\partial\mathcal{C}_\mathcal{L})$ and (2.17) holds, then $\mu = 0$.*
  - *if $A_d u_n \in \text{int } \mathcal{C}_\mathcal{L} \cup (-\text{int } \mathcal{C}_\mathcal{L})$ and (2.17) holds, then $\mu \in \left[0, \frac{u_n^T A_d^T Q A_d u_n}{\lambda_n}\right]$.*
- *Let $\mathcal{I} = \{i \mid A_d u_i \notin \mathcal{C}_\mathcal{L} \cup (-\mathcal{C}_\mathcal{L})\}$. If the set $\mathcal{I} \subseteq \mathcal{I}(n-1)$ is nonempty, then*
  - *if $A_d u_n \in \partial\mathcal{C}_\mathcal{L} \cup (-\partial\mathcal{C}_\mathcal{L})$, then μ satisfying (2.17) does not exist.*
  - *if $A_d u_n \in \text{int } (\mathcal{C}_\mathcal{L}) \cup (-\text{int}(\mathcal{C}_\mathcal{L}))$, then*
    - *if there exist $i^* \in \mathcal{I}$, such that $\frac{u_{i^*}^T A_d^T Q A_d u_{i^*}}{\lambda_{i^*}} > \frac{u_n^T A_d^T Q A_d u_n}{\lambda_n}$, then μ satisfying (2.17) does not exist.*

$*$ *otherwise, if (2.17) holds, then* $\mu \in \Big[\max_{i\in\mathcal{I}}\Big\{\frac{u_i^T A_d^T Q A_d u_i}{\lambda_i}\Big\}, \frac{u_n^T A_d^T Q A_d u_n}{\lambda_n}\Big]$.

We now consider the invariance condition of a Lorenz cone $\mathcal{C}_\mathcal{L}$ given as in (7.6), which is a convex cone and can handle expansive systems.

Lemma 2.2.29. [68] *A Lorenz cone* $\mathcal{C}_\mathcal{L}$*, given as in (7.6), can be written as* $T\mathcal{C}_\mathcal{L}^*$*, where* $\mathcal{C}_\mathcal{L}^*$ *is the standard Lorenz cone and* T *is the nonsingular matrix,*

$$T = \Big[\frac{u_1}{\sqrt{\lambda_1}}, ..., \frac{u_{n-1}}{\sqrt{\lambda_{n-1}}}, \frac{u_n}{\sqrt{-\lambda_n}}\Big]. \tag{2.18}$$

Lemma 2.2.30. *A Lorenz cone* $\mathcal{C}_\mathcal{L}$ *given as in (7.6) is an invariant set for the discrete system (1.1) if and only if the standard Lorenz cone* $\mathcal{C}_\mathcal{L}^*$ *is an invariant set for the following discrete system*

$$x_{k+1} = T^{-1} A_d T x_k, \tag{2.19}$$

where T *is defined by (2.18).*

Proof. The Lorenz cone $\mathcal{C}_\mathcal{L}$ is an invariant set for (1.1) if and only if $A_d\mathcal{C}_\mathcal{L} \subseteq \mathcal{C}_\mathcal{L}$. This holds if and only if $A_d T\mathcal{C}_\mathcal{L}^* \subseteq T\mathcal{C}_\mathcal{L}^*$, which is equivalent to $T^{-1}A_d T\mathcal{C}_\mathcal{L}^* \subseteq \mathcal{C}_\mathcal{L}^*$. □

The invariance condition of a Lorenz cone for discrete systems is presented in Theorem 2.2.31. Although we have developed such invariance condition independently, it was brought to our attention recently that the invariance condition is the same as the one proposed by Aliluiko and Mazko in [1]. But our proof is more straightforward.

Theorem 2.2.31. *A Lorenz cone* $\mathcal{C}_\mathcal{L}$ *(or* $-\mathcal{C}_\mathcal{L}$*), given as in (7.6), is an invariant set for the discrete system (1.1) if and only if*

$$u_n^T A_d u_n \geq 0,\ u_n^T A_d Q^{-1} A_d^T u_n \leq 0 \text{ and } \exists\, \mu \geq 0, \text{ such that } A_d^T Q A_d - \mu Q \preceq 0, \tag{2.20}$$

where u_n *is the eigenvector corresponding to the unique negative eigenvalue* λ_n *of* Q*.*

Proof. Since $A_d\mathcal{C}_\mathcal{L} \subseteq \mathcal{C}_\mathcal{L}$ if and only $A_d(-\mathcal{C}_\mathcal{L}) \subseteq -\mathcal{C}_\mathcal{L}$, we only present the proof for $\mathcal{C}_\mathcal{L}$. For an arbitrary $x \in \mathcal{C}_\mathcal{L}$, by Theorem 2.2.25, we have that $A_d x \in \mathcal{C}_\mathcal{L}$ or $A_d x \in -\mathcal{C}_\mathcal{L}$ if and

only if condition (2.16) is satisfied. To ensure that only $A_d x \in \mathcal{C}_{\mathcal{L}}$ holds, some additional conditions should be added.

According to Lemma 2.2.30, we may consider $\mathcal{C}_{\mathcal{L}}^*$ and the discrete system (2.19), where the coefficient matrix, denoted by $\tilde{A}$, can be explicitly written as

$$\tilde{A} = T^{-1} A_d T = \begin{bmatrix} u_1^T A_d u_1 & \cdots & \sqrt{-\frac{\lambda_1}{\lambda_n}} u_1^T A_d u_n \\ \vdots & \ddots & \vdots \\ \sqrt{-\frac{\lambda_n}{\lambda_1}} u_n^T A_d u_1 & \cdots & u_n^T A_d u_n \end{bmatrix}.$$

Then, according to Theorem 2.2.25, condition (2.16) is equivalent to

$$\exists\, \mu \geq 0, \text{ such that } (T^{-1}AT)^T \tilde{I} T^{-1} A T - \mu \tilde{I} \preceq 0, \tag{2.21}$$

where $\tilde{I} = \text{diag}\{1, ..., 1, -1\}$. Note that $T^T Q T = \tilde{I}$, condition (2.21) is equivalent to

$$\exists\, \mu \geq 0, \text{ such that } A_d^T Q A_d - \mu Q \preceq 0.$$

Recall that we denote the i-th row of a matrix M by M_i^T. Also, the second constraint in the formulae of $\mathcal{C}_{\mathcal{L}}^*$ requires that for every $x \in \mathcal{C}_{\mathcal{L}}^*$ the last coordinate in x is nonnegative. Since $\tilde{A}\mathcal{C}_{\mathcal{L}}^* \subseteq \mathcal{C}_{\mathcal{L}}^*$, we have $\tilde{A}_n^T x \geq 0, \text{ for all } x \in \mathcal{C}_{\mathcal{L}}^*$. Note that $\mathcal{C}_{\mathcal{L}}^*$ is a self-dual cone, we have $\tilde{A}_n^T x \geq 0, \text{ for all } x \in \mathcal{C}_{\mathcal{L}}^*$ if and only if $\tilde{A}_n \in \mathcal{C}_{\mathcal{L}}^*$. Now we have

$$\tilde{A}_n^T = \sqrt{-\lambda_n}\Big(\frac{1}{\sqrt{\lambda_1}} u_n^T A_d u_1, \frac{1}{\sqrt{\lambda_2}} u_n^T A_d u_2, ..., \frac{1}{\sqrt{-\lambda_n}} u_n^T A_d u_n\Big) = \sqrt{-\lambda_n} u_n^T A_d T. \tag{2.22}$$

Substituting the value of $\tilde{A}_n^T$ given by the right side of (2.22) into the first inequality in the formulae of $\mathcal{C}_{\mathcal{L}}^*$, we have

$$-\lambda_n (T^T A_d^T u_n)^T \tilde{I} (T^T A_d^T u_n) \leq 0. \tag{2.23}$$

Since $\lambda_n < 0$ and $T\tilde{I}T^T = \sum_{i=1}^n \frac{u_i u_i^T}{\lambda_i} = Q^{-1}$, where the second equality is due to the spectral decomposition of Q^{-1}, we have that (2.23) is equivalent to $u_n^T A_d Q^{-1} A_d^T u_n \leq 0$. Also, substituting (2.22) into the second inequality in the formulae of $\mathcal{C}_{\mathcal{L}}^*$ yields $u_n^T A_d u_n \geq 0$. The proof is complete. □

Remark 2.2.32. *The inequality system* $u_n^T A_d Q^{-1} A_d^T u_n \leq 0$ *and* $u_n^T A_d u_n \geq 0$ *holds if and only if* $u_n^T A_d x \geq 0$*, for all* $x \in \mathcal{C}_\mathcal{L}$.

Proof. Since $x^T Q x \leq 0$ can be written as $x^T U \Lambda^{\frac{1}{2}} \tilde{I} \Lambda^{\frac{1}{2}} U^T x \leq 0$, we have $x \in \mathcal{C}_\mathcal{L}$ if and only if $\Lambda^{\frac{1}{2}} U^T x \in \mathcal{C}_\mathcal{L}^*$. Similarly, since $Q^{-1} = U \Lambda^{-\frac{1}{2}} \tilde{I} \Lambda^{-\frac{1}{2}} U^T$, we have that $u_n^T A_d Q^{-1} A_d^T u_n \leq 0$ can be written as $u_n^T A_d U \Lambda^{-\frac{1}{2}} \tilde{I} \Lambda^{-\frac{1}{2}} U^T A_d^T u_n \leq 0$, which yields $\Lambda^{-\frac{1}{2}} U^T A_d^T u_n \in \mathcal{C}_\mathcal{L}^* \cup (-\mathcal{C}_\mathcal{L}^*)$. Since the set $\mathcal{C}_\mathcal{L}^*$ is a self-dual cone, we have $(\Lambda^{-\frac{1}{2}} U^T A_d^T u_n)^T (\Lambda^{\frac{1}{2}} U^T x) \geq 0$, which can be simplified to $u_n^T A_d x \geq 0$, for all $x \in \mathcal{C}_\mathcal{L}$. □

The normal plane of the eigenvector u_n that contains the origin separates $\mathbb{R}^n$ into two half spaces. Corollary 2.2.32 presents a geometrical interpretation that A_d transforms the Lorenz cone $\mathcal{C}_\mathcal{L}$ to the half space that contains eigenvector u_n, i.e., $A_d \mathcal{C}_\mathcal{L} \subseteq \{y \mid u_n^T y \geq 0\}$. Moreover, note that $u_n^T A_c x = (A_d^T u_n)^T x$, which shows that the vector $A_d^T u_n$ is in the dual cone of $\mathcal{C}_\mathcal{L}$.

Corollary 2.2.33. *If condition (2.20) holds, then*

$$0 \leq \mu \leq \frac{u_n^T A_d^T Q A_d u_n}{\lambda_n}. \tag{2.24}$$

Proof. The proof is analogous to the one given in the proof of Corollary 2.2.27. □

The interval for the scalar μ in condition (2.24) is wider and simpler than the one presented in Corollary 2.2.27. Analogous to Corollary 2.2.28, we present an intuitive geometrical interpretation of μ for Lorenz cones.

Corollary 2.2.34. *The relationship between the vector* $A_c u_n$*, and the scalars* $u_n^T A_d^T Q A_d u_n$*, and* μ *are as follows:*

- *If* $A_d u_n \notin \mathcal{C}_\mathcal{L} \cup (-\mathcal{C}_\mathcal{L})$*, then* μ *satisfying (2.24) does not exist.*
- *If* $A_d u_n \in \partial \mathcal{C}_\mathcal{L} \cup (-\partial \mathcal{C}_\mathcal{L})$ *and (2.24) holds, then* $\mu = 0$.
- *If* $A_d u_n \in \text{int } (\mathcal{C}_\mathcal{L}) \cup (-\text{int } (\mathcal{C}_\mathcal{L}))$ *and (2.24) holds, then* $\mu \in \left[0, \frac{u_n^T A_d^T Q A_d u_n}{\lambda_n}\right]$.

2.2.3.2 Invariance Conditions for Continuous Systems

Now we consider the invariance condition of Lorenz cones for continuous systems. We also need to analyze the eigenvalue of a sum of two symmetric matrices for the invariance conditions for continuous systems. The following lemma is a useful tool in our analysis. It shows the fact that the spectrum of a matrix is stable under a small perturbation by another matrix. Since the statement is obvious, we omit the proof.

Lemma 2.2.35. *Let M and N be two symmetric matrices. Then*

- *if there exists a $\hat{\tau} > 0$, such that $M + \tau N \preceq 0$, for $0 < \tau \leq \hat{\tau}$, then $M \preceq 0$.*
- *if $M \prec 0$, then there exists a $\hat{\tau} > 0$, such that $M + \tau N \preceq 0$, for $0 < \tau \leq \hat{\tau}$.*

Similar to the case for discrete system, we first consider the invariance condition of the nonconvex set $\mathcal{C}_{\mathcal{L}} \cup (-\mathcal{C}_{\mathcal{L}})$ for the continuous system.

Theorem 2.2.36. *The nonconvex set $\mathcal{C}_{\mathcal{L}} \cup (-\mathcal{C}_{\mathcal{L}})$ is an invariant set for the continuous system (1.2) if and only if*

$$\exists\, \eta \in \mathbb{R},\ \textit{such that } A_c^T Q + Q A_c - \eta Q \preceq 0. \tag{2.25}$$

Proof. For the "if" part, i.e., condition (2.25) holds, then for every $x \in \partial\mathcal{C}_{\mathcal{L}} \cup (-\partial\mathcal{C}_{\mathcal{L}})$, we have $(A_c x)^T Q x = (A_c x)^T Q x - \frac{\eta}{2} x^T Q x = \frac{1}{2} x^T (A_c^T Q + Q A_c - \eta Q) x \leq 0$. Thus, by Nagumo's Theorem 7.2.5, the set $\mathcal{C}_{\mathcal{L}} \cup (-\mathcal{C}_{\mathcal{L}})$ is an invariant set for continuous system.

Next, we prove the "only if" part. According to Theorem 2.2.1, there exists a $\hat{\tau} > 0$, such that for every $0 \leq \Delta t \leq \hat{\tau}$, $\mathcal{C}_{\mathcal{L}} \cup (-\mathcal{C}_{\mathcal{L}})$ is also an invariant set for $x_{k+1} = (I - A_c \Delta t)^{-1} x_k$. By Theorem 2.2.25 and $(I - A_c \Delta t)^{-1} = I + A_c \Delta t + A_c^2 \Delta t^2 + \cdots$, we have

$$\exists\ \mu(\Delta t) \geq 0,\ \text{such that } \frac{1 - \mu(\Delta t)}{\Delta t} Q + (A_c^T Q + Q A_c) + \Delta t K(\Delta t) \preceq 0,$$

where $K(\Delta t) = (A_c^T Q A_c + (A_c^2)^T Q + Q A_c^2) + \Delta t((A_c^2)^T Q A_c + A_c^T Q A_c^2 + (A_c^3)^T Q + Q A_c^3) + \cdots$. Since Q and A_c are constant matrices, and applying the fact that $\|M\| = \|M^T\|$, $\|M+N\| \leq$

$\|M\| + \|N\|$ and $\|MN\| \leq \|M\|\|N\|$, we have

$$\begin{aligned}\|K(\Delta t)\| &\leq \sum_{i=3}^{\infty} i\|Q\|\|A_c\|^{i-1}(\Delta t)^{i-3} = \|Q\|\|A_c\|^2 \sum_{i=0}^{\infty}(i+3)(\Delta t\|A_c\|)^i \\ &= \|Q\|\|A_c\|^2 \frac{3 - 2\Delta t\|A_c\|}{(1-\Delta t\|A_c\|)^2} \leq 8\|Q\|\|A_c\|^2,\end{aligned}$$

where $\Delta t \leq \frac{5}{4}\|A\|^{-1}$ such that $(3 - 2\Delta t\|A_c\|)/(1 - \Delta t\|A_c\|)^2 \leq 8$. Also, applying the relationship between spectral radius $\rho(A_c)$ and its induced norm, $\rho(A_c) \leq \|A_c\|$ (see [21]), to $K(\Delta t)$, we have

$$|\lambda_i(K(\Delta t))| \leq \rho(K(\Delta t)) \leq \|K(\Delta t)\| \leq 8\|Q\|\|A_c\|^2, \text{ for } i \in \mathcal{I}(n),$$

i.e., the eigenvalues of $K(\Delta t)$ are bounded. Let us denote $\eta(\Delta t) = \frac{\mu(\Delta t)-1}{\Delta t}$.Then according to Lemma 2.2.35 and taking $\Delta t \to 0$, we have

$$A_c^T Q + Q A_c - \eta(\Delta t) Q \preceq 0. \tag{2.26}$$

According to (2.26), we have $\eta(\Delta t)$ is bounded for all Δt. Therefore[4], we can take a subsequence $\{\Delta t_\ell\}$ such that $\eta(\Delta t_\ell) \to \eta$ as $\Delta t_\ell \to 0$, which yields (2.25). The proof is complete. □

The approach in the proof of Theorem 2.2.36 can be also used to prove Theorem 2.2.24. The only remaining invariance condition is the one of a Lorenz cone for continuous system.

Theorem 2.2.37. *A Lorenz cone $\mathcal{C}_\mathcal{L}$ (or $-\mathcal{C}_\mathcal{L}$) is an invariant set for the continuous system (1.2) if and only if (2.25) holds.*

Proof. Consider the continuous system with $x_0 \in \mathcal{C}_\mathcal{L}$, according to Theorem 2.2.36, the trajectory $x(t)$ will stay in $\mathcal{C}_\mathcal{L} \cup (-\mathcal{C}_\mathcal{L})$ if condition (2.25) is satisfied. If $x(t)$ would move over to $-\mathcal{C}_\mathcal{L}$, then $x(t)$ must go through the origin, i.e., $x(t^*) = 0$ for some $t^* \geq 0$. Note that $x(t) = e^{A_c(t-t^*)}x(t^*) = 0$ for any $t > t^*$, i.e., the origin is an equilibrium point, which means $\mathcal{C}_\mathcal{L}$ is an invariant set for the continuous system. Thus the theorem is immediate. □

[4]Here we use the fact that every bounded sequence in a Euclidean space has a convergent subsequence, see, e.g., [61].

In fact, a direct proof of Theorem 2.2.37 can be given as follows: one can also prove that the second and third conditions in (2.20) hold by choosing sufficiently small Δt. To be specific, for the second condition in (2.20), we have

$$u_n^T(I-\Delta t A_c)^{-1}u_n \geq 0, \text{ if and only if } \|u_n\|^2+\sum_{i=1}^{\infty}(\Delta t)^i u_n^T A_c^i u_n \geq 0, \tag{2.27}$$

where the second term, when $\Delta t < \|A\|^{-1}$, can be bounded as follows: $\left|\sum_{i=1}^{\infty}(\Delta t)^i u_n^T A_c^i u_n\right| \leq \|u_n\|^2 \frac{\Delta t\|A_c\|}{(1-\Delta t\|A_c\|)}$. Thus, we can choose $\Delta t < 0.5\|A_c\|^{-1}$, such that condition (2.27) holds. Similarly, the third condition in (2.20) can be transformed to

$$u_n^T(I-\Delta t A_c)^{-1}Q^{-1}(I-\Delta t A_c)^{-T}u_n \leq 0, \text{ if and only if } \frac{1}{\lambda_n}\|u_n\|^2+K(\Delta t)\leq 0, \tag{2.28}$$

where we use the fact that u_n is the eigenvector corresponding to the eigenvalue λ_n^{-1} of Q^{-1}, and $K(\Delta t) = \Delta t u_n^T(A_cQ^{-1}+Q^{-1}A_c^T)u_n + (\Delta t)^2 u_n^T(A_cQ^{-1}A_c + A_c^2Q^{-1}+Q^{-1}A_c^{2T})u_n + \cdots$. We note that inertia$\{Q\} = \{n-1,0,1\}$ implies inertia$\{Q^{-1})\} = \{n-1,0,1\}$, then we have that Q^{-1} exists, which yields the following: $|K(\Delta t)| \leq \|u\|^2(2\Delta t\|A_c\|\|Q^{-1}\| + 3\Delta t^2\|A_c\|^2\|Q^{-1}\|+\cdots) = \|u\|^2\|Q^{-1}\|\frac{2\Delta t\|A_c\|-(\Delta t\|A_c\|)^2}{(1-\Delta t\|A_c\|)^2}$. We can choose $\Delta t \leq \min\{0.5\|A_c\|^{-1}, (\|A_c\|(1-4\lambda_n\|Q^{-1}\|)^{-1}\}$, such that (2.28) holds. In fact,

$$\begin{aligned}\frac{1}{\lambda_n}\|u_n\|^2+K(\Delta t) &\leq \|u_k\|^2\Big(\frac{1}{\lambda_n}+\|Q^{-1}\|\frac{2\Delta t\|A_c\|-(\Delta t\|A_c\|)^2}{(1-\Delta t\|A_c\|)^2}\Big)\\ &\leq \|u\|^2\Big(\frac{1}{\lambda_n}+\|Q^{-1}\|\frac{4\Delta t\|A_c\|}{1-\Delta t\|A_c\|}\Big)\leq 0.\end{aligned}$$

Condition (2.25) is the same as the one presented in [68], where the proof is much more complicated than ours. Finding the value of η in Theorem 2.2.36 and 2.2.37 is essentially a semidefinite optimization problem. For example, we can use the following semidefinite optimization problem:

$$\max\{\eta\in\mathbb{R} \mid A_c^TQ+QA_c-\eta Q \preceq 0\}. \tag{2.29}$$

When the optimal solution η^* of (2.29) exists, then by Theorem 2.2.37 we can claim that the Lorenz cone is an invariant set for the continuous system. Various celebrated SDO solvers, e.g., SeDuMi, CVX, and SDPT3, can be used to solve SDO problem (2.29).

Corollary 2.2.38. *If condition (2.25) holds, then*

$$\max_{1\le i\le n-1}\left\{u_i^T(A_c^T+A_c)u_i\right\}\le\eta\le u_n^T(A_c^T+A_c)u_n. \tag{2.30}$$

Proof. The proof is similar to the one presented in the proof of Corollary 2.2.27 by noting that $u_i^T(A_c^TQ+QA_c)u_i=2(A_cu_i)^TQu_i$, and $Qu_i=\lambda_iu_i$. □

We now present some simple examples to illustrate the invariance conditions presented in Section 2.2. Since it is straightforward for discrete systems, we only present examples for continuous systems. The following two examples consider polyhedral sets for continuous systems.

Example 2.2.39. *Consider the polyhedron* $\mathcal{P}=\{(\xi,\eta)\,|\,\xi+\eta\le 1,-\xi+\eta\le 1,\xi-\eta\le 1,-\xi-\eta\le 1\}$, *and the continuous system* $\dot{\xi}=-\xi,\dot{\eta}=-\eta$.

The solution of the system is $\xi(t)=\xi_0e^{-t},\eta(t)=\eta_0e^{-t}$, so $(\xi(t),\eta(t))\in\mathcal{P}$ for all $t\ge 0$, i.e., the polyhedron is an invariant set for the continuous system provided that $(\xi_0,\eta_0)\in\mathcal{P}$. This can also be verified by Theorem 2.2.8. We have

$$H=-I_4,\quad G=\begin{bmatrix}1&1\\-1&1\\1&-1\\-1&-1\end{bmatrix},\quad b=\begin{bmatrix}1\\1\\1\\1\end{bmatrix},\quad A_c=-I_2,$$

which satisfy $HG=GA_c$ and $Hb\le 0$. Thus Theorem 2.2.8 yields that $\mathcal{P}$ is an invariant set for this continuous system.

Example 2.2.40. *Consider the polyhedral cone* $\mathcal{C}_\mathcal{P}$ *generated by the extreme rays* $x^1=(1,1,1)^T,x^2=(-1,1,1)^T,x^3=(1,-1,1)^T$, *and* $x^4=(-1,-1,1)^T$, *and the continuous system* $\dot{\xi}=\xi,\dot{\eta}=\eta,\dot{\zeta}=\zeta$.

The solution of the system is $\xi(t)=\xi_0e^t,\ \eta(t)=\eta_0e^t,\ \zeta(t)=\zeta_0e^t$, thus one can easily verify that the polyhedral cone is an invariant set for this continuous system provided that

$(\xi_0, \eta_0, \zeta_0) \in \mathcal{C}_{\mathcal{P}}$. This can also be verified by Corollary 2.2.16. We have

$$X = \begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \ \tilde{L} = I_4, \ A_c = I_3,$$

which satisfy that $X\tilde{L} = A_c X$. Thus Corollary 2.2.16 yields that $\mathcal{C}_{\mathcal{P}}$ is an invariance set for this continuous system.

The following two examples consider ellipsoids and Lorenz cones for continuous systems.

Example 2.2.41. *Consider the ellipsoid* $\mathcal{E} = \{(\xi, \eta) \mid \xi^2 + \eta^2 \leq 1\}$*, and the system* $\dot{\xi} = -\eta, \dot{\eta} = \xi$.

The solution of the system is $\xi(t) = \alpha \cos t + \beta \sin t$ and $\eta(t) = \alpha \sin t - \beta \cos t$, where α, β are two parameters depending on the initial condition. The solution trajectory is a circle, thus the system is invariant on this ellipsoid. Also, we have

$$A_c = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad Q = I_2, \quad A_c^T Q + Q A_c = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \preceq 0,$$

which shows that, according to Theorem 2.2.24, the ellipsoid is an invariant set for this continuous system.

Example 2.2.42. *Consider the Lorenz cone* $\mathcal{C}_{\mathcal{L}} = \{(\xi, \eta, \zeta) \mid \xi^2 + \eta^2 \leq \zeta^2, \zeta \geq 0\}$*, and the system* $\dot{\xi} = \xi - \eta, \dot{\eta} = \xi + \eta, \dot{\zeta} = \zeta$.

The solution is $\xi(t) = e^t(\alpha \cos t + \beta \sin t)$, $\eta(t) = e^t(\alpha \sin t - \beta \cos t)$ and $\zeta(t) = \gamma e^t$, where α, β, γ are three parameters depending on the initial condition. It is easy to verify that this Lorenz cone is an invariant set for the continuous system. Also, by letting $\eta \leq -2$, we have

$$A_c = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad Q = I_3, \quad A_c^T Q + Q A_c + \eta Q = \begin{bmatrix} \eta + 2 & 0 & 0 \\ 0 & \eta + 2 & 0 \\ 0 & 0 & \eta + 2 \end{bmatrix} \preceq 0,$$

which shows that, according to Theorem 2.2.37, the Lorenz cone is an invariant set for this continuous system.

Chapter 3

Steplength Threshold for Invariance Preserving

3.1 Introduction

In this chapter, steplength thresholds for invariance preserving of three types of discretization methods on a polyhedron are considered. First, 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. Second, 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 latter two types of discretization methods and the forward Euler method is studied.

In this chapter, 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. Let us rewrite Definition 7.1.1 as follows:

Definition 3.1.1. *A polyhedron $\mathcal{P}$ in $\mathbb{R}^n$ is defined as*

$$\mathcal{P} = \{x \in \mathbb{R}^n \,|\, g_1^T x \leq b_1, g_2^T x \leq b_2, ..., g_m^T x \leq b_m\} := \{x \in \mathbb{R}^n \,|\, Gx \leq b\}, \tag{3.1}$$

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 (3.1), 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., [8, 19]. A novel and unified approach to derive these invariance conditions is proposed in Chapter 2. We will use to Theorem 2.2.8 and Theorem 2.2.4 in our analysis.

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 in Theorem 2.2.8 and Theorem 2.2.4 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 3.1.2. *Assume a polyhedron $\mathcal{P}$ is an invariant set for the continuous system (1.2), 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 τ is a **uniform invariance preserving steplength threshold** of this discretization method on $\mathcal{P}$.*

For simplicity, we use steplength threshold to indicate uniform invariance steplength threshold for simplicity throughout this chapter. The steplength threshold in Definition

4.1.1 implies that any value smaller than this threshold is also a valid steplength threshold. 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 Δt in this interval, which means that there are infinitely many values to be considered. 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 τ_1. We also note that the steplength threshold in Definition 4.1.1 is uniform on $\mathcal{P}$, i.e., τ needs to be a valid steplenth threshold for every initial point in $\mathcal{P}$. This is another key reason why the problem of finding a valid steplength threshold is not an easy problem.

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 chapter:

> *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* $\mathcal{P}$.

3.2 Computing Steplength Threshold

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: forward Euler method, Taylor approximation type discretization methods, and rational function type discretizatin methods. For the forward Euler method, we derive the largest steplength threshold for invariance preserving. 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.

3.2.1 Forward Euler Method

As an 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.,

$$\mathcal{P} = \text{conv}\{x^1, x^2, ..., x^\ell\} = \Big\{ x \,|\, x = \sum_{i=1}^{\ell} \lambda_i x^i, \quad \sum_{i=1}^{\ell} \lambda_i = 1, \quad \lambda_i \geq 0 \Big\}, \tag{3.2}$$

where $\{x^i\}$ are the vertices of $\mathcal{P}$. For polytope, we have a simple invariance condition, which relies on a simple form of tangent cone. A sufficient and necessary condition under which a polytope is an invariant set for the continuous system is presented in Lemma 3.2.1.

Lemma 3.2.1. *[10] The polytope $\mathcal{P}$ defined as in (3.2) is an invariant set for the continuous system (1.2) 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 at x^i, which can be given*

$$\mathcal{T}_{\mathcal{P}}(x^i) = \{ y \,|\, y = \sum_{j \neq i} \gamma_j (x^j - x^i), \ \gamma_j \geq 0 \}. \tag{3.3}$$

Corollary 3.2.2. *The polyhedron $\mathcal{P}$ defined as in (3.2) is an invariant set for the continuous system (1.2) if and only if there exist $\gamma_j^{(i)} \geq 0, j = 1, 2, ..., \ell$, such that*

$$A_c x^i = \sum_{j \neq i} \gamma_j^{(i)} (x^j - x^i), \ \textit{for all } i = 1, 2, ..., \ell. \tag{3.4}$$

Let $\epsilon^i = \big(\sum_{j \neq i} \gamma_j^{(i)} \big)^{-1}$ for $i = 1, 2, ..., \ell$ (let $\epsilon^i = \infty$, when $\sum_{j \neq i} \gamma_j^{(i)} = 0$), then

$$x^i + \Delta t A_c x^i \in \mathcal{P} \textit{ for any } \Delta t \in [0, \epsilon^i]. \tag{3.5}$$

Proof. According to Lemma 3.2.1 and Equation (3.3), Equation (3.4) is immediate. According to (3.4) and $\epsilon^i \sum_{j \neq i} \gamma_j^{(i)} = 1$, we have

$$\epsilon^i A_c x^i = \sum_{j \neq i} \epsilon^i \gamma_j^{(i)} (x^j - x^i) = \sum_{j \neq i} \epsilon^i \gamma_j^{(i)} x^j - \sum_{j \neq i} \epsilon^i \gamma_j^{(i)} x^i = \sum_{j \neq i} \epsilon^i \gamma_j^{(i)} x^j - x^i. \tag{3.6}$$

According to (3.6), 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. □

We now consider the calculation of ϵ^i, where ϵ^i is defined as in Corollary 3.2.2. By the formula of ϵ^i, we need to compute $\gamma_j^{(i)}, j = 1, 2, ..., \ell$, such that (3.4) is satisfied. In fact, this can be achieved by solving the following optimization problem:

$$\min \Big\{ \sum_{j\neq i} \gamma_j^{(i)} \mid \sum_{j\neq i} \gamma_j^{(i)} (x^j - x^i) = A_c x^i,\ \gamma_j^{(i)} \geq 0. \Big\} \tag{3.7}$$

Since $x^1, x^2, ..., x^k$, and A_c are known, optimization problem (3.7) 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 (3.7). The advantage by using the current objective function in (3.7) is that this optimization problem yields the largest ϵ^i that satisfies (3.4). This is since the objective function in (3.7) is $(\epsilon^i)^{-1}$. Thus, the value of ϵ^i obtained by solving the optimization problem (3.7) is the largest possible value of ϵ^i.

An alternative is presented by the following discussion. Equation (3.3) 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 for $i = 1, 2, ..., \ell$:

$$\max \Big\{ \tau^i \mid \sum_{j=1}^{\ell} u_j^{(i)} x^j = x^i + \tau^i A_c x^i,\ \sum_{j=1}^{\ell} u_j^{(i)} = 1,\ u_j^{(i)} \geq 0 \Big\}. \tag{3.8}$$

Optimization problems (3.7) and (3.8) are equivalent problems, i.e., we claim that τ^i is equal to ϵ^i. Observing that $\sum_{j=1}^{n} \beta_j^{(i)} = 1$ for the first constraint in (3.8), we have

$$\tau^i A_c x^i = \sum_{j=1}^{\ell} u_j^{(i)} x^j - \sum_{j=1}^{\ell} u_j^{(i)} x^i = \sum_{j=1}^{\ell} \tau^i \frac{u_j^{(i)}}{\tau^i} x^j - \sum_{j=1}^{\ell} \tau^i \frac{u_j^{(i)}}{\tau^i} x^i = \tau^i \sum_{j=1}^{\ell} \frac{u_j^{(i)}}{\tau^i} (x^j - x^i), \tag{3.9}$$

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 (3.7).

According to the argument for ϵ^i above, we have the following theorem.

Theorem 3.2.3. *Assume that the polytope $\mathcal{P}$ defined as in (3.2) is an invariant set for the continuous system (1.2), and the forward Euler method is applied to (1.2). Then, $\tau = \min_{i=1,2,\ldots,\ell}\{\epsilon^i\}$, where ϵ^i is defined as in Corollary 3.2.2, 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 3.2.2 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. □

3.2.2 Taylor Approximation Type Discretization Methods

We now consider the Taylor approximation type discretization methods. Note that the solution of the continuous system (1.2) 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:

$$e^{A_c \Delta t} \approx I + A_c \Delta t + \frac{1}{2!} A_c^2 \Delta t^2 + \cdots + \frac{1}{p!} A_c^p \Delta t^p = \sum_{i=0}^{p} \frac{1}{i!} A_c^i \Delta t^i := A_d. \tag{3.10}$$

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 (3.10). In fact, the Taylor approximation type methods form a family of discretization methods. For example, $p = 1$ corresponds to the forward Euler method, $p = 2$ corresponds to Heun's method, the midpoint method, or generalized Runge-Kutta 2nd order methods, $p = 3$ corresponds to the classical 3rd order Runge-Kutta method, $p = 4$ corresponds to classical 4th order Runge-Kutta method, etc, see, e.g., [32].

3.2.2.1 Existence of Steplength Threshold

Our approach to derive steplength threshold is based on the invariance conditions presented in Theorem 2.2.8 and Theorem 2.2.4. The basic idea is that we build the relationship between these two invariance conditions of the continuous and discrete systems. In fact,

conditions in Theorem 2.2.8 and Theorem 2.2.4 are essentially linear feasibility problems [60]. The unknowns in the two invariance conditions are the matrix H and $\tilde{H}$ given by Theorem 2.2.8 and Theorem 2.2.4, respectively. Thus, the key is to find relationship between those matrices.

Lemma 3.2.4. *[36] Assume H satisfies the condition in Theorem 2.2.8, 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 γ in Lemma 3.2.4 is not unique, e.g., any value greater than a valid γ is also valid. We will show more about the effect of γ to the steplength threshold in Section 3.2.3, and the way to derive a larger steplength threshold based on γ is also presented.

Lemma 3.2.5. *Assume H satisfies the condition in Theorem 2.2.8, and define*

$$\tilde{H}(\Delta t) = I + H\Delta t + \frac{1}{2!}H^2\Delta t^2 + \cdots + \frac{1}{p!}H^p\Delta t^p = \sum_{i=0}^{p}\frac{1}{i!}H^i\Delta t^i. \tag{3.11}$$

a). For the γ and $\hat{H}$ given in Lemma 3.2.4, we have

$$\tilde{H}(\Delta t) = f_0(\Delta t)I + f_1(\Delta t)\hat{H} + ... + f_p(\Delta t)\hat{H}^p, \tag{3.12}$$

where

$$f_i(\Delta t) = \sum_{k=i}^{p}\frac{(-1)^{k-i}}{k!}\binom{k}{i}\gamma^{k-i}\Delta t^k,\ \textit{for } i = 0, 1, ..., p, \tag{3.13}$$

and

$$\sum_{i=0}^{p}\gamma^i f_i(\Delta t) = 1. \tag{3.14}$$

b). Let $\tau = \min_{i=0,...,p}\{\tau_i\}$, where τ_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 the condition in Theorem 2.2.4, where A_d is defined by (3.10).

Proof. a). According to Lemma 3.2.4, there exists $\gamma > 0$, such that $\hat{H} = H + \gamma I \geq 0$. The matrix $\tilde{H}(\Delta t)$ given by (3.11) is represented in terms of Δt. By substituting $H = \hat{H} - \gamma I$

into (3.11), we now reformulate $\tilde{H}(\Delta t)$ in terms of $\hat{H}$, i.e.,

$$\begin{aligned}\tilde{H}(\Delta t) = I + (\hat{H} - \gamma I)\Delta t + \frac{1}{2!}(\hat{H}^2 - 2\gamma\hat{H} + \gamma^2 I)\Delta t^2 + \cdots \\ + \frac{1}{p!}(\hat{H}^p - p\gamma\hat{H}^{p-1} + \cdots + (-1)^p\gamma^p I)\Delta t^p.\end{aligned} \tag{3.15}$$

According to (3.15), 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 (3.13).

We note that $\sum_{i=0}^{p}\gamma^i f_i(\Delta t)$ is equivalent to replacing I and $\hat{H}$ in (3.12) by 1 and γ, respectively. Then, according to (3.15), we have

$$\sum_{i=0}^{p}\gamma^i f_i(\Delta t) = \sum_{i=0}^{p}\frac{1}{i!}(\gamma\Delta t)^i\sum_{k=0}^{i}(-1)^k\binom{i}{k}. \tag{3.16}$$

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 (3.16) equals to 1, thus (3.14) follows immediately.

b). We note that for every i the first term of $f_i(\Delta t)$ given as in (3.13) is $\frac{1}{i!}\Delta t^i$. Then we can write

$$f_i(\Delta t) = \frac{\Delta t^i}{i!}\big(1 + \mathcal{O}(\Delta t)\big). \tag{3.17}$$

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

$$\tau = \min_{i=0,1,...,p}\{\tau_i\}, \tag{3.18}$$

thus we have $f_i(\Delta t) \geq 0$ for all $\Delta t \in [0, \tau]$ and $i = 0, 1, ..., p$. According to (3.12), 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 τ is defined by (3.18). Thus, we have proved that the first condition in Theorem 2.2.4 is satisfied.

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

$$H^iG = H^{i-1}(HG) = H^{i-1}GA_c = H^{i-2}(HG)A_c = H^{i-2}GA_c^2 = ... = GA_c^i. \tag{3.19}$$

Then, according to (3.19), and substituting (3.11) and (3.10), we have

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

Thus, we have proved that the second condition in Theorem 2.2.4 is satisfied.

Since H satisfies the condition in Theorem 2.2.8, 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

$$\Big(\frac{\hat{H}}{\gamma}\Big)^i b \leq b, \text{ i.e., } \hat{H}^i b \leq \gamma^i b, \text{ for any } i = 1, 2, ..., p. \tag{3.20}$$

Then, according to (3.20) and (3.14), 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 Theorem 2.2.4 is satisfied. The proof is complete. □

Lemma 3.2.5 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 (3.11), which is derived from the Taylor approximation (3.10). We note that Kraaijevanger [46] uses a simpler approach to study the polynomial approximation to exponential function related to numerical methods for solving initial value problems, which is motivated by positivity and contractivity problems. In fact, one can show that the functions given as in (3.13) are coming from Taylor expansion of an exponential function. One may refer to [46] for more detailed discussions. According to Lemma 3.2.5, Theorem 2.2.8 and Theorem 2.2.4, we have the following theorem.

Theorem 3.2.6. *Assume a polyhedron $\mathcal{P}$, given as in (3.1), is an invariant set for the continuous system (1.2), and a Taylor approximation type discretization method (3.10) is applied to the continuous system (1.2). Then, the steplength threshold $\tau > 0$ as given in Lemma 3.2.5 is a valid steplength threshold for invariance preserving for the given Taylor approximation type discretization method (3.10) on $\mathcal{P}$.*

According to the proof of Lemma 3.2.5, we have that a valid τ requires $f_i(\Delta t) \geq 0$ for all $\Delta t \in [0, \tau]$ and all $i = 0, 1, ..., p$, where $f_i(\Delta t)$ is given by (3.13). Since each $f_i(\Delta t)$ can be represented in the form of (3.17), the following corollary is immediate.

Corollary 3.2.7. *The value of τ given in Theorem 4.3.9 (or Lemma 3.2.5) is a valid steplength threshold for invariance preserving on $\mathcal{P}$ for the Taylor approximation type discretization methods (3.10). To compute τ, one needs to find the first positive zeros of finitely many polynomial functions in the form*

$$f(\Delta t) = 1 + \alpha_1 \Delta t + \alpha_2 \Delta t^2 + ... + \alpha_q \Delta t^q, \quad \alpha_q \neq 0, \tag{3.21}$$

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

In fact, Lemma 3.2.5 can be extended to a more general case for polynomial approximation rather than Taylor type discretization methods.

Theorem 3.2.8. *Assume H satisfies the condition in Theorem 2.2.8, and define*

$$\tilde{H}(\Delta t) = I + \sigma_1 H \Delta t + \sigma_2 H^2 \Delta t^2 + \cdots + \sigma_p H^p \Delta t^p = \sum_{i=0}^{p} \sigma_i H^i \Delta t^i. \tag{3.22}$$

a). For the γ and $\hat{H}$ given in Lemma 3.2.4, we have

$$\tilde{H}(\Delta t) = f_0(\Delta t) I + f_1(\Delta t) \hat{H} + ... + f_p(\Delta t) \hat{H}^p, \tag{3.23}$$

where

$$f_i(\Delta t) = \sum_{k=i}^{p} (-1)^{k-i} \sigma_k \binom{k}{i} \gamma^{k-i} \Delta t^k, \text{ for } i = 0, 1, ..., p, \tag{3.24}$$

and

$$\sum_{i=0}^{p} \gamma^i f_i(\Delta t) = 1. \tag{3.25}$$

b). Let $\tau = \min_{i=0,\ldots,p}\{\tau_i\}$, *where* τ_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 the condition in Theorem 2.2.4, where* A_d *is defined by (3.10).*

Proof. a). According to Lemma 3.2.4, there exists a $\gamma > 0$, such that $\hat{H} = H + \gamma I \geq 0$. The matrix $\tilde{H}(\Delta t)$ given by (3.22) is represented in terms of Δt. By substituting $H = \hat{H} - \gamma I$ into (3.22), we now reformulate $\tilde{H}(\Delta t)$ in terms of $\hat{H}$, i.e.,

$$\begin{aligned} \tilde{H}(\Delta t) = I + \sigma_1(\hat{H} - \gamma I)\Delta t + \sigma_2(\hat{H}^2 - 2\gamma\hat{H} + \gamma^2 I)\Delta t^2 + \cdots \\ + \sigma_p(\hat{H}^p - p\gamma\hat{H}^{p-1} + \cdots + (-1)^p\gamma^p I)\Delta t^p. \end{aligned} \tag{3.26}$$

According to (3.26), 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 (3.24).

We note that $\sum_{i=0}^{p} \gamma^i f_i(\Delta t)$ is equivalent to replacing I and $\hat{H}$ in (3.23) by 1 and γ, respectively. Then, according to (3.26), we have

$$\sum_{i=0}^{p} \gamma^i f_i(\Delta t) = \sum_{i=0}^{p} \alpha_i(\gamma\Delta t)^i \sum_{k=0}^{i}(-1)^k\binom{i}{k}. \tag{3.27}$$

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 (3.27) equals to 1, thus (3.25) follows immediately.

The proof for Part b) is the same as the one presented for Part b) in Lemma 3.2.5, thus we are not repeating the proof here. □

3.2.2.2 Computing Steplength Threshold

We now consider the value of τ, i.e., the steplength threshold. In this section, we present an algorithm to numerically compute τ. In particular, this algorithm aims to find the first

positive zero of a polynomial function in the form of (3.21).

Lemma 3.2.9. *Let $f(\Delta t)$ be given as in (3.21). 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 (3.21). 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 3.2.9. 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 τ^* that satisfies Lemma 3.2.9 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 3.2.10. *Let $f(\Delta t)$ be given as in (3.21) 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 that 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 3.2.10 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 by (3.21).

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 3.2.11. *[70] Let $f(t)$ be a univariate polynomial function. The* ***Sturm sequence*** *$\{s_i(t)\}, i = 1, 2, ...,$ of $f(t)$ is defined as*

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

where $s'(t)$ is the derivative of $s(t)$ with respect to t, and $s_i(t)$ is the negative of the remainder of the division when $s_{i-2}(t)$ is divided 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 3.2.12. *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 ν_1 to ν_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 3.2.13. *[70]* ***(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 3.2.10 and Theorem 3.2.13, we now propose our algorithm to numerically find the first positive zero of $f(\Delta t)$ where $f(\Delta t)$ is defined by (3.21). Let us denote $\#f[\delta]$ the number of positive zeros of $f(\Delta t)$ at the interval $[0, \delta]$. The value of $\#f[\delta]$ can be computed by the Sturm Theorem 3.2.13. 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 factor 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 3.2.10.

Step 1: **[Initial Setting]** Set $t_l = t^\circ$, $t_r = t^*$, and $\epsilon > 0$ 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 3.2.10. 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 τ, such that $f(t) \geq 0$ for all $[0, \tau]$, thus we return t_l, i.e., the left end of the interval.

Remark 3.2.14. *If all coefficients $\sigma_i \geq 0$ for $i = 1, 2, ..., p$ in (3.22), then the algorithm is also applicable to compute a valid steplength threshold for invariance preserving for the polynomial approximation (3.22).*

3.2.3 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 (3.10). 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 when applied to the continuous system, yield the discrete system

$$x_{k+1} = r(A_c\Delta t)x_k, \tag{3.28}$$

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

$$r(t) = \frac{g(t)}{h(t)} = \frac{\lambda_0 + \lambda_1 t + \cdots + \lambda_p t^p}{\mu_0 + \mu_1 t + \cdots + \mu_q t^q}, \tag{3.29}$$

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 [29], etc.

Definition 3.2.15. *[33] Let $r(t)$ be given as in (3.29), and let M be a matrix. Assume $h(M)$ is nonsingular, then*[1]

$$r(M) := (h(M))^{-1}g(M) = g(M)(h(M))^{-1}. \tag{3.30}$$

3.2.3.1 Existence of Steplength Threshold

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

Definition 3.2.16. *[64] Let $r(t) : \mathbb{R} \to \mathbb{R}$. If $\rho = \max\{\kappa \,|\, r^{(i)}(t) \geq 0$ for all $i = 1, 2, ...,$ and $t \in [-\kappa, 0]\}$, where $r^{(i)}(t)$ is the i^{th} derivative of $r(t)$, then ρ is called the **radius of absolute monotonicity** of $r(t)$.*

[1]The definition of a matrix function may refer to Chapter 1 in [33]. By using the definition of matrix functions, and the fact that both $h(x)$ and $g(x)$ are polynomial functions, one can easily verify that $g(M)h(M) = h(M)g(M)$. Multiplying this identity from left and right by the inverse of $h(M)$ gives the commuting relationship in (3.30).

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., [38, 46, 64].

Theorem 3.2.17. *Assume $r(t)$ is a rational function with $r(0) = 1$. Let ρ be the radius of absolute monotonicity of $r(t)$. Assume $\rho > 0$, and assume a polyhedron $\mathcal{P}$ be given as in (3.1) is an invariant set for the continuous system (1.2), and the rational function type discretization method given as in (3.28) is applied to the continuous system (1.2). Then $\tau = \frac{\rho}{\gamma}$, where γ is given in Lemma 3.2.4, is a valid steplength threshold for invariance preserving of the rational function type discretization method given as in (3.28) on $\mathcal{P}$.*

Proof. The framework of this proof is similar to the one presented for Lemma 3.2.5. Since $\mathcal{P}$ is an invariant set for the continuous system, according to Lemma 3.2.4, there exists an H, and $\gamma > 0$, such that

$$H + \gamma I \geq 0,\ HG = GA_c,\ \text{and}\ Hb \leq 0. \tag{3.31}$$

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

$$\tilde{H}(\Delta t) \geq 0,\ \tilde{H}(\Delta t)G = Gr(A_c\Delta t),\ \text{and}\ \tilde{H}(\Delta t)b \leq b. \tag{3.32}$$

Let $\tilde{H}(\Delta t) = r(H\Delta t)$. Now we prove that $\tilde{H}(\Delta t)$ satisfies (3.32).

For the first condition in (3.32), we use the Taylor expansion of $r(t)$ at the value $-\rho$ as

$$r(t) = \sum_{i=0}^{\infty} \frac{r^{(i)}(-\rho)}{i!}(t+\rho)^i. \tag{3.33}$$

By substituting $t = H\Delta t$ into (3.33) we have

$$\tilde{H}(\Delta t) = r(H\Delta t) = \sum_{i=0}^{\infty} \frac{r^{(i)}(-\rho)}{i!}(H\Delta t + \rho I)^i = \sum_{i=0}^{\infty} \frac{r^{(i)}(-\rho)}{i!}(\Delta t)^i \left(H + \frac{\rho}{\Delta t} I\right)^i. \tag{3.34}$$

Since ρ is the radius of absolute monotonicity of $r(t)$, we have $\frac{r^{(i)}(-\rho)}{i!} \geq 0$ for all i. Also,

according to (3.31), 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 (3.34), we have $\tilde{H}(\Delta t) \geq 0$ for $\Delta t \leq \frac{\rho}{\gamma}$, thus the first condition in (3.32) is satisfied.

According to Definition 3.2.15, the second condition in (3.32) 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.,

$$g(H\Delta t) G h(A_c \Delta t) = h(H\Delta t) G g(A_c \Delta t). \tag{3.35}$$

According to (3.29), we have

$$\begin{aligned} h(H\Delta t) G g(A_c \Delta t) &= \sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_i \mu_j H^i G H^j \Delta t^{i+j}, \\ g(H\Delta t) G h(A_c \Delta t) &= \sum_{j=1}^{q} \sum_{i=1}^{p} \lambda_i \mu_j H^j G H^i \Delta t^{i+j}. \end{aligned} \tag{3.36}$$

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

$$H^i G A_c^j = G A_c^{i+j} = H^{i+j} G = H^j G A_c^i. \tag{3.37}$$

According to (3.36) and (3.37), we have that (3.35) is true, i.e., the second condition (3.32) is satisfied.

For the third condition in (3.32) 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 (3.32) is also satisfied. The proof is complete. □

The assumption $r(0) = 1$ in Theorem 3.2.17 is a fundamental condition for most dis-

cretization methods. This is since for steplength $\Delta t = 0$ the coefficient matrix of the discrete system is the identity matrix.

3.2.3.2 Computing Steplength Threshold

The steplength threshold given in Theorem 3.2.17 is related to ρ and γ. Recall that γ is given in Lemma 3.2.4, thus we only consider the computation of ρ.

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 ρ 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 ρ is the negative of the first negative real zero of this polynomial function. Then an algorithm similar to the one presented in Section 3.2.2.2 can be proposed to numerically compute ρ. Due to the space consideration, we are not repeating the algorithm here.

3.2.4 Parameter and Steplength Threshold

According to Theorem 4.3.9 and Theorem 3.2.17, we have that the parameter γ plays an important role to derive a large valid steplength threshold. In this section, we consider the effect of γ to the steplength threshold.

3.2.4.1 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 (3.13) can be written as

$$f_i(\Delta t) = \Delta t^i \sum_{k=i}^{p} \frac{(-1)^{k-i}}{k!} \binom{k}{i} (\gamma \Delta t)^{k-i}, \text{ for } i = 0, 1, ..., p, \tag{3.38}$$

which means that smaller γ will yield larger steplength threshold for Taylor type discretization method given as in (3.10). Similarly, according to Theorem 3.2.17, we also have that smaller γ will yield larger steplength threshold for the rational function type discretization methods (3.28). Thus we prefer the smallest possible γ, which in fact can be computed by

solving the following optimization problem

$$\min\{\gamma \,|\, H + \gamma I \geq 0,\ HG = GA_c,\ \text{and}\ Hb \leq 0\}. \tag{3.39}$$

In optimization problem (3.39), the variables are H and γ, while G, A_c and b are known, thus problem (3.39) is a linear optimization problem, which can be easily solved by existing optimization algorithms, e.g., simplex methods [6] or interior point methods [60]. If IPMs are used, then (3.39) can be solved in polynomial time. In particular, if there exists an $H \geq 0$ such that $HG = GA_c$ and $Hb \leq 0$, then the optimal solution, denoted by γ^*, of (3.39) is nonpositive. In this case, according to (3.38), we have $f_i(\Delta t) \geq 0$ for all $\Delta t \geq 0$. Then according to the proof of Lemma 3.2.5, we have that the steplength threshold for invariance preserving for Taylor approximation type discretization methods (3.10) on polyhedron $\mathcal{P}$ is infinity. Similarly, if $\gamma^* \leq 0$, according to Theorem 3.2.17, we have that the steplength threshold for invariance preserving for rational function type discretization methods (3.28) on polyhedron $\mathcal{P}$ is also infinity. Thus, we have the following theorem.

Theorem 3.2.18. *If the optimal solution of (3.39) is nonpositive, then the steplength threshold for invariance preserving on the polyhedron $\mathcal{P}$ is infinity for both Taylor approximation type discretization methods (3.10) and rational function type discretization methods (3.28) (For rational functional type discretization method, we assume the radius of absolute monotonicity is positive).*

One should note that the steplength thresholds given in Theorem 4.3.9 and Theorem 3.2.17 may not be the largest steplength thresholds. For example, for 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, Δt^*, 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 Δt^* 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 Δt^* is a local minimum. Then we have to make Δt to be larger, and use an algorithm similar to the one presented in Section 3.2.2.2 to find the next zero of $f(\Delta t)$, and verify again if that is a local minimum. This procedure used to be repeated until the first valid positive zero is found.

3.2.4.2 Relation to the Forward Euler Method

The following lemma presents the relationship between γ that satisfies the constraints in (3.39) 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 3.2.19. *Assume $\gamma > 0$. 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

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

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 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, completes the proof. □

We highlight that the forward Euler method is used to analyze invariance in continuous dynamical systems in [14, 15]. In [14], 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 [14]. The value of γ^{-1} in Lemma 3.2.19 can be considered as the step size of the forward Euler method for preserving the invariance of polyhedral set $\mathcal{P}$, and the value of γ is easily quantified. The existence of a step size for preserving the contractivity of a set is also presented in [14] for the forward Euler method. A similar result to Lemma 3.2.19 is presented in [15], which is an extension of [23], 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 [15] for (A, B)-invariance condition is computed in a similar way to the one given as in Lemma 3.2.19.

Chapter 4

Theory of Invariance Preserving

In this chapter, we consider the existence of local and uniform invariance preserving steplength thresholds on a set when a discretization method is applied to a linear or nonlinear dynamical system. For the forward or backward Euler method, the existence of local and uniform invariance preserving steplength thresholds is proved when the invariant sets are polyhedra, ellipsoids, or Lorenz cones. Further, we also quantify the steplength thresholds of the backward Euler methods on these sets for linear dynamical systems. Finally, we present our main results on the existence of uniform invariance preserving steplength threshold of general discretization methods on general convex sets, compact sets, and proper cones both for linear and nonlinear dynamical systems.

4.1 Introduction

In this chapter, we consider discrete and continuous linear dynamical systems which are respectively given as in (1.1) and (1.2). For simplicity, we use A to indicate the matrix A_c in the continuous system (1.2) in this chapter.

The following definition introduces the concepts of invariance preserving and steplength threshold.

Definition 4.1.1. *Assume a set $\mathcal{S}$ is an invariant set for the continuous system (1.2), and a discretization method is applied to the continuous system to yield a discrete system.*

- *For a given $x_k \in \mathcal{S}$, if there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \mathcal{S}$ for $\Delta t \in [0, \tau(x_k)]$,*

where x_{k+1} is obtained by using the discretization method, then the discretization method is ***locally invariance preserving at*** *x_k, and $\tau(x_k)$ is a* ***local invariance preserving steplength threshold*** *for this discretization method* ***at*** *x_k.*

- *If there exists a $\tau > 0$, such that $\mathcal{S}$ is also an invariant set for the discrete system for any steplength $\Delta t \in [0, \tau]$, then the discretization method is* ***uniformly invariance preserving on*** *$\mathcal{S}$ and τ is a* ***uniform invariance preserving steplength threshold*** *for this discretization method* ***on*** *$\mathcal{S}$.*

The forward and backward Euler methods are simple first order discretization methods that are usually applied to solve ordinary differential equations numerically with initial conditions. The forward Euler method, which is an explicit method, is conditionally stable. On the other hand, the backward Euler method, which is an implicit method, is unconditionally stable, see, e.g., [32].

4.2 Local Steplength Threshold

In this section, we prove the existence of an invariance preserving local steplength threshold when the invariant sets are polyhedra, ellipsoids, and Lorenz cones.

4.2.1 Existence of Local Steplength Threshold

We first consider polyhedral sets and the forward and backward Euler methods for linear systems.

Lemma 4.2.1. *Assume that a polyhedron $\mathcal{P}$, given as in (7.1), is an invariant set for the continuous system (1.2), and $x_k \in \mathcal{P}$. Then there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \mathcal{P}$ for all $\Delta t \in [0, \tau(x_k)]$, where x_{k+1} is obtained by the forward Euler method.*

Proof. Since $\text{int}(\mathcal{P})$ is an open set, we have that the statement is true for $x_k \in \text{int}(\mathcal{P})$. For any $x_k \in \partial\mathcal{P}$, we have $x_{k+1} = (I + A\Delta t)x_k = x_k + \Delta t A x_k$. According to Nagumo's Theorem 7.2.5, see e.g., [52], we have $Ax_k \in \mathcal{T}_{\mathcal{P}}(x_k)$. Then the statement is also true for $x_k \in \partial\mathcal{P}$. □

In fact, the proof of Lemma 4.2.1 is also applicable for nonlinear systems, thus a similar

conclusion about the local steplength threshold can be obtained for nonlinear systems too. Now we turn our attention to the backward Euler method.

Lemma 4.2.2. *Assume that a polyhedron $\mathcal{P}$, given as in (7.1), is an invariant set for the continuous system (1.2), and $x_k \in \mathcal{P}$. Then there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \mathcal{P}$ for all $\Delta t \in [0, \tau(x_k)]$, where x_{k+1} is obtained by the backward Euler method.*

Proof. Since $\mathcal{P}$ is an invariant set for the continuous system (1.2), we have $G(e^{At}x) \leq b$ for all $t \geq 0$. By substituting $e^{At} = I + At + \frac{1}{2}A^2t^2 + \cdots$, we have $Gx + tGAx + \frac{t^2}{2!}GA^2x + \cdots \leq b$ for all $t \geq 0$, which, for all $t \geq 0$, can be written as

$$G_i^T x + tG_i^T Ax + \tfrac{t^2}{2!}G_i^T A^2 x + \tfrac{t^3}{3!}G_i^T A^3 x + \cdots \leq b_i, \text{ for } i \in \mathcal{I}(n). \tag{4.1}$$

For the backward Euler method we need to prove that for given $Gx_k \leq b$ there exists a $\tau(x_k) > 0$, such that $G(I - A\Delta t)^{-1}x_k \leq b$, for $\Delta t \in [0, \tau(x_k)]$, which, by using $(I - A\Delta t)^{-1} = I + A\Delta t + A^2\Delta t + \cdots$, is equivalent to

$$G_i^T x_k + \Delta t G_i^T Ax_k + (\Delta t)^2 G_i^T A^2 x_k + (\Delta t)^3 G_i^T A^3 x_k + \cdots \leq b_i, \text{ for } i \in \mathcal{I}(n), \tag{4.2}$$

for $\Delta t \in [0, \tau(x_k)]$. For $i \in \mathcal{I}(n)$, we denote the bound for Δt by $\tau_i(x_k) \geq 0$, such that (4.2) holds. We have the following three cases:

- If $G_i^T x_k < b_i$, then $\tau_i(x_k) > 0$ due to the fact that $\text{int}(\mathcal{P})$ is an open set.
- If there exists an $\ell \geq 1$, such that $G_i^T x_k = b, G_i^T Ax_k = 0, ..., G_i^T A^{\ell-1}x = 0$, and $G_i^T A^\ell x_k < 0$, then according to (4.2), we have $\tau_i(x_k) > 0$.
- If neither of the above two cases is true, then we have $G_i^T x_k = b, G_i^T A^j x_k = 0$ for all $j = 1, 2, ...$, which yields $\tau_i(x_k) = \infty$.

Let $\tau(x_k) = \min_{i \in \mathcal{I}(n)}\{\tau_i(x_k)\}$. Since $\mathcal{I}(n) = n$ is finite, we have $\tau(x_k) > 0$. Clearly, when $\Delta t \in [0, \tau(x_k)]$, we have $x_{k+1} = (I - A\Delta t)^{-1}x_k \in \mathcal{P}$. The proof is complete. □

We now consider ellipsoids and Lorenz cones. If the trajectory of the continuous system is on the boundary of a given ellipsoid or Lorenz cone, then according to the fact that the

forward Euler method yields the tangent line of the trajectory at the given point x_k, we have that the forward Euler method is not invariance preserving for any $\Delta t > 0$. Thus, we only consider the backward Euler method for ellipsoids and Lorenz cones.

Lemma 4.2.3. *Assume that an ellipsoid $\mathcal{E}$, given as in (7.5), is an invariant set for the continuous system (1.2), and $x_k \in \mathcal{E}$. Then there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \text{int}(\mathcal{E})$ for all $\Delta t \in [0, \tau(x_k)]$, where x_{k+1} is obtained by the backward Euler method.*

Proof. It is easy to show that $Ax_k = 0$ implies $x_{k+1} = x_k$, thus we consider the case of $Ax_k \neq 0$. Since $\text{int}(\mathcal{E})$ is an open set, it is trivial to find $\tau(x_k) > 0$ for $x_k \in \text{int}(\mathcal{E})$. Thus we consider only the case when $x_k \in \partial\mathcal{E}$, i.e., $x_k^T Q x_k = 1$.

Since $\mathcal{E}$ is an invariant set for the continuous system, we have $x_k^T (e^{At})^T Q (e^{At}) x_k \leq 1$ for all $t \geq 0$. By substituting $e^{At} = I + At + \frac{1}{2}A^2t^2 + \mathcal{O}(t^3)$, we have

$$x_k^T Q x_k + t x_k^T (A^T Q + QA) x_k + t^2 (\tfrac{1}{2} x_k^T (A^{2T} + A^2) x_k + (Ax_k)^T Q (Ax_k)) + \mathcal{O}(t^3) \leq 1$$

for all $t \geq 0$, which, by noting that $x_k^T Q x_k = 1$, is equivalent to

$$x_k^T (A^T Q + QA) x_k + t (\tfrac{1}{2} x_k^T (A^{2T} Q + QA^2) x_k + (Ax_k)^T Q (Ax_k)) + \mathcal{O}(t^2) \leq 0 \tag{4.3}$$

for all $t \geq 0$. If $x_k^T (A^T Q + QA) x_k = 0$, then (4.3) implies

$$\tfrac{1}{2} x_k^T (A^{2T} Q + QA^2) x_k + (Ax_k)^T Q (Ax_k) \leq 0. \tag{4.4}$$

Since $Ax_k \neq 0$ and $Q > 0$, then $(Ax_k)^T Q (Ax_k) \succ 0$, which, according to (4.4), yields

$$\tfrac{1}{2} x_k^T (A^{2T} Q + QA^2) x_k < 0. \tag{4.5}$$

For the discrete system obtained by the backward Euler method, by using $(I - A\Delta t)^{-1} = I + A\Delta t + A^2 \Delta t + \mathcal{O}((\Delta t)^3)$, we have

$$\begin{aligned} x_{k+1}^T Q x_{k+1} =& 1 + \Delta t x_k^T (A^T Q + QA) x_k \\ &+ (\Delta t)^2 (x_k^T (A^{2T} + A^2) x_k + (Ax_k)^T Q (Ax_k)) + \mathcal{O}((\Delta t)^3). \end{aligned} \tag{4.6}$$

Then we consider the following two cases:

- If $x_k^T(A^TQ + QA)x_k < 0$, then $x_{k+1}^T Q x_{k+1} < 1$ for sufficiently small Δt.
- If $x_k^T(A^TQ + QA)x_k = 0$, then, according to (4.5), we have

$$x_k^T(A^{2T}Q + QA^2)x_k + (Ax_k)^T Q(Ax_k) \leq \tfrac{1}{2}x_k^T(A^{2T}Q + QA^2)x_k < 0,$$

  i.e., the coefficient of $(\Delta t)^2$ in (4.6) is negative, which yields $x_{k+1}^T Q x_{k+1} < 1$ for sufficiently small Δt.

Thus, there exists a $\tau(x_k) > 0$ such that $x_{k+1} \in$ int $(\mathcal{E})$ for all $\Delta t \in [0, \tau(x_k)]$. The proof is complete. □

Now we are ready to extend the result of Lemma 4.2.3 to the case of Lorenz cones.

Lemma 4.2.4. *Assume that a Lorenz cone $\mathcal{C}_\mathcal{L}$, given as in (7.6), is an invariant set for the continuous system (1.2), and $x_k \in \mathcal{C}_\mathcal{L}$. Then there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \mathcal{C}_\mathcal{L}$ for all $\Delta t \in [0, \tau(x_k)]$, where x_{k+1} is obtained by the backward Euler method.*

Proof. Since $x_k = 0$ implies $x_{k+1} = 0$, we consider only the case of $x_k \neq 0$. The idea of the proof is similar to the proof of Lemma 4.2.3. Since inequality (4.3) also holds for $\mathcal{C}_\mathcal{L}$, we have $x_k^T(A^TQ + QA)x_k \leq 0$. If $x_k^T(A^TQ + QA)x_k = 0$, then $(Ax_k)^T Q x_k = 0$, i.e., the inner product of Ax_k and Qx_k is 0. This shows that Ax_k is in the tangent plane of $\mathcal{C}_\mathcal{L}$ at x_k, since Qx_k is the normal direction at x_k with respect to $\mathcal{C}_\mathcal{L}$. The intersection of the tangent plane and the cone is a half line, thus we consider the following two cases:

- If $Ax_k \in \partial\mathcal{C}_\mathcal{L}$, i.e., $(Ax_k)^T Q(Ax_k) = 0$, then Ax_k is in the intersection of the cone $\mathcal{C}_\mathcal{L}$ and the tangent plane of cone $\mathcal{C}_\mathcal{L}$ at x_k. Also, since this intersection is a half line, we have $Ax_k = \lambda_k x_k$ for some $\lambda_k > 0$, i.e., the vector x_k is an eigenvector of A. Thus, we have $x_{k+1} = x_k + \lambda_k \Delta t x_k + (\lambda_k \Delta t)^2 x_k + \cdots = \frac{x_k}{1-\lambda_k \Delta t}$, for $\Delta t \in [0, \lambda_k^{-1})$, which implies that $x_{k+1} \in \partial\mathcal{C}_\mathcal{L}$ for all $\Delta t \in [0, \lambda_k^{-1})$.
- If $Ax_k \notin \partial\mathcal{C}_\mathcal{L}$, i.e., $(Ax_k)^T Q(Ax_k) > 0$, then the rest of the proof is analogous to the proof of Lemma 4.2.3, which leads to the conclusion that $x_{k+1} \in \text{int}(\mathcal{C}_\mathcal{L})$.

Thus, there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \mathcal{C}_\mathcal{L}$ for all $\Delta t \in [0, \tau(x_k)]$. The proof is complete. □

4.2.2 Computation of Local Steplength Threshold

Lemma 4.2.3 and 4.2.4 show the existence of a valid steplength threshold such that x_{k+1} obtained by the backward Euler method is also in the invariant set. In fact, given $x_k \in \mathcal{E}$ (or $\mathcal{C}_\mathcal{L}$), we can quantify the steplength threshould.

For simplicity we consider only the case of $\mathcal{E}$. To ensure $x_{k+1} \in \mathcal{E}$, we need

$$x_{k+1}^T Q x_{k+1} = x_k^T Q x_k + \Delta t x_k^T (A^T Q + QA) x_k + \cdots \leq 1. \tag{4.7}$$

We introduce the following notations to represent the sum of the remaining infinitely many terms starting from the first, second, and third term in (4.7), respectively.

$$\begin{aligned}
\sigma_1 &= \Delta t x_k^T (A^T Q + QA) x_k + \sigma_2, \\
\sigma_2 &= (\Delta t)^2 x_k^T (A^{2T} Q + A^T Q A + Q A^2) x_k + \sigma_3, \\
\sigma_3 &= (\Delta t)^3 x_k^T (A^{3T} Q + A^{2T} Q A + A^T Q A^2 + Q A^3) x_k + \mathcal{O}((\Delta t)^4).
\end{aligned}$$

Now we use the fact that $\|M + N\| \leq \|M\| + \|N\|$ and $\|MN\| \leq \|M\|\|N\|$, where M and N are matrices of appropriate dimensions. For simplicity we denote $\|A\|\Delta t$ by $\tilde{\alpha}$. We can bound σ_1 as

$$|\sigma_1| \leq \|Q\|\|x_k\|^2 \left(2\tilde{\alpha} + 3\tilde{\alpha}^2 + \cdots\right) = \|Q\|\|x_k\|^2 \tfrac{2\tilde{\alpha} - \tilde{\alpha}^2}{(1-\tilde{\alpha})^2}, \tag{4.8}$$

where (4.8) holds when $\tilde{\alpha} \leq 1$, i.e., $\Delta t \leq \frac{1}{\|A\|}$. Similarly, for σ_2 and σ_3, we have

$$|\sigma_2| \leq \|Q\|\|x_k\|^2 \tfrac{3\tilde{\alpha}^2 - 2\tilde{\alpha}^3}{(1-\tilde{\alpha})^2}, \text{ and } |\sigma_3| \leq \|Q\|\|x_k\|^2 \tfrac{4\tilde{\alpha}^3 - 3\tilde{\alpha}^4}{(1-\tilde{\alpha})^2}, \tag{4.9}$$

where $\Delta t \leq \frac{1}{\|A\|}$. We now consider the following three cases.

1). If $x_k^T Q x_k := \delta_1 < 1$, i.e., $x_k \in \text{int}(\mathcal{E})$, then to ensure that (4.7) holds, we let

$|\sigma_1| \leq 1 - \delta_1$, which is true when

$$\frac{2\tilde{\alpha}-\tilde{\alpha}^2}{(1-\tilde{\alpha})^2} \leq \frac{1-\delta_1}{\|Q\|\|x_k\|^2}, \text{ i.e., } \Delta t \leq \frac{1}{\|A\|}(1 - \frac{1}{\sqrt{1+\beta_1}}) := \gamma_1, \tag{4.10}$$

where $\beta_1 = (1 - \delta_1)\|Q\|^{-1}\|x_k\|^{-2} > 0$.

2). If $x_k^T Q x_k = 1$, i.e., $x_k \in \partial\mathcal{E}$, and $x_k^T(A^T Q + QA)x_k := -\delta_2 < 0$, then to ensure that (4.7) holds, we let $|\sigma_2| \leq \delta_2 \Delta t$, which is true when

$$\frac{3\tilde{\alpha}^2-2\tilde{\alpha}^3}{(1-\tilde{\alpha})^2} \leq \frac{\delta_2\|A\|\Delta t}{\|A\|\|Q\|\|x_k\|^2}, \text{ i.e., } \Delta t \leq \frac{1}{\|A\|}(\frac{2\beta_2+3-\sqrt{4\beta_2+9}}{2\beta_2+4}) := \gamma_2, \tag{4.11}$$

where $\beta_2 = \delta_2\|A\|^{-1}\|Q\|^{-1}\|x_k\|^{-2} > 0$.

3). If neither of the previous two cases hold, then according to (4.2.1) we have $x_k^T(A^{2T}Q + A^T QA + QA^2)x_k := -\delta_3 < 0$. Then to ensure that (4.7) holds, we let $|\sigma_3| \leq \delta_3(\Delta t)^2$, which is true when

$$\frac{4\tilde{\alpha}^3-3\tilde{\alpha}^4}{(1-\tilde{\alpha})^2} \leq \frac{\delta_3(\|A\|\Delta t)^2}{\|A\|^2\|Q\|\|x_k\|^2}, \text{ i.e., } \Delta t \leq \frac{1}{\|A\|}(\frac{\beta_3+2-\sqrt{\beta_3+4}}{\beta_3+3}) := \gamma_3, \tag{4.12}$$

where $\beta_3 = \delta_3\|A\|^{-2}\|Q\|^{-1}\|x_k\|^{-2} > 0$.

Clearly, we have $\gamma_1, \gamma_2, \gamma_3 \in (0, \frac{1}{\|A\|})$, which is consistent with conditions (4.8) and (4.9). The analysis for a cone $\mathcal{C}_\mathcal{L}$ can be done analogously. The results are summarized in the following lemma.

Lemma 4.2.5. *Assume that an ellipsoid $\mathcal{E}$, given as in (7.5) (or a Lorenz cone $\mathcal{C}_\mathcal{L}$, given as in (7.6)), is an invariant set for the continuous system (1.2), and $x_k \in \mathcal{E}$ (or $x_k \in \mathcal{C}_\mathcal{L}$). Then $x_{k+1} \in int(\mathcal{E})$ (or $int(\mathcal{C}_\mathcal{L})$), where x_{k+1} is obtained by the backward Euler method with*

- $\Delta t \in [0, \gamma_1)$, *if* $x_k \in \text{int}(\mathcal{E})$ *(or* $\text{int}(\mathcal{C}_\mathcal{L})$*),*
- $\Delta t \in [0, \gamma_2)$, *if* $x_k \in \partial\mathcal{E}$ *(or* $\partial\mathcal{C}_\mathcal{L}$*) and* $(Ax_k)^T Q x_k < 0$,
- $\Delta t \in [0, \gamma_3)$, *if* $x_k \in \partial\mathcal{E}$ *(or* $\partial\mathcal{C}_\mathcal{L}$*) and* $(Ax_k)^T Q x_k = 0$,

where γ_1, γ_2 and γ_3 are defined as in (4.10), (4.11), and (4.12), respectively.

Note that γ_1, γ_2, or γ_3 might be quite small. Let us consider an ellipsoid as an example. If x_k is sufficiently close to the boundary, then we have $x_k^T Q x_k := \delta_1 \approx 1$, which yields that $\gamma_1 \approx 0$.

We now present two simple examples, in which the forward Euler method is not invariance preserving, while the backward Euler method is invariance preserving.

Example 4.2.6. *Consider the ellipsoid* $\mathcal{E} = \{(\xi, \eta) \mid \xi^2 + \eta^2 \leq 1\}$ *and the system* $\dot{\xi} = -\eta, \dot{\eta} = \xi$.

The solution of this system is $\xi(t) = \alpha \cos t + \beta \sin t$ and $\eta(t) = \alpha \sin t - \beta \cos t$, where α, β are two parameters that depend on the initial condition. The solution trajectory is a circle, thus $\mathcal{E}$ is an invariant set for the system. If we apply the forward Euler method, the discrete system is $\xi_{k+1} = \xi_k - \Delta t \eta_k$, $\eta_{k+1} = \Delta t \xi_k + \eta_k$. Thus, we obtain $\xi_{k+1}^2 + \eta_{k+1}^2 = (1 + (\Delta t)^2)(\xi_k^2 + \eta_k^2) > \xi_k^2 + \eta_k^2$, which yields $(\xi_{k+1}, \eta_{k+1}) \notin \mathcal{E}$ for every $\Delta t > 0$ when $(\xi_k, \eta_k) \in \partial\mathcal{E}$. If we apply the backward Euler method, the discrete system is $\xi_{k+1} = \frac{1}{1+(\Delta t)^2}(\xi_k - \Delta t \eta_k)$, $\eta_{k+1} = \frac{1}{1+(\Delta t)^2}(\Delta t \xi_k + \eta_k)$. Thus we obtain that $\xi_{k+1}^2 + \eta_{k+1}^2 = \frac{1}{1+(\Delta t)^2}(\xi_k^2 + \eta_k^2) \leq \xi_k^2 + \eta_k^2$, which yields $(\xi_{k+1}, \eta_{k+1}) \in \mathcal{E}$ for every $\Delta t \geq 0$, when $(\xi_k, \eta_k) \in \mathcal{E}$.

Example 4.2.7. *Consider the Lorenz cone* $\mathcal{C}_\mathcal{L} = \{(\xi, \eta, \zeta) \mid \xi^2 + \eta^2 \leq \zeta^2, \zeta \geq 0\}$ *and the system* $\dot{\xi} = \xi - \eta, \dot{\eta} = \xi + \eta, \dot{\zeta} = \zeta$.

The solution of the system is $\xi(t) = e^t(\alpha \cos t + \beta \sin t)$, $\eta(t) = e^t(\alpha \sin t - \beta \cos t)$ and $\zeta(t) = \gamma e^t$, where α, β, γ are three parameters depending on the initial condition. It is easy to show that $\mathcal{C}_\mathcal{L}$ is an invariant set for the system. If we apply the forward Euler method, the discrete system is

$$\begin{pmatrix} \xi_{k+1} \\ \eta_{k+1} \\ \zeta_{k+1} \end{pmatrix} = \begin{pmatrix} 1 + \Delta t & -\Delta t & 0 \\ \Delta t & 1 + \Delta t & 0 \\ 0 & 0 & 1 + \Delta t \end{pmatrix} \begin{pmatrix} \xi_k \\ \eta_k \\ \zeta_k \end{pmatrix}.$$

However, if we choose any $(\xi_k, \eta_k, \zeta_k) \in \partial\mathcal{C}_\mathcal{L}$, then we have $(\xi_{k+1}, \eta_{k+1}, \zeta_{k+1}) \notin \mathcal{C}_\mathcal{L}$, since $\xi_{k+1}^2 + \eta_{k+1}^2 = (1 + (1 + \Delta t)^2)(\xi_k^2 + \eta_k^2) > (1 + \Delta t)^2(\xi_k^2 + \eta_k^2) = \zeta_{k+1}^2$, for all $\Delta t > 0$. If we

apply the backward Euler method, the discrete system is

$$\begin{pmatrix} \xi_{k+1} \\ \eta_{k+1} \\ \zeta_{k+1} \end{pmatrix} = \frac{1}{\omega} \begin{pmatrix} (1-\Delta t)^2 & -\Delta t(1-\Delta t) & 0 \\ \Delta t(1-\Delta t) & (1-\Delta t)^2 & 0 \\ 0 & 0 & (1-\Delta t)^2 + \Delta t^2 \end{pmatrix} \begin{pmatrix} \xi_k \\ \eta_k \\ \zeta_k \end{pmatrix},$$

where $\omega = (1-\Delta t)((1-\Delta t)^2 + \Delta t^2)$. If we choose any point $(\xi_k, \eta_k, \zeta_k) \in \mathcal{C}_{\mathcal{L}}$, then we have $(\xi_{k+1}, \eta_{k+1}, \zeta_{k+1}) \in \mathcal{C}_{\mathcal{L}}$, since $\xi_{k+1}^2 + \eta_{k+1}^2 = \frac{1}{\omega^2}((1-\Delta t)^4 + \Delta t^2(1-\Delta t)^2)(\xi_k^2 + \eta_k^2) \leq \frac{1}{\omega^2}((1-\Delta t)^2 + \Delta t^2)^2(\xi_k^2 + \eta_k^2) \leq \frac{1}{\omega^2}((1-\Delta t)^2 + \Delta t^2)^2 z_k^2 = \zeta_{k+1}^2$ for all $\Delta t > 0$, and $\zeta_{k+1} = \frac{1}{\omega}((1-\Delta t)^2 + \Delta t^2)\zeta_k \geq 0$ for all $\Delta t \in [0, 1)$.

4.3 Uniform Steplength Threshold

In the analysis of Section 4.2, the invariance preserving steplength threshold depends on the given x_k. However, such a changing steplength threshold is not practical, i.e., with such changing threshold one has to sequentially modify the value of invariance preserving Δt as x_k is changing. Thus, it is important to obtain a uniform steplength threshold for invariance preserving that depends only on the given invariant set.

4.3.1 Uniform Steplength Threshold for Linear Systems

We first consider polyhedral sets and the forward Euler method. Note that a similar results for polytopes is presented in [11]. For the forward Euler method, Theorem 3.2.3 already presents the existence of uniform steplength threshold threshold. Thus, we now consider the polyhedron and the backward Euler method. Note that a similar result can be found in [39].

Theorem 4.3.1. *Assume that a polyhedron $\mathcal{P}$, given as in (3.1), is an invariant set for the continuous system (1.2). Then there exists a $\hat{\tau} > 0$, such that for every $x_k \in \mathcal{P}$ and $\Delta t \in [0, \hat{\tau}]$, we have $x_{k+1} \in \mathcal{P}$, where x_{k+1} is obtained by the backward Euler method, i.e., $\mathcal{P}$ is an invariant set for the discrete system.*

Proof. Let $\bar{\tau} = \sup\{\tau \,|\, I - A\Delta t \text{ is nonsingular for every } \Delta t \in [0, \tau]\}$. Recall that, see, e.g., [58] p. 44, the relative interior and the relative boundary of a set $\mathcal{S}$ are denoted by $\text{ri}(\mathcal{S})$

and $\mathrm{rb}(\mathcal{S})$, respectively. Note that $\mathcal{P}$ is a closed set, thus for every $x_k \in \mathcal{P}$, one has either $x_k \in \mathrm{ri}(\mathcal{P})$ or $x_k \in \mathrm{rb}(\mathcal{P})$. We consider the following two cases:

Case 1). $x_k \in \mathrm{ri}(\mathcal{P})$. For every $\tau > 0$, we can reformulate $x_{k+1} = (I - A\Delta t)^{-1}x_k$ as $x_{k+1} + \frac{\Delta t}{\tau}x_{k+1} = x_k + \frac{\Delta t}{\tau}(x_{k+1} + \tau A x_{k+1})$, i.e.,

$$x_{k+1} = \tfrac{\tau}{\tau+\Delta t}x_k + \tfrac{\Delta t}{\tau+\Delta t}(x_{k+1} + \tau A x_{k+1}). \tag{4.13}$$

Note that $\frac{\tau}{\tau+\Delta t} + \frac{\Delta t}{\tau+\Delta t} = 1$, $\frac{\tau}{\tau+\Delta t} > 0$, and $\frac{\Delta t}{\tau+\Delta t} > 0$, thus x_{k+1} is a convex combination of x_k and $\bar{x} = x_{k+1} + \tau A x_{k+1}$. Further, we observe that $\bar{x}$ is the vector obtained by applying the forward Euler method at x_{k+1} with steplength τ.

Now we are going to prove that $x_{k+1} \in \mathrm{ri}(\mathcal{P})$ for every $\Delta t \in [0, \bar{\tau})$. This proof is by contradiction. Let us assume that there exists a $\bar{\tau}_1 \in [0, \bar{\tau})$, such that $x_{k+1} = (I - A\bar{\tau}_1)x_k \in \mathrm{rb}(\mathcal{P})$. We now choose a $\tau > 0$, which is not larger than the threshold given in Theorem **??**, thus we have $\bar{x} \in \mathcal{P}$ and

$$x_{k+1} = \tfrac{\tau}{\tau+\bar{\tau}_1}x_k + \tfrac{\bar{\tau}_1}{\tau+\bar{\tau}_1}\bar{x}, \tag{4.14}$$

which, by noting that $x_k \in \mathrm{ri}(\mathcal{P})$, implies that $x_{k+1} \in \mathrm{ri}(\mathcal{P})$. This contradicts to the assumption that $x_{k+1} \in \mathrm{rb}(\mathcal{P})$.

Case 2). $x_k \in \mathrm{rb}(\mathcal{P})$. There exists a $y \in \mathrm{ri}(\mathcal{P})$, such that $\bar{x}_k^\epsilon = x_k + \epsilon y \in \mathrm{ri}(\mathcal{P})$, for every $\epsilon \in (0, 1)$. By a similar discussion as in Case 1), we have that $\bar{x}_{k+1}^\epsilon = (I - A\Delta t)\bar{x}_k^\epsilon \in \mathrm{ri}(\mathcal{P})$, for every $\Delta t \in [0, \bar{\tau})$. By letting $\epsilon \to 0$, we have that $\bar{x}_{k+1}^\epsilon \to x_{k+1} \in \mathcal{P}$.

We prove that every $\hat{\tau} \in (0, \bar{\tau})$ satisfies the theorem. The proof is complete. □

Remark 4.3.2. *The proof of Theorem* 4.3.1 *also quantifies the value of the invariance preserving uniform steplength threshold $\hat{\tau}$, i.e., $\hat{\tau} \in (0, \bar{\tau})$, where $\bar{\tau} = \sup\{\tau \,|\, I - A\Delta t$ is nonsingular for every $\Delta t \in [0, \tau]\}$.*

Corollary 4.3.3. *Assume that a polyhedral cone $\mathcal{C}_\mathcal{P}$ is an invariant set for the continuous system (1.2). Then there exists a $\hat{\tau} > 0$, such that for every $x_k \in \mathcal{C}_\mathcal{P}$ and $\Delta t \in [0, \hat{\tau}]$, we have $x_{k+1} \in \mathcal{C}_\mathcal{P}$, where x_{k+1} is obtained by the backward Euler method, i.e., $\mathcal{C}_\mathcal{P}$ is an invariant set for the discrete system.*

Theorem 4.3.4. *Assume that an ellipsoid $\mathcal{E}$, given as in (7.5), is an invariant set for the continuous system (1.2). Then there exists a $\hat{\tau} > 0$, such that for every $x_k \in \mathcal{E}$ and $\Delta t \in [0, \hat{\tau}]$, we have $x_{k+1} \in \mathcal{E}$, where x_{k+1} is obtained by the backward Euler method, i.e., $\mathcal{E}$ is an invariant set for the discrete system.*

Proof. In the backward Euler method, the coefficient matrix is $(I - A\Delta t)^{-1}$, where Δt is the steplength. Given any $x_k \in \mathcal{E}$, according to Lemma 4.2.3, there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \text{int}(\mathcal{E})$ for every $\Delta t \in (0, \tau(x_k)]$. In our proof, we need to bound the magnitude of the coefficient matrix $(I - A\Delta t)^{-1}$. We consider the eigenvalues of $(I - A\Delta t)^{-1}$, which are $(1 - \lambda_i(A)\Delta t)^{-1}$, for $i = 1, 2, ..., n$. To bound $(1 - \lambda_i(A)\Delta t)^{-1}$, we need $|\lambda_i(A)\Delta t| < 1$. Note that any positive $\tau < \tau(x_k)$ is also a bound for Δt, thus, for example, we can choose $0 < \Delta t \leq \hat{\tau}(x_k) := \min\{\tau(x_k), \frac{1}{2|\lambda_i(A)|}\} = \min\{\tau(x_k), \frac{1}{2\rho(A)}\}$, where $\rho(A)$ is the spectral radius (see, e.g. [36]) of A, which yields $|1 - \lambda_i(A)\Delta t| \geq \frac{1}{2}$. Thus, we need to have that $\|(I - A\Delta t)^{-1}\|$ is uniformly bounded by 2 on $\mathcal{E}$ for every $\Delta t \in (0, \hat{\tau}(x_k)]$.

Since $x_{k+1} = (I - A\hat{\tau}(x_k))^{-1}x_k \in \text{int}(\mathcal{E})$, we can choose a positive $R(x_{k+1})$, such that the open ball $\delta(x_{k+1}, R(x_{k+1})) \subset \text{int}(\mathcal{E})$. It is easy to verify that the open ball $\delta(x_k, \frac{1}{2}R(x_{k+1}))$ is mapped into $\delta(x_{k+1}, R(x_{k+1}))$ by the backward Euler method. This is because for $\tilde{x}_k \in \delta(x_k, \frac{1}{2}R(x_{k+1}))$, we apply the backward Euler method at $\tilde{x}_k$ with $\hat{\tau}(x_k)$ to yield $\tilde{x}_{k+1} = (I - A\hat{\tau}(x_k))^{-1}\tilde{x}_k$. Then we have

$$\|\tilde{x}_{k+1} - x_{k+1}\| \leq \|(I - A\hat{\tau}(x_k))^{-1}\|\|\tilde{x}_k - x_k\| \leq 2\|\tilde{x}_k - x_k\| \leq R(x_{k+1}),$$

i.e., $\tilde{x}_{k+1} \in \delta(x_{k+1}, R(x_{k+1})) \subset \text{int}(\mathcal{E})$. Therefore, we have that $\hat{\tau}(x_k)$ is a uniform bound for Δt at every point in $\delta(x_k, \frac{1}{2}R(x_{k+1}))$.

Obviously, $\cup_{x_k \in \mathcal{E}}\delta(x_k, \frac{1}{2}R(x_{k+1}))$ is an open cover of the ellipsoid $\mathcal{E}$. Since $\mathcal{E}$ is a compact set, according to [61], there exists a finite subcover $\cup_{k=1}^{m}\delta(x_k, \frac{1}{2}R(x_{k+1}))$ of $\mathcal{E}$. For each open ball $\delta(x_k, \frac{1}{2}R(x_{k+1}))$, there is a uniform bound $\hat{\tau}(x_k)$, thus, we have that $\hat{\tau} = \min_{k=1,...,m}\{\hat{\tau}(x_k)\}$ is an invariance preserving uniform bound for Δt for the backward Euler method at every point in $\mathcal{E}$. The proof is complete. □

We now consider to quantify a uniform steplength threshold of the backward Euler

method for invariance preserving for ellipsoids. We need some technical results.

Lemma 4.3.5. [36] *Let $M \succ 0$ ($M \succeq 0$, $M \prec 0$, or $M \preceq 0$) and N be a nonsingular matrix. Then $N^T MN \succ 0$ ($N^T MN \succeq 0$, $N^T MN \prec 0$, or $N^T MN \preceq 0$).*

Lemma 4.3.6. *If $Q \succ 0, A^T Q + QA \preceq 0$, then*

- *for $P = QA$, we have $x^T Px \leq 0$ for every $x \in \mathbb{R}^n$.*
- *for every $t \geq 0$, $I - At$ is nonsingular.*

Proof. For $x \neq 0$, $2x^T Px = 2x^T(QA)x = x^T(A^T Q + QA)x \leq 0$, that proves the first part.

For the second part, since $I - At = I - tQ^{-1}P = Q^{-1}(Q - tP)$, the singularity of $I - At$ is equivalent to that of $Q - tP$. Assume that the latter one is singular. Then there exists an $x \neq 0$, such that $(Q - tP)x = 0$. Then $0 = x^T(Q - tP)x = x^T Qx - tx^T Px > 0$, where the last inequality is due to $Q \succ 0$ and the first part. This is a contradiction, thus the proof is complete. □

The following theorem presents a uniform invariance preserving steplength threshold of the backward Euler method for ellipsoids. The form of the threshold coincides with the one for polyhedra given in Remark 4.3.2. Further, the uniform steplength threshold is proved to be ∞.

Theorem 4.3.7. *Assume that an ellipsoid $\mathcal{E}$ is an invariant set for the continuous system (1.2). Let $\bar{\tau} = \sup\{\tau \,|\, I - A\Delta t$ is nonsingular for every $\Delta t \in [0, \tau]\}$. Then $\bar{\tau} = \infty$, and thus for every $x_k \in \mathcal{C}$ and $\Delta t \geq 0$ we have that $x_{k+1} \in \mathcal{E}$, where x_{k+1} is obtained by the backward Euler method, i.e., $\mathcal{E}$ is an invariant set for the discrete system.*

Proof. We have that $\mathcal{E}$ is an invariant set for the discrete and continuous systems if and only if $A^T QA - Q \preceq 0$ and $A^T Q + QA \preceq 0$, respectively. Then by Lemma 4.3.6, we have that $\hat{\tau} = \infty$. It is easy to see that the theorem is equivalent to that $A^T Q + QA \preceq 0$ implies that

$$(I - tA)^{-T} Q(I - tA)^{-1} - Q \preceq 0 \tag{4.15}$$

holds for every $t \geq 0$, According to Lemma 4.3.5, to prove (4.15) is equivalent to prove

$Q-(I-tA)^TQ(I-tA)\preceq 0$, i.e.,

$$A^TQ+QA-tA^TQA\preceq 0. \tag{4.16}$$

Since $Q\succ 0$, we have $A^TQA\succeq 0$, thus (4.16) is true. The proof is complete. □

Remark 4.3.8. *In Theorem 4.3.7, if we assume A is a symmetric matrix, then we have $\bar{\tau}=\frac{1}{\lambda_1(A)}$, where $\lambda_1(A)$ is the largest real eigenvalue of A.*

By using an analogous discussion as the one presented in the proof of Theorem 4.3.7, one can show that other discretization methods, e.g., Padé[1,1], Padé[2,2], etc., see e.g., [2], also allow some uniform invariance preserving steplength thresholds.

To establish a uniform invariance preserving steplength threshold for the backward Euler method for the Lorenz cone $\mathcal{C}_{\mathcal{L}}$, we first consider the case when no eigenvector of the coefficient matrix A in (1.2) is on the boundary of $\mathcal{C}_{\mathcal{L}}$.

Theorem 4.3.9. *Assume that a Lorenz cone $\mathcal{C}_{\mathcal{L}}$, given as in (7.6), is an invariant set for the continuous system (1.2), and no eigenvector of the coefficient matrix A in (1.2) is on $\partial(\mathcal{C}_{\mathcal{L}})$. Then there exists a $\hat{\tau}>0$, such that for every $x_k\in\mathcal{C}_{\mathcal{L}}$ and $\Delta t\in[0,\hat{\tau}]$, we have $x_{k+1}\in\mathcal{C}_{\mathcal{L}}$, where x_{k+1} is obtained by the backward Euler method, i.e., $\mathcal{C}_{\mathcal{L}}$ is an invariant set for the discrete system.*

Proof. If $x_k=0$, then for every $\Delta t\geq 0$ we have $x_{k+1}=0\in\mathcal{C}_{\mathcal{L}}$. We now consider the case when $x_k\neq 0$. Our proof has two steps.

The first step of the proof is considering a uniform bound for Δt on a base (see [3] page 66) of the Lorenz cone $\mathcal{C}_{\mathcal{L}}$. For every $x_k\in\mathcal{C}_{\mathcal{L}}$, to have $x_{k+1}=(I-A\Delta t)^{-1}x_k\in\mathcal{C}_{\mathcal{L}}$, we need to have

$$\begin{aligned} x_{k+1}^TQx_{k+1}=&x_k^TQx_k+\Delta t x_k^T(A^TQ+QA)x_k\\ &+(\Delta t)^2x_k^T(A^{2T}Q+A^TQA+QA^2)x_k+\cdots\leq 0. \end{aligned} \tag{4.17}$$

Let us take a hyperplane $\mathcal{H}$ such that $\mathcal{H}$ intersected with $\mathcal{C}_{\mathcal{L}}$ is a compact set $0\notin\mathcal{C}_{\mathcal{L}}^+=\mathcal{H}\cap\mathcal{C}_{\mathcal{L}}$. In fact, $\mathcal{C}_{\mathcal{L}}^+$ is a base[1] of the Lorenz cone $\mathcal{C}_{\mathcal{L}}$. For every $x_k\in\mathcal{C}_{\mathcal{L}}^+$, we consider the following

[1]In practice, a possible way to obtain a base can be done as follows: we first take a hyperplane through

four cases:

Case 1): In this case, $x_k \in \text{int}(\mathcal{C}_\mathcal{L}) \cap \mathcal{C}_\mathcal{L}^+$, thus we have $x_k^T Q x_k < 0$. Consequently, due to (4.17), $x_{k+1} \in \text{int}(\mathcal{C}_\mathcal{L})$ for sufficiently small Δt.

Case 2): In this case, $x_k \in \partial(\mathcal{C}_\mathcal{L}) \cap \mathcal{C}_\mathcal{L}^+$, and $x_k^T(A^TQ + QA)x_k < 0$, thus we have $x_k^T Q x_k = 0$. Since the constant term is zero and the first order term is negative in (4.17), we have $x_{k+1} \in \text{int}(\mathcal{C}_\mathcal{L})$ for sufficiently small Δt.

Case 3): In this case, $x_k \in \partial(\mathcal{C}_\mathcal{L}) \cap \mathcal{C}_\mathcal{L}^+$, $x_k^T(A^TQ + QA)x_k = 0$, and $Ax_k \notin \partial(\mathcal{C}_\mathcal{L}) \cup (-\partial(\mathcal{C}_\mathcal{L}))$, thus we have $x_k^T Q x_k = 0$, and $x_k^T(A^{2T}Q + A^TQA + QA^2)x_k < 0$. The last inequality is due to the proof of Lemma 4.2.4. Since the constant term is zero, the first order term is also zero, and the second order term is negative in (4.17), we have $x_{k+1} \in \text{int}(\mathcal{C}_\mathcal{L})$ for sufficiently small Δt.

Case 4): In this case, $x_k \in \partial(\mathcal{C}_\mathcal{L}) \cap \mathcal{C}_\mathcal{L}^+$, $x_k^T(A^TQ + QA)x_k = 0$, and $Ax_k \in \partial(\mathcal{C}_\mathcal{L}) \cup (-\partial(\mathcal{C}_\mathcal{L}))$. However, since x_k is nonzero, we have seen in the proof of Lemma 4.2.4 that in this case x_k is an eigenvector of A. This violates the assumption of this theorem, thus this case is not possible.

Therefore, for every $x_k \in \mathcal{C}_\mathcal{L}^+$, there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \text{int}(\mathcal{C}_\mathcal{L})$ for every $\Delta t \in (0, \tau(x_k)]$. Also, note that $\mathcal{C}_\mathcal{L}^+$ is a compact set, thus, according to a similar argument as in the proof of Theorem 4.3.4, we have a uniform bound for Δt, denoted by $\hat{\tau}(\mathcal{C}_\mathcal{L}^+)$, on $\mathcal{C}_\mathcal{L}+$, such that for every $x_k \in \mathcal{C}_\mathcal{L}^+$, we have $x_{k+1} \in \mathcal{C}_\mathcal{L}$ for every $\Delta t \in [0, \hat{\tau}(\mathcal{C}_\mathcal{L}^+)]$.

The second step of the proof is extending the uniform bound of the steplength from $\mathcal{C}_\mathcal{L}^+$ to $\mathcal{C}_\mathcal{L}$. Let $0 \neq x_k \in \mathcal{C}_\mathcal{L}$. Then, because $\mathcal{C}_\mathcal{L}^+$ is a base of $\mathcal{C}_\mathcal{L}$, there exists a scalar $\gamma > 0$, such that $\gamma x_k = \tilde{x}_k \in \mathcal{C}_\mathcal{L}^+$. Then we have

$$x_{k+1} = (I - A\Delta t)^{-1} x_k = (I - A\Delta t)^{-1} \gamma^{-1} \tilde{x}_k = \gamma^{-1} \tilde{x}_{k+1}.$$

Since $\tilde{x}_{k+1} \in \mathcal{C}_\mathcal{L}$ for every $\Delta t \in [0, \tau(\mathcal{C}_\mathcal{L}^+)]$, we have $x_{k+1} \in \mathcal{C}_\mathcal{L}$, for every $\Delta t \in [0, \tau(\mathcal{C}_\mathcal{L}^+)]$.

Therefore, $\tau(\mathcal{C}_\mathcal{L}^+)$ is a uniform bound for the steplength Δt for the backward Euler method at every point of $\mathcal{C}_\mathcal{L}$. The proof is complete. □

the origin that intersects $\mathcal{C}$ only by the origin. Then shift the hyperplane to x^*, where x^* is an interior point of $\mathcal{C}$. The intersection of the shifted hyperplane and $\mathcal{C}$ is a base of $\mathcal{C}$. The base of $\mathcal{C}$ is a compact set.

Now, in a more general setting, we consider the uniform invariance preserving steplength threshold on a general proper cone for linear dynamical systems.

Definition 4.3.10. [49] *A convex cone $\mathcal{C}$ is called **proper** if it is nonempty, closed, and pointed.*

We recall the concept of a matrix to be cross-positive on a proper cone, which is first proposed by Schneider and Vidyasagar in [62].

Definition 4.3.11. [62] *Let $\mathcal{C} \in \mathbb{R}^n$ be a proper cone and $\mathcal{C}^*$ be the dual cone of $\mathcal{C}$. The matrix $M \in \mathbb{R}^{n\times n}$ is called **cross-positive** on $\mathcal{C}$ if for all $x \in \mathcal{C}, y \in \mathcal{C}^*$ with $x^T y = 0$, the inequality $x^T My \geq 0$ holds.*

The properties of cross-positive matrices are thoroughly studied in [62]. The following lemma, which directly follows from Theorem 2 and Lemma 6 in [62], is useful in our analysis.

Lemma 4.3.12. [62] *Let $\mathcal{C} \in \mathbb{R}^n$ be a proper cone, and denote the following two sets of matrices: $\Sigma_{\mathcal{C}} = \{M \,|\, M$ is cross-positive on $\mathcal{C}\}$, and $\Pi_{\mathcal{C}} = \{M \,|\, (M + \alpha I)(\mathcal{C}\backslash\{0\}) \subseteq \text{int}(\mathcal{C})$ for some $\alpha \geq 0\}$. Then the closure of $\Pi_{\mathcal{C}}$ is $\Sigma_{\mathcal{C}}$.*

Lemma 4.3.13. *Let $\mathcal{C} \in \mathbb{R}^n$ be a proper cone, and denote $\Omega_{\mathcal{C}} = \{M \,|\, M\mathcal{C} \subseteq \mathcal{C}\}$. Then $\Omega_{\mathcal{C}}$ is closed.*

Proof. Let $\{M_i\}$ be a sequence of matrices in $\Omega_{\mathcal{C}}$, such that $\lim_{i\to\infty} M_i = M$. We choose an arbitrary $x \in \mathcal{C}$. For every i, since $M_i\mathcal{C} \subseteq \mathcal{C}$, we have $M_i x = y_i \in \mathcal{C}$. Since $\mathcal{C}$ is closed, we have $Mx = \lim_{i\to\infty} M_i x = \lim_{i\to\infty} y_i = \bar{y} \in \mathcal{C}$. The proof is complete. □

The existence of a uniform invariance preserving steplength threshold for a proper cone is presented in the following theorem.

Theorem 4.3.14. *Assume that a proper cone $\mathcal{C} \in \mathbb{R}^n$ is an invariant set for the continuous system (1.2). Then there exists a $\hat{\tau} > 0$, such that for every $x_k \in \mathcal{C}$ and $\Delta t \in [0, \hat{\tau}]$, we have $x_{k+1} \in \mathcal{C}$, where x_{k+1} is obtained by the backward Euler method, i.e., $\mathcal{C}$ is an invariant set for the discrete system.*

Proof. Since $\mathcal{C}$ is an invariant set for the continuous system, we have $e^{At}\mathcal{C} \subseteq \mathcal{C}$ for every $t \geq 0$. According to Theorem 3 in [62], this is equivalent to that the coefficient matrix A is cross-positive on $\mathcal{C}$. Then by Lemma 4.3.12, there exists a sequence of matrices $\{A_i\}$, where $(A_i + \alpha_i I)\mathcal{C} \subseteq \mathcal{C}$ for some $\alpha_i \geq 0$, such that $\lim_{i\to\infty} A_i = A$. For simplicity, we introduce the notation $B_i = A_i + \alpha_i I$, then $B_i\mathcal{C} \subseteq \mathcal{C}$.

Then we consider $(I - A\Delta t)^{-1}$, i.e., the coefficient matrix of the discrete system obtained by using the backward Euler method. Let $\bar{\tau} = \sup\{\tau \,|\, I - A\Delta t$ is nonsingular for every $\Delta t \in [0, \tau]\}$, then we have $\bar{\tau} > 0$. Since $\lim_{i\to\infty} A_i = A$ for every $0 < \epsilon_1 < \epsilon_2 < \bar{\tau}$, there exists an integer $\bar{n} > 0$, such that for every $i > \bar{n}$, we have $I - A_i\Delta t$ is nonsingular for $\Delta t \in [0, \tau_i]$, where $\tau_i \in (\bar{\tau} - \epsilon_2, \bar{\tau} - \epsilon_1)$. Since $\{\tau_i\}_{i>\bar{n}}$ is bounded, it has a convergent subsequence $\{\tau_{i^*}\}$, i.e., $\lim_{i^*\to\infty} \tau_{i^*} = \hat{\tau} \in [\bar{\tau} - \epsilon_2, \bar{\tau} - \epsilon_1]$. Thus, we have $0 < \hat{\tau} < \bar{\tau}$, and $I - A\Delta t$ is nonsingular for $\Delta t \in [0, \hat{\tau}]$. For every i^* we have

$$(I - A_{i^*}\Delta t)^{-1} = ((1 + \alpha_{i^*}\Delta t)I - B_{i^*}\Delta t)^{-1} = \tfrac{1}{1+\alpha_{i^*}\Delta t}(I - B_{i^*}\tfrac{\Delta t}{1+\alpha_{i^*}\Delta t})^{-1}. \tag{4.18}$$

Since $B_{i^*}\mathcal{C} \subseteq \mathcal{C}$ and $\frac{\Delta t}{1+\alpha_{i^*}\Delta t} > 0$, we have

$$(I - B_{i^*}\tfrac{\Delta t}{1+\alpha_{i^*}\Delta t})^{-1}\mathcal{C} = (I + \tfrac{\Delta t}{1+\alpha_{i^*}\Delta t}B_{i^*} + (\tfrac{\Delta t}{1+\alpha_{i^*}\Delta t})^2 B_{i^*}^2 + ...)\mathcal{C} \subseteq \mathcal{C}. \tag{4.19}$$

Since $1 + \alpha_{i^*}\Delta t > 0$, by (4.18) and (4.19), we have $(I - A_{i^*}\Delta t)^{-1}\mathcal{C} \subseteq \mathcal{C}$ for $\Delta t \in [0, \tau_{i^*}]$.

Finally, since $(I - A\Delta t)^{-1} = \lim_{i^*\to\infty}(I - A_{i^*}\Delta t)^{-1}$, according to Lemma 4.3.13, we have $(I - A\Delta t)^{-1}\mathcal{C} \subseteq \mathcal{C}$ for $\Delta t \in [0, \hat{\tau}]$. The proof is complete. □

In fact, according to the proof of Theorem 4.3.14, we can also give the exact value of a uniform bound for the steplength for a proper cone.

Corollary 4.3.15. *Assume that a proper cone $\mathcal{C} \in \mathbb{R}^n$ is an invariant set for the continuous system (1.2). Let $\bar{\tau} = \sup\{\tau \,|\, I - A\Delta t$ is nonsingular for every $\Delta t \in [0, \tau]\}$ and $\hat{\tau} \in (0, \bar{\tau})$. Then for every $x_k \in \mathcal{C}$ and $\Delta t \in [0, \hat{\tau})$, we have $x_{k+1} \in \mathcal{C}$, where x_{k+1} is obtained by the backward Euler method, i.e., $\mathcal{C}$ is an invariant set for the discrete system.*

Proof. For every $\Delta t \in [0, \hat{\tau})$, we choose $0 < \epsilon_1 < \epsilon_2 < \bar{\tau} - \Delta t$. Then, by an argument similar to the proof of Theorem 4.3.14, we have that $\hat{\tau} \in [\bar{\tau} - \epsilon_2, \bar{\tau} - \epsilon_1]$ is a uniform bound of the

steplength. Note that $\Delta t < \bar{\tau} - \epsilon_2 \leq \hat{\tau}$, and we can choose $\epsilon_2 > 0$ sufficiently small, then the corollary is immediate. □

Let us take an example to illustrate Corollary 4.3.15.

Example 4.3.16. *Consider the cone* $\mathcal{C} = \{(\xi, \eta) \mid \xi^2 \leq \eta^2, \eta \geq 0\}$, *and the system* $\dot{\xi} = 3\xi - \eta, \dot{\eta} = -\xi + 3\eta$.

The solution of the system is $\xi(t) = \frac{1}{2}(\alpha e^{2t} - \beta e^{4t}), \eta(t) = \frac{1}{2}(\alpha e^{2t} + \beta e^{4t})$, where α, β depend on the initial condition. Clearly, $\mathcal{C}$ is an invariant set for the system. It is easy to compute $\tau = \sup\{\, \Delta t \,|\, I - A\Delta t \text{ is nonsingular}\} = \frac{1}{4}$. When the backward Euler method is applied, we have $\xi_{k+1} = \gamma((1 - 3\Delta t)\xi_k - \Delta t \eta_k), \eta_{k+1} = \gamma(-\Delta t \xi + (1 - 3\Delta t)\eta_k)$, where $\gamma = ((1 - 3\Delta t)^2 + (\Delta t)^2)^{-1}$. To ensure that $\xi_{k+1}^2 \leq \eta_{k+1}^2$, we let $(1 - 3\Delta t)^2 - (\Delta t)^2 \geq 0$, which yields that $\Delta t \leq \frac{1}{4}$. Note that the other solution that $\Delta t \geq \frac{1}{2}$ is not applicable.

Since a Lorenz cone is a proper cone, the following corollary is immediate.

Corollary 4.3.17. *Assume that a Lorenz cone* $\mathcal{C}_\mathcal{L}$, *given as in (7.6), is an invariant set for the continuous system (1.2). Then there exists a* $\hat{\tau} > 0$ *such that for every* $x_k \in \mathcal{C}_\mathcal{L}$ *and* $\Delta t \in [0, \hat{\tau}]$ *we have* $x_{k+1} \in \mathcal{C}_\mathcal{L}$, *where* x_{k+1} *is obtained by the backward Euler method, i.e.,* $\mathcal{C}_\mathcal{L}$ *is an invariant set for the discrete system. Moreover,* $\hat{\tau} \in [0, \bar{\tau})$, *where* $\bar{\tau}$ *is given as in Corollary 4.3.15.*

4.3.2 General Results for Uniform Steplength Threshold

The property that the backward Euler method has a uniform invariance preserving steplength threshold for plays a significant role in the proof of Theorem 4.3.1, thus we now generalize the conclusion to closed and convex sets. By a similar proof of Theorem 4.3.1, the following theorem is immediate.

Theorem 4.3.18. *Let* $\mathcal{S}$ *be a closed and convex set. Assume that* $\mathcal{S}$ *is an invariant set for the continuous system (1.2), and let* $\bar{\tau} = \sup\{\tau \,|\, I - A\Delta t$ *is nonsingular for every* $\Delta t \in [0, \tau]\}$. *Assume that there exists a* $\tilde{\tau} > 0$, *such that for every* $x_k \in \mathcal{S}$ *and* $\Delta t \in [0, \tilde{\tau}]$, *we have* $x_k + \Delta t A x_k \in \mathcal{S}$. *Then for every* $x_k \in \mathcal{S}$ *and* $\Delta t \in [0, \bar{\tau})$, *we have* $x_{k+1} \in \mathcal{S}$, *where*

x_{k+1} is obtained by the backward Euler method, i.e., $\mathcal{S}$ is an invariant set for the discrete system.

The compactness of an ellipsoid plays an important role in the proof of Theorem 4.3.4. Now we generalize Theorem 4.3.4 to compact sets and a general class of discretization methods.

Theorem 4.3.19. *Let a set $\mathcal{S}$, and a discretization method $x_{k+1} = D(\Delta t, x_k)$ be given. Assume that the following conditions hold:*

1. *The set $\mathcal{S}$ is a compact set.*
2. *For every $x_k \in \mathcal{S}$, there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \text{int}(\mathcal{S})$ for every $\Delta t \in (0, \tau(x_k)]$.*
3. *The Lipschitz condition holds for $D(\Delta t, x)$ with respect to x, i.e., there exists an $L > 0$, such that*

$$\|D(\Delta t, \tilde{x}) - D(\Delta t, x)\| \leq L\|\tilde{x} - x\|, \text{ for } x, \tilde{x} \in \mathcal{S}. \tag{4.20}$$

Then there exists a $\hat{\tau} > 0$, such that for every $x_k \in \mathcal{S}$ and $\Delta t \in [0, \hat{\tau}]$, we have $x_{k+1} \in \mathcal{S}$, i.e., $\mathcal{S}$ is an invariant set for the discrete system.

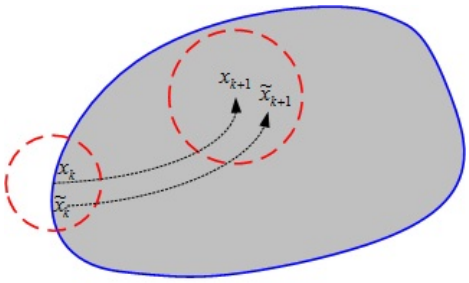

Figure 4.1: The idea of the proof of Theorem 4.3.19.

Proof. Note that every positive $\tau < \tau(x_k)$ is also a bound for Δt at x_k. Then let us define $\hat{\tau}(x_k) = \min\{\tau(x_k), \tilde{\tau}\}$, according to Condition 2, we have $x_{k+1} = D(\hat{\tau}(x_k), x_k) \in \text{int}(\mathcal{S})$. Thus we can choose an $R(x_{k+1}) > 0$, such that the open ball $\delta(x_{k+1}, R(x_{k+1})) \subset \text{int}(\mathcal{S})$. According to Condition 3, there exists $0 < L < \infty$, such that $\|D(\Delta t, x_k)\| \leq L$ for all $x_k \in \mathcal{S}$ and $\Delta t \in [0, \tilde{\tau}]$.

It is easy to verify that by the discretization method the open ball $\delta(x_k, \frac{1}{L}R(x_{k+1}))$ is mapping into $\delta(x_{k+1}, R(x_{k+1}))$, see Figure 4.1. This is because for every $\tilde{x}_k \in \delta(x_k, \frac{1}{L}R(x_{k+1}))$, the discretization method applied to $\tilde{x}_k$ with steplength $\hat{\tau}(x_k)$ yields $\tilde{x}_{k+1} = D(\hat{\tau}(x_k), \tilde{x}_k)$. Then we have

$$\|\tilde{x}_{k+1} - x_{k+1}\| \leq \|D(\hat{\tau}(\tilde{x}_k), \tilde{x}_k) - D(\hat{\tau}(x_k), x_k)\| \leq L\|\tilde{x}_k - x_k\| \leq R(x_{k+1}), \tag{4.21}$$

i.e., $\tilde{x}_{k+1} \in \delta(x_{k+1}, R(x_{k+1})) \subset \text{int}(\mathcal{S})$. Therefore, we have that $\hat{\tau}(x_k)$ is a uniform bound for Δt at every point in $\delta(x_k, \frac{1}{L}R(x_{k+1}))$.

Obviously, $\cup_{x_k \in \mathcal{S}} \delta(x_k, \frac{1}{L}R(x_{k+1}))$ is an open cover of $\mathcal{S}$. Since $\mathcal{S}$ is a compact set, there exists a finite subcover $\cup_{k=1}^{m} \delta(x_k, \frac{1}{L}R(x_{k+1}))$ of $\mathcal{S}$. A uniform bound for Δt is the smallest $\hat{\tau}(x_k)$ of the finite number of open balls $\delta(x_k, \frac{1}{L}R(x_{k+1}))$, thus, we have that $\hat{\tau} = \min_{k=1,\ldots,m}\{\hat{\tau}(x_k)\}$ is a uniform bound for Δt for the discretization method at every point in $\mathcal{S}$. The proof is complete. □

According to Theorem 4.3.19, the following corollary is immediate.

Corollary 4.3.20. *Let a set $\mathcal{S}$, and a discretization method $x_{k+1} = D(\Delta t)x_k$ be given. Assume that the following conditions hold:*

1. *The set $\mathcal{S}$ is a compact set.*
2. *For every $x_k \in \mathcal{S}$, there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \text{int}(\mathcal{S})$ for every $\Delta t \in (0, \tau(x_k)]$.*
3. *There exists a $\tilde{\tau} > 0$, such that $\|D(\Delta t)\|$ is uniformly bounded for every $x \in \mathcal{S}$ and $\Delta t \in [0, \tilde{\tau}]$.*

Then there exists a $\hat{\tau} > 0$, such that for every $x_k \in \mathcal{S}$ and $\Delta t \in [0, \hat{\tau}]$, we have $x_{k+1} \in \mathcal{S}$, i.e., $\mathcal{S}$ is an invariant set for the discrete system.

The assumption in Theorem 4.3.9 that no eigenvector of the coefficient matrix is on the boundary of $\mathcal{C}_\mathcal{L}$ excludes the case that $x_{k+1} \in \partial(\mathcal{C}_\mathcal{L})$. We now generalize Theorem 4.3.9 to proper cones.

Theorem 4.3.21. *Let a set $\mathcal{C}$, and a discretization method $x_{k+1} = D(\Delta t, x_k)$ be given. Assume that the following conditions hold:*

1. *The set $\mathcal{C}$ is a proper cone.*

2. *For every $0 \neq x_k \in \mathcal{C}$, there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \text{int}(\mathcal{C})$ for every $\Delta t \in (0, \tau(x_k)]$.*

3. *The Lipschitz condition holds for $D(\Delta t, x)$ with respect to x, i.e., there exists an $L > 0$, such that*

$$\|D(\Delta t, \tilde{x}) - D(\Delta t, x)\| \leq L\|\tilde{x} - x\|, \text{ for } x, \tilde{x} \in \mathcal{S}. \tag{4.22}$$

4. *The function $D(\Delta t, x)$ is homogeneous of degree $p \geq 1$ with respect to x, i.e.,*

$$D(\Delta t, \alpha x) = \alpha^p D(\Delta t, x), \text{ for } \alpha \in \mathbb{R}, x \in \mathcal{C}. \tag{4.23}$$

Then there exists a $\hat{\tau} > 0$, such that for every $x_k \in \mathcal{C}$ and $\Delta t \in [0, \hat{\tau}]$ we have $x_{k+1} \in \mathcal{C}$, i.e., $\mathcal{C}$ is an invariant set for the discrete system.

Proof. If $x_k = 0$ then for every $\Delta t \geq 0$ we have $x_{k+1} = 0 \in \mathcal{C}$. We now consider the case when $x_k \neq 0$. Our proof has two steps.

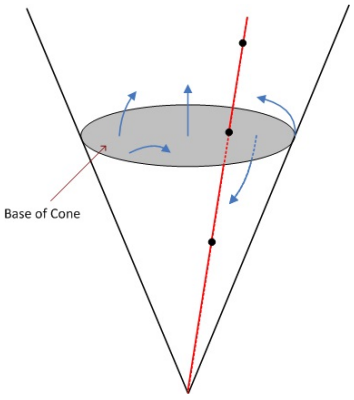

Figure 4.2: The idea of the proof of Theorem 4.3.21.

The first step of the proof is considering a uniform bound for Δt on a base of $\mathcal{C}$. Since $\mathcal{C}$ is a proper cone, we can take a hyperplane $\mathcal{H}$ to intersect it with $\mathcal{C}$ to generate a base of $\mathcal{C}$ denoted by $\mathcal{C}^+ = \mathcal{H} \cap \mathcal{C}$. For every $x_k \in \mathcal{C}^+$, there exists a $\tau(x_k) > 0$, such that

$x_{k+1} \in \text{int}(\mathcal{C})$ for every $0 < \Delta t \leq \tau(x_k)$. Note that $\mathcal{H}$ and $\mathcal{C}$ are closed sets, thus $\mathcal{C}^+$ is also a closed set. Also, assume that $\mathcal{C}^+$ is unbounded, then $\mathcal{C} \cap (-\mathcal{C})$ contains a half line, which contradicts that $\mathcal{C}$ is a pointed proper cone. Therefore, $\mathcal{C}^+$ is a compact set. Then, according to a similar argument as in the proof of Theorem 4.3.20, we have a uniform bound for Δt, denoted by $\hat{\tau}(\mathcal{C}^+)$, on $\mathcal{C}^+$, such that for every $x_k \in \mathcal{C}^+$, we have $x_{k+1} \in \mathcal{C}$ for every $\Delta t \in [0, \hat{\tau}(\mathcal{C}^+)]$, see Figure 4.2.

The second step of the proof is extending the uniform bound of the steplength Δt from $\mathcal{C}^+$ to $\mathcal{C}$. Let $0 \neq x_k \in \mathcal{C}$. Then, because $\mathcal{C}^+$ is a base of $\mathcal{C}$, there exists a scalar $\gamma > 0$ such that $\gamma x_k = \tilde{x}_k \in \mathcal{C}^+$. Then we have

$$x_{k+1} = D(\Delta t, x_k) = D(\Delta t, \gamma^{-1}\tilde{x}_k) = \gamma^{-p}\tilde{x}_{k+1}. \tag{4.24}$$

Since $\tilde{x}_{k+1} \in \mathcal{C}$ for every $\Delta t \in [0, \tau(\mathcal{C}^+)]$, we have $x_{k+1} \in \mathcal{C}$ for every $\Delta t \in [0, \tau(\mathcal{C}^+)]$.

Therefore, $\tau(\mathcal{C}^+)$ is a uniform bound for the steplength Δt for the discretization method at every point on $\mathcal{C}$. The proof is complete. □

Corollary 4.3.22. *Let a set $\mathcal{C}$, and a discretization method $x_{k+1} = D(\Delta t)x_k$ be given. Assume that the following conditions hold:*

1. *The set $\mathcal{C}$ is a proper cone.*
2. *For every $0 \neq x_k \in \mathcal{C}$, there exists a $\tau(x_k) > 0$, such that $x_{k+1} \in \text{int}(\mathcal{C})$ for every $\Delta t \in (0, \tau(x_k)]$.*
3. *There exists a $\tilde{\tau} > 0$, such that $\|D(\Delta t)\|$ is uniformly bounded for every $x \in \mathcal{S}$ and $\Delta t \in [0, \tilde{\tau}]$.*

Then there exists a $\hat{\tau} > 0$, such that for every $x_k \in \mathcal{C}$ and $\Delta t \in [0, \hat{\tau}]$, we have $x_{k+1} \in \mathcal{C}$, i.e., $\mathcal{C}$ is an invariant set for the discrete system.

Chapter 5

Invariance Conditions for Nonlinear Systems

5.1 Introduction

In this chapter, we consider invariance conditions for some classical convex sets for discrete and continuous nonlinear systems. This chapter is a generalization of Chapter 2. Discrete and continuous systems in general form are respectively described by the following equations:

$$x_{k+1} = f_d(x_k), \tag{5.1}$$

$$\dot{x}(t) = f_c(x(t)), \tag{5.2}$$

where $f_d, f_c : \mathbb{R}^n \to \mathbb{R}^n$ are continuous functions, and $x_k \in \mathbb{R}^n$ for $k = 1, 2, ...$ and $x(t) \in \mathbb{R}^n$ for $t \geq 0$ are *state variables*.

5.2 Invariance Conditions

5.2.1 Invariance Conditions for Discrete Systems

First, an invariance condition of polyhedral sets for discrete systems is presented as follows.

Theorem 5.2.1. *Let a polyhedron $\mathcal{P}$ be given as in (7.1) and the discrete system be given*

as in (5.1), respectively. Assume $b_i - G_i^T f_d(x)$ are convex functions for all $i \in \mathcal{I}(m)$. Then $\mathcal{P}$ is an invariant set for the discrete system (5.1) if and only if there exists a matrix $H \geq 0$, such that

$$HGx - Gf_d(x) \geq Hb - b, \quad \textit{for all } x \in \mathbb{R}^n. \tag{5.3}$$

Proof. We have that $\mathcal{P}$ is an invariant set for the discrete system (5.1) if and only if $\mathcal{P} \subseteq \mathcal{P}' = \{x \,|\, Gf_d(x) \leq b\}$. The latter one means that for every $i \in \mathcal{I}(m)$, the system $G_i^T f_d(x) > b_i$ and $Gx \leq b$ has no solution. Let us assume to the contrary that there exists an x^* and i^*, such that $G_{i^*}^T f_d(x^*) > b_{i^*}$ and $Gx^* \leq b$. Then we have $x^* \in \mathcal{P}$ but $x^* \notin \mathcal{P}'$, which contradicts to $\mathcal{P} \subseteq \mathcal{P}'$. Also, since $b_i - G_i^T f_d(x)$ is a convex function, then, according to the convex Farkas Lemma 7.2.3, we have that there exists a vector $H_i \geq 0$ such that

$$b_i - G_i^T f_d(x) + H_i^T(Gx - b) \geq 0, \quad \text{for all } x \in \mathbb{R}^n.$$

Writing H_i^T for all $i \in \mathcal{I}(m)$ together into a matrix H, we have $H \geq 0$ and

$$b - Gf_d(x) + H(Gx - b) \geq 0, \quad \text{for all } x \in \mathbb{R}^n,$$

which is the same as (5.3). □

In Theorem 5.2.1, we require $b_i - G_i^T f_d(x)$ are convex functions for all i. This means that $G_i^T f_d(x)$ are concave functions for all i. Clearly, when all $f_d(x)$ are affine functions, i.e., the system is a linear system, we have $G_i^T f_d(x)$ are convex functions. To verify if condition (5.3) is true, we can use the following optimization problems.

Remark 5.2.2. *Consider the following m optimization problems*

$$\max_{H_i \geq 0} \min_{x \in \mathbb{R}^n} \{H_i^T Gx - G_i^T f_d(x) - H_i^T b + b_i\} \quad i \in \mathcal{I}(m). \tag{5.4}$$

If the global optimal objective values of the m optimization problems in (5.4) are all non-negative, we can claim that condition (5.3) holds.

If the system in Theorem 5.2.1 is a linear dynamical system, one can easily to derive the invariance condition presented in Chapter 2.

Example 5.2.3. *Let the discrete system be given as* $\xi_{k+1} = -\xi_k + 2\eta_k - \xi_k^2, \eta_{k+1} = -2\xi_k - \eta_k + \eta_k^2$, *and the polyhedron be given as* $\mathcal{P} = \{(\xi, \eta) \,|\, \xi - \eta \leq -10, 2\xi - \eta \leq 10, \xi - 2\eta \leq -20\}$.

We first directly show that $\mathcal{P}$ is an invariant set for the discrete system, i.e., $(\xi_{k+1}, \eta_{k+1}) \in \mathcal{P}$ for all $(\xi_k, \eta_k) \in \mathcal{P}$. For simplicity, we only prove the first constraint, i.e., $\xi_{k+1} - \eta_{k+1} \leq -10$. In fact, we have $\xi_{k+1} - \eta_{k+1} = -\xi_k^2 - \eta_k^2 + \xi_k + 3\eta_k = -\xi_k^2 - (\eta_k - 2.5)^2 + \xi_k - 2\eta_k + 6.25 \leq \xi_k - 2\eta_k + 6.25 \leq -20 + 6.25 \leq -10$. The other two constraints can be proved in a similar manner. On the other hand, one can show that the assumption in Theorem 5.2.1 is satisfied for this example. Then we can find a suitable $H \geq 0$ such that condition (5.3) holds. One can easily verify that $H = [0, 0, 1; 0, 0, 0; 1, 0, 1]$ satisfies condition (5.3). Then according to Theorem 5.2.1, we have that $\mathcal{P}$ is an invariant set for the discrete system.

We now consider invariance condition of ellipsoids for the discrete system (5.1).

Theorem 5.2.4. *Let an ellipsoid* $\mathcal{E}$ *be given as in (7.5) and the discrete system be given as in (5.1), respectively. Assume* $(f_d(x))^T Q f_d(x)$ *is a concave function. Then* $\mathcal{E}$ *is an invariant set of the discrete system (5.1) if and only if there exists a* $\beta \geq 0$, *such that*

$$\beta x^T Q x - (f_d(x))^T Q f_d(x) \geq \beta - 1, \quad \text{for all } x \in \mathbb{R}^n. \tag{5.5}$$

Proof. The ellipsoid $\mathcal{E}$ is an invariant set for the discrete system if and only if $\mathcal{E} \subseteq \mathcal{E}'$, where $\mathcal{E}' = \{x \,|\, (f_d(x))^T Q f_d(x) \leq 1\}$. We also note that $\mathcal{E} \subseteq \mathcal{E}'$ is equivalent to $(\mathbb{R}^n \setminus \mathcal{E}') \cap \mathcal{E} = \emptyset$, i.e., the inequality system $1 - (f_d(x))^T Q f_d(x) < 0$ and $x^T Q x - 1 \leq 0$ has no solution. Since $(f_d(x))^T Q f_d(x)$ is a concave function, we have that $1 - (f_d(x))^T Q f_d(x)$ is a convex function. Note that $x^T Q x - 1$ is also a convex function, according to Theorem 7.2.3, there exists a $\beta \geq 0$, such that

$$-(f_d(x))^T Q f_d(x) + 1 + \beta(x^T Q x - 1) \geq 0, \quad \text{for all } x \in \mathbb{R}^n,$$

which is the same as (5.5). □

Remark 5.2.5. *If we choose* $x = 0$ *in condition (5.5), then we have* $\beta \leq 1 - (f_d(0))^T Q f_d(0)$, *which can be considered as an upper bound of* β.

Remark 5.2.6. *Consider the following optimization problem*

$$\max_{\beta \geq 0} \min_{x \in \mathbb{R}^n} \{\beta x^T Q x - (f_d(x))^T Q f_d(x) - \beta + 1\}. \tag{5.6}$$

If the optimal objective value of optimization problem (5.6) is nonnegative, we can claim that condition (5.5) holds.

The parameter β presented in Corollary **??** can be eliminated. In fact, one can show that $A_d^T Q A_d - \beta Q \preceq 0$ for $\beta \in [0, 1]$ and $Q \succ 0$ is equivalent to $A_d^T Q A_d - Q \preceq 0$, see Corollary 2.2.20.

Example 5.2.7. *Let the discrete system be* $\xi_{k+1} = \frac{\sqrt{\xi_k + \eta_k}}{2}, \eta_{k+1} = \frac{\sqrt{\xi_k - 3\eta_k}}{2}$, *and the ellipsoid be given as* $\mathcal{E} = \{(\xi, \eta) \,|\, \xi^2 + \eta^2 \leq 1\}$.

For any $(\xi_k, \eta_k) \in \mathcal{E}$, we have $\xi_{k+1}^2 + \eta_{k+1}^2 = \frac{\xi_k - \eta_k}{2} \leq \frac{\sqrt{2}}{2}\sqrt{\xi_k^2 + \eta_k^2} < 1$, which shows that $\mathcal{E}$ is an invariant set for the discrete system. On the other hand, let $f(x) = (f_1(x), f_2(x))^T = (\frac{\sqrt{\xi_k + \eta_k}}{2}, \frac{\sqrt{\xi_k - 3\eta_k}}{2})^T$ and $Q = [1, 0; 0, 1]$. Then we have that $f(x)^T Q f(x)$ is a concave function. If we choose $\beta = \frac{1}{4}$, then condition (5.5) yields $(\xi_k - 1)^2 + (\eta_k - 1)^2 + 1 \geq 0$ for any $(\xi_k, \eta_k) \in \mathbb{R}^2$. This also shows that $\mathcal{E}$ is an invariant set for the discrete system according to Theorem 5.2.4.

We now consider invariance conditions for more general convex sets for discrete system (5.1). Let a convex set be given:

$$\mathcal{S} = \{x \in \mathbb{R}^n \,|\, g(x) \leq 0\}, \tag{5.7}$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is a convex function. Then we have the following theorem, which gives invariance condition for the convex set (5.7) for discrete system (5.1).

Theorem 5.2.8. *Let the convex set* $\mathcal{S}$ *be given as in (5.7) and the discrete system be given as in (5.1), respectively. Assume that there exists* $x^0 \in \mathbb{R}^n$ *such that* $g(x) < 0$, *and that* $g(f_d(x))$ *is a concave function. Then* $\mathcal{S}$ *is an invariant set for the discrete system if and*

only if there exists an $\alpha \geq 0$, such that

$$\alpha g(x) - g(f_d(x)) \geq 0, \quad \textit{for all } x \in \mathbb{R}^n. \tag{5.8}$$

Moreover, if $g(x)$ and $g(f_d(x))$ are quadratic functions, then the assumption that $g(f_d(x))$ is a concave function is not required.

Proof. The major tool used in this proof is the convex Farkas Lemma, i.e., Theorem 7.2.3. Note that to ensure $\mathcal{S}$ is an invariant set for the discrete system, we need to prove $\mathcal{S} \subseteq \mathcal{S}' = \{x \,|\, g(f_d(x)) \leq 0\}$, i.e., $(\mathbb{R}^n \setminus \mathcal{S}') \cap \mathcal{S} = \emptyset$. Then the following inequality system has no solution:

$$-g(f_d(x)) < 0, \quad g(x) \leq 0.$$

According to Theorem 7.2.3, there exists an $\alpha \geq 0$, such that

$$-g(f_d(x)) + \alpha g(x) \geq 0, \quad \text{for } x \in \mathbb{R}^n,$$

which is the same as (5.8). For the case of quadratic functions, we can use a similar argument and the S-Lemma to prove the last statement. □

Remark 5.2.9. *The set $\mathcal{S}$ given as in (5.7) is represented by only a single convex function. In fact, the first statement in Theorem 5.2.8 can be easily extended to the set which is presented by several convex functions, e.g., polyhedral sets.*

The first statement in Theorem 5.2.8 requires $g(f_d(x))$ is a concave function given that $g(x)$ is a convex function. Let us consider x defined in one dimensional space as an example[1] to illustrate this case is indeed possbile. Since $f_d(x)$ is a convex function, we have $f_d''(x) \geq 0$ for all $x \in \mathbb{R}$. For simplicity, we denote $h(x) = -g(f_d(x))$. Then we have

$$h''(x) = -g''(f_d(x))(f_d(x))^2 - g'(f_d(x))f_d''(x). \tag{5.9}$$

If $h''(x) > 0$ for all $x \in \mathbb{R}$, then $h(x)$ is a convex function, i.e., $g(f_d(x))$ is a concave function.

[1]The example uses the following theorem: if $\tilde{g}(x)$ is a nondecreasing function, and $\tilde{f}(x)$ is a convex function, then $\tilde{g}(\tilde{f}(x))$ is a convex function.

We now find a sufficient condition such that $h'(x) > 0$ for all $x \in \mathbb{R}$. Assume that $g(x)$ is a decreasing convex nonlinear function and $g(x)$ has no lower bound, we have $g'(x) < 0$ and $g''(x) > 0$ for all $x \in \mathbb{R}$. Assume $f_d(x)$ is a concave function, we have $f_d''(x) < 0$. This yields $-\frac{g'(f_d(x))}{g_d''(f(x))} > 0 \geq \frac{(f'(x))^2}{f''(x)}$, i.e., $h''(x) > 0$.

Remark 5.2.10. *Consider the following optimization problem:*

$$\max_{\alpha \geq 0} \min_{x \in \mathbb{R}^n} \{\alpha g(x) - g(f_d(x))\}. \tag{5.10}$$

If the optimal objective value of optimization problem (5.10) is nonnegative, we can claim that condition (5.8) holds.

We can prove that optimization problem (5.10) can be transformed to a nonlinear optimization problem. From here, we assume that $g(x)$ in (5.7) is continuously differentiable. We illustrate this idea below:

Theorem 5.2.11. *Optimization problem (5.10) is equivalent to the nonlinear optimization problem*

$$\max_{x,\alpha} \{\alpha g(x) - g(f_d(x)) \,|\, \alpha \nabla_x g(x) - \nabla_x g(f_d(x)) = 0, \alpha \geq 0\}. \tag{5.11}$$

Proof. Since $\alpha \geq 0$, and the functions $g(x)$ and $-g(f_d(x))$ are both convex functions, we have that $\alpha g(x) - g(f_d(x))$ is also a convex function. Also, for $\alpha \geq 0$, the optimization problem

$$\min_{x \in \mathbb{R}^n} \{\alpha g(x) - g(f_d(x))\}, \tag{5.12}$$

is a convex optimization problem in $\mathbb{R}^n$, thus problem (5.12) has a Wolfe dual, see, e.g., [20, 78] given as follows:

$$\max_{x \in \mathbb{R}^n} \{\alpha g(x) - g(f_d(x)) \,|\, \alpha \nabla_x g(x) - \nabla_x g(f_d(x)) = 0\}. \tag{5.13}$$

Consequently, problem (5.10) is equivalent to the nonlinear optimization problem (5.11). □

Remark 5.2.12. *One can use a proof similar to the one presented in Theorem 5.2.11 to derive equivalent nonlinear optimization problems for the optimization problems (5.4) and*

(5.6).

5.2.2 Invariance Conditions for Continuous Systems

In this section, we consider invariance conditions for continuous systems in the form of (5.2).

First, we consider an invariance condition for polyhedral sets given as in (7.1) for continuous system (5.2). For simplicity, we assume that the origin is in the polyhedral set, thus we have $\mathcal{P} = \{x \in \mathbb{R}^n \,|\, Gx \leq b\} = \{x \in \mathbb{R}^n \,|\, g_i^T x \leq b_i, i = 1, 2, ..., m\}$, where $b > 0$.

Theorem 5.2.13. *Let a polyhedral set be given as $\mathcal{P} = \{x \in \mathbb{R}^n \,|\, g_i^T x \leq b_i, i = 1, 2, ..., m\}$, where $b > 0$, and let $\mathcal{P}^i = \{x \in \mathbb{R}^n \,|\, g_i^T x = b_i$ and $x \in \mathcal{P}\}$ for $i = 1, 2, ..., m$. Then $\mathcal{P}$ is an invariant set for the continuous system (5.2) if and only if*

$$g_i^T f_c(x) \leq 0, \text{ for all } x \in \mathcal{P}^i \tag{5.14}$$

holds for all $i = 1, 2, ..., m$.

Proof. Let $x \in \partial\mathcal{P}$. Then we have that x is in the relative interior of a face, on the relative boundary, or a vertex of $\mathcal{P}$. There exists an index set $\mathcal{I}_x$ such that $x \in \cap_{i \in \mathcal{I}_x} \mathcal{P}^i$. We note that $\mathcal{T}_{\mathcal{P}}(x) = \{y \in \mathbb{R}^n \,|\, g_i^T y \leq 0, i \in \mathcal{I}_x\}$, then, according to Nagumo Theorem 7.2.5, the theorem is immediate. □

Remark 5.2.14. *Let us assume a polyhedral set $\mathcal{P}$ be given as in the statement of Theorem 5.2.13. Consider the following m optimization problems:*

$$\max\{g_i^T f_c(x) \,|\, g_i^T x = b_i \text{ and } x \in \mathcal{P}\}, i = 1, 2, ..., m. \tag{5.15}$$

If the optimal objective values of the all m optimization problems (5.15) are nonpositive, then we can claim that (5.14) holds.

Invariance conditions for ellipsoids or Lorenz cones for continuous system (5.2) is presented in the following theorem.

Theorem 5.2.15. *An ellipsoid $\mathcal{E}$ given as in (7.5) (or a Lorenz cone $\mathcal{C}_{\mathcal{L}}$ given as in (7.6)) is an invariant set for the continuous system (5.2) if and only if*

$$(f_c(x))^T Qx \leq 0, \text{ for all } x \in \partial\mathcal{E} \text{ (or } x \in \partial\mathcal{C}_{\mathcal{L}}). \tag{5.16}$$

Proof. For simplicity, we only consider $\mathcal{E}$. The proof for $\mathcal{C}_{\mathcal{L}}$ is analogous. Note that $\partial\mathcal{E} = \{x \,|\, x^T Qx = 1\}$, thus the outer normal vector of $\mathcal{E}$ at $x \in \partial\mathcal{E}$ is $f_d(x)$. Then we have that the tangent cone is given as $\mathcal{T}_{\mathcal{E}}(x) = \{y \,|\, y^T Qx \leq 0\}$, thus this theorem follows by the Nagumo Theorem 7.2.5. □

Remark 5.2.16. *Let us consider an ellipsoid $\mathcal{E}$ and the following optimization problem:*

$$\max\{(f_c(x))^T Qx \,|\, x^T Qx = 1\}. \tag{5.17}$$

If the global optimal objective value of optimization problem (5.17) is nonpositive, then we can claim that condition (5.16) holds.

Theorem 5.2.17. *Let the convex set $\mathcal{S}$ be given as in (5.7) and function $g(x)$ is continuously differentiable. Then $\mathcal{S}$ is an invariant set for the continuous system (5.2) if and only if*

$$(\nabla g(x))^T f_c(x) \leq 0, \qquad \text{for all } x \in \partial\mathcal{S}. \tag{5.18}$$

Proof. The outer normal vector at $x \in \partial\mathcal{S}$ is $\nabla g(x)$. Since $\mathcal{S}$ is a convex set, we have

$$\mathcal{T}_{\mathcal{S}}(x) = \{y \,|\, (\nabla g(x))^T y \leq 0\}. \tag{5.19}$$

The proof is immediate by applying Nagumo's Theorem 7.2.5. □

Remark 5.2.18. *Consider the following optimization problem:*

$$\max\{\alpha \,|\, \alpha = (\nabla g(x))^T f_c(x), \ g(x) = 0\}. \tag{5.20}$$

If the optimal objective value of optimization problem (5.20) is nonpositive, then we can claim that condition (5.18) holds.

We note that the nonlinear optimization problems presented in (5.17) and (5.20) can be hard to solve, since the optimization problems could be not convex optimization problems.

5.2.3 General Results

In this section, invariance conditions without assumptions are presented. First, we present the invariance condition for polyhedral sets for discrete systems.

Theorem 5.2.19. *Let the polyhedron $\mathcal{P}$ be given as in (7.1) and the discrete system be given as in (5.1). Then $\mathcal{P}$ is an invariant set for the discrete system (5.1) if and only if there exists a matrix $H \geq 0$, such that*

$$HGx - Gf_d(x) \geq Hb - b, \quad \textit{for all } x \in \mathcal{P}. \tag{5.21}$$

Proof. Sufficiency: Condition (5.21) can be reformulated as $b - Gf_d(x) \geq H(b - Gx)$, where $x \in \mathcal{P}$, i.e., $b - Gx \geq 0$. Since $H \geq 0$, we have $b - Gf_d(x) \geq 0$, i.e., $f_d(x) \in \mathcal{P}$ for all $x \in \mathcal{P}$. Thus $\mathcal{P}$ is an invariant set for the discrete system. *Necessity:* Assume $\mathcal{P}$ is an invariant set for the discrete system, then we have that $b - Gx \geq 0$ implies $b - Gf_d(x) \geq 0$. Thus, we can choose $H = 0$. □

We now present the invariance condition for ellipsoidal sets for discrete systems.

Theorem 5.2.20. *Let the ellipsoid $\mathcal{E}$ be given as in (7.5) and the discrete system be given as in (5.1), respectively. Then $\mathcal{E}$ is an invariant set for the discrete system if and only if there exists a $\beta \geq 0$, such that*

$$\beta x^T Q x - (f_d(x))^T Q f_d(x) \geq \beta - 1, \quad \textit{for all } x \in \mathcal{E}. \tag{5.22}$$

Proof. Sufficiency: Condition (5.22) can be reformulated as $1 - (f_d(x))^T Q f_d(x) \geq \beta(1 - x^T Q x)$, where $x \in \mathcal{E}$. Thus we have $1 - (f_d(x))^T Q f_d(x) \geq 0$, i.e., $f_d(x) \in \mathcal{E}$. Thus $\mathcal{E}$ is an invariant set for the discrete system. *Necessity:* It is immediate by choosing $\beta = 0$. □

We now present the invariance condition for convex sets for discrete systems.

Theorem 5.2.21. *Let the convex set $\mathcal{S}$ be given as in (5.7) and the discrete system be given as in (5.1), respectively. Then $\mathcal{S}$ is an invariant set for the discrete system if and only if there exists an $\alpha \geq 0$, such that*

$$\alpha g(x) - g(f_d(x)) \geq 0, \qquad \textit{for all } x \in \mathcal{S}. \tag{5.23}$$

Proof. Sufficiency: Condition (5.23) can be reformulated as $\alpha g(x) \geq g(f_d(x))$, where $x \in \mathcal{S}$, i.e., $g(x) \leq 0$. According to $\alpha \geq 0$, we have $g(f_d(x)) \leq 0$, i.e., $f_d(x) \in \mathcal{S}$. Thus $\mathcal{S}$ is an invariant set for the discrete system. *Necessity:* It is immediate by choosing $\beta = 0$. □

We note that there is no convexity assumption for the involved in Theorem 5.2.19, Theorem 5.2.20, and Theorem 5.2.21, we cannot use Wolfe duality to derive the invariance conditions. The absence of convexity assumptions makes the theorems stronger, however the nonlinear feasibility problems (5.21), (5.22), and (5.23) are nonconvex, thus their verification is significantly harder than solving convex feasibility problems and requires harder optimization problems to be solved.

Chapter 6

Conclusions and Future Research

In this chapter, we conclude this thesis and propose future research topics.

6.1 Conclusions

Invariant sets are important both in the theory and computational practice of dynamical systems. In this thesis, we studied four fundamental questions arising in this field. The first question is: how can we efficiently verify whether a set is an invariant set for a linear continuous or discrete system. To answer this question, we derived sufficient and necessary conditions for several classical sets, which have wide range of applications, to be an invariant set for a linear system by presenting a novel, simple, and unified approach, which relies on optimization theory. Our sufficient and necessary conditions are more straightforward to use than using the definition directly for the verification of invariant sets. The second question is that when a discretization method is applied to solve a continuous system, how can we ensure that the discretization method is invariance preserving, i.e., the invariant set for the continuous system is also an invariant set for the discretized system. To answer this question, we studied three classic types of discretization methods for linear system, and proposed novel approaches to calculate a valid or largest uniform steplength threshold for invariance preserving. These methods have the potential to significantly influence computational practice as they enable us to identify a pre-specified steplength threshold for invariance preservation. The third question is an extension of the second question: how

can we generalize the existence of uniform steplength threshold for invariance preserving for general sets and systems. To answer this question, we established a novel theory to ensure positive local and uniform steplength threshold for invariance preserving on a set when a discretization method is applied to a linear or nonlinear dynamical system. Our methodology not only applied to classic sets, discretization methods, and dynamical systems, but also extended to more general sets, discretization methods, and dynamical systems. The last question is: how can we extend the first question to nonlinear systems. To answer this question, we used optimization methodology to derive invariance conditions for some classic sets for nonlinear dynamical systems.

In Chapter 2, we explore invariance conditions for four classic convex sets, for both linear discrete and continuous systems. In particular, these four convex sets are polyhedra, polyhedral cones, ellipsoids, and Lorenz cones, all of which have a wide range of applications in control theory. In this chapter, we present a novel, simple and unified method to derive invariance conditions for linear dynamical systems. We first consider discrete systems, followed by continuous systems, since invariance conditions of the latter one are derived by using invariance condition of the former one. For discrete systems, to derive invariance conditions, we introduce the Theorems of Alternatives, i.e., the Farkas Lemma and the S-lemma. We also show that by applying the S-lemma one can extend invariance conditions to any set represented by a quadratic inequality. The connection between discrete systems and continuous systems is built by using the forward or backward Euler methods, while invariance is preserved with sufficiently small step size. Then we use elementary methods to derive invariance conditions for continuous systems. In Chapter 2 we not only present invariance conditions of the four convex sets for continuous and discrete systems by using simple proofs, but also establishes a framework, which may be used for other convex sets as invariant sets, to derive invariance conditions for both continuous and discrete systems.

In Chapter 3, we consider invariance preserving steplength thresholds on polyhedron, when the discrete system is obtained by using special classes of discretization methods. Many real world problems are studied by developing dynamical system models. In practice, continuous systems are usually solved by using discretization methods. We particularly study three classes of discretization methods, which are: the forward Euler method, Taylor

approximation type, and rational function type. For the forward Euler method we prove that the largest steplength threshold can be obtained by solving a finite number of linear optimization problems. For the second 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 last 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 a method analogous to the one used in the first case.

In Chapter 4, we propose a theory of the existence of local or uniform invariance preserving steplength thresholds for large class of discretization methods. Existing results usually rely on the assumption that the explicit Euler method has an invariance preserving steplength threshold. In this chapter, first we study the existence and the quantification of local and uniform invariance preserving steplength threshold for Euler methods on special sets, namely, polyhedra, ellipsoids, or Lorenz cones. Our novel proofs are using only elementary concepts. We also extend our results and proofs to general convex sets, compact sets, and proper cones when a general discretization method is applied to linear or nonlinear dynamical systems. Conditions for the existence of a uniform invariance preserving steplegnth threshold for discretization methods on these sets are presented. In practice, one can use our results as criteria to check if a discretization method is invariance preserving with a uniform steplength threshold.

In Chapter 5, we derive invariance condition of some classical sets for nonlinear dynamical systems by utilizing methodology analogous to the one presented in Chapter 2. This is motivated by the fact that most problems in the real world often show nonlinearity characteristics which are modeled by nonlinear dynamical systems. First, the Theorems of Alternatives, i.e., the nonlinear Farkas lemma and the S-lemma, together with Nagumo's Theorem are utilized to derive invariance conditions for discrete and continuous systems. Only some standard assumptions are needed to establish invariance of some broadly used convex sets, including polyhedral and ellipsoidal sets. Second, we establish optimization framework to computationally verify the invariance conditions. Finally, we derive the in-

variance conditions for these classic sets for nonlinear systems without any conditions.

6.2 Future Research

We now consider future research directions. One interesting direction is to design novel invariant sets so that the steplength threshold $\hat{\tau}$ for invariance preserving is large. We can consider polyhedral cones or Lorenz cones. We are also interested in deriving invariance conditions for the intersections of some convex sets, e.g., the intersections of several ellipsoids.

6.2.1 Research Direction 1:

For the design of novel invariant conic sets for discrete systems, Horváth [37] has constructed an invariant polyhedral cone in $\mathbb{R}^n_+$, i.e., in the positive orthant of $\mathbb{R}^n$.

Lemma 6.2.1. *[37] The following two statements are true:*

- *Let $s \in \mathbb{R}^n$ and $s > 0$, then for every $v \in \mathbb{R}^n$ there exists the minimum real number, denoted by $\sigma_s(v)$, such that $\sigma_s(v)s \pm v \geq 0$.*
- *Let $\{s_k\}$ be a basis of $\mathbb{R}^n$, $s_1 > 0$. Suppose a vector $v = \sum_{k=1}^n \eta_k s_k \in \mathbb{R}^n$ is given, where $\eta_1 \geq 0$, and $\sum_{k=1}^n \sigma_{s_1}(s_k)|\eta_k| \leq 2\eta_1$. Then $v \geq 0$.*

An intuitive explanation of Lemma 6.2.1 is given as follows: since $s_1 > 0$ means that s_1 is in the interior of $\mathbb{R}^n_+$, to ensure $v \in \mathbb{R}^n_+$, the weight on s_1 is required to dominate that on the other basis vectors. Thus, by Lemma 6.2.1, the following set is constructed.

$$\mathcal{C} = \Big\{ x \in \mathbb{R}^n | v = \sum_{k=1}^n \eta_k s_i, \ \sum_{k=1}^n \sigma_{s_1}(s_k)|\eta_k| \leq 2\eta_1 \Big\}. \tag{6.1}$$

Although the author mentions that the set $\mathcal{C}$ defined in (6.1) is, in fact, a polyhedral cone, we present the following theorem to prove this statement, and this theorem explicitly gives the extreme ray of the polyhedral cone. For simplicity, we do not present the proof.

Theorem 6.2.2. *The set $\mathcal{C}$ defined as (6.1) is a polyhedral cone in $\mathbb{R}^n$, and its extreme rays are $\sigma_{s_1}(s_k)s_1 \pm s_k$, for $k = 2, ..., n$, where $\sigma_{s_1}(s_k)$ is defined in Lemma 6.2.1.*

Then, the invariance condition under which a matrix A leaves $\mathcal{C}$ invariant is presented in [37]. In fact, this is the invariance condition for the discrete system.

Theorem 6.2.3. *[37] Assume that the eigenvalues of A are real and denoted by $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. The set of the eigenvectors $\{s_k\}$, where s_k corresponds to λ_k, is a basis of $\mathbb{R}^{n\times n}$. Let $s_1 > 0$, and the cone $\mathcal{C}$ be defined as (6.1). If $\lambda_1 \geq \max_{k\geq 2} |\lambda_k|$, then $A\mathcal{C} \subseteq \mathcal{C}$.*

The steplength threshold for invariance preserving is also studied in the following theorem.

Theorem 6.2.4. *[37] Let $A, \{\lambda_k\}, \{s_k\}$, and $\mathcal{C}$ be given as in Theorem 6.2.3, and assume $\lambda_1 \leq 0$. Let $\delta \in (0, \infty]$, and assume a rational function $r(t)$ is non-negative and strictly increasing on $[-\delta, 0]$, and $r(-\delta) \geq |r(t)|$ for all $t \in (-\infty, -\delta)$. Then $r(A\Delta t)\mathcal{C} \subseteq \mathcal{C}$ for every $\Delta t \in [0, -\delta/\lambda_1]$.*

According to the discussion above, we now propose the potential method to design novel Lorenz cones in $\mathbb{R}^n_+$ by investigating the eigen-structure of the coefficient matrix of the dynamical system. Applying similar ideas as in [37], we may use the eigenvalues and eigenvectors of the coefficient matrix to design Lorenz cones. We use the same notations as in Theorem 6.2.3. Our idea is presented as follows:

1. similar to Theorem 6.2.3, let $s_1 > 0$.
2. for every $s_k, k \geq 2$, we choose $\tilde{\sigma}_k$, such that $\tilde{\sigma}_k s_1 \pm s_k \geq 0$.
3. use $\{\tilde{\sigma}_k s_1 \pm s_k\}_{k\geq 2}$ to design an ellipsoid $\mathcal{E}$ in $\mathbb{R}^n_+$.
4. the Lorenz cone $\mathcal{C}$ is designed by choosing $\mathcal{E}$ as its base and the origin as its vertex.

Two difficulties immediately arise:

- How to derive the explicit description of the constructed ellipsoid in Step 3? The difficulty is mainly due to that the constructed ellipsoid has to be in $\mathbb{R}^n_+$.
- How to compute $\hat{\tau}$ for a given cone?

6.2.2 Research Direction 2

For the Lorenz cone, we are also interested in considering it in $\mathbb{R}^n_+$. This is since several problems are usually introduced and studied in $\mathbb{R}^n_+$ in practice. The so called Dikin ellipsoid which is originally developed in optimization theory has caught our attention.

Definition 6.2.5. *[22] Let $\mathcal{F} = \mathbb{R}^n_+$ and $\bar{x} \in \text{int}(\mathcal{F})$. The **Dikin ellipsoid** around $\bar{x}$ is*

$$\mathcal{E}_{\bar{x},\delta} = \left\{x \in \mathbb{R}^n \mid (x-\bar{x})^T H(x-\bar{x}) \leq \delta \leq 1\right\} = \left\{x \in \mathbb{R}^n \mid \sum_{i=1}^{n} \frac{(x_i - \bar{x}_i)^2}{\bar{x}_i^2} \leq \delta \leq 1\right\}, \quad (6.2)$$

where $H = \text{diag}\{\bar{x}_1^{-2}, \bar{x}_2^{-2}, ..., \bar{x}_n^{-2}\}$.

One of the most significant property of a Dikin ellipsoid is presented as follows.

Theorem 6.2.6. *Let $\mathcal{F} = \mathbb{R}^n_+$, $\bar{x} \in \text{int}(\mathcal{P})$, then $\mathcal{E}_{\bar{x},\delta} \subseteq \mathcal{F}$.*

According to the discussion above, we now propose a potential method to design novel Lorenz cones in $\mathbb{R}^n_+$ by using Dikin ellipsoids. By Theorem 6.2.6, we have that $x \in \mathcal{E}_{\bar{x},\delta}$ implies $x \geq 0$. Our idea is presented as follows:

1. use a hyperplane to intersect the Dikin ellipsoid through its center, resulting an ellipsoidal intersection.
2. the intersection is then used as the base of the cone and the origin as its vertex.

Two difficulties immediately arise:

- How to derive an explicit description of the constructed cone? Although the formulaes of the hyperplane and the Dikin ellipsoid are given, the formula of the constructed cone is not trivial to obtain especially in high dimensional space.
- How to compute $\hat{\tau}$ for the constructed Lorenz cone? We have deep analysis for polyhedral sets, but Lorenz cones are quite different from polyhedral cones. To solve this difficulty, we might use optimization techniques. By solving appropriate optimization problems, we can obtain a valid threshold.

Assume Lorenz cones in $\mathbb{R}^n_+$ have been designed by either the first or the second method given as above, a challenge for the constructed Lorenz cones is given as follows:

For a given discretization method, how to choose a "good" invariant cone, where "good" means having large $\hat{\tau}$? We need to properly choose the position and shape of the constructed cone, and then compute $\hat{\tau}$ to choose the cone with large $\hat{\tau}$.

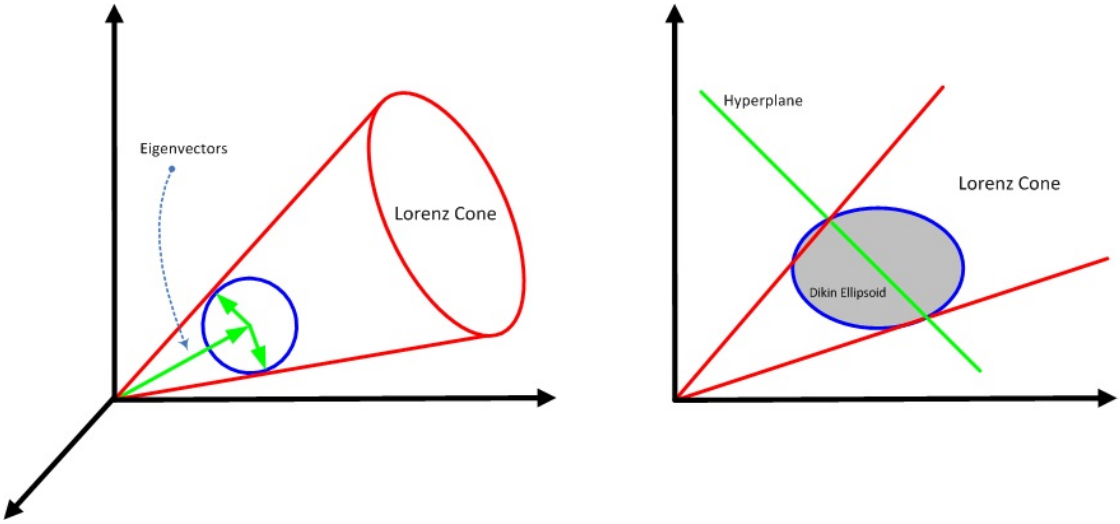

Figure 6.1: Two different ways to design novel invariant sets.

Chapter 7

Appendix

In this Appendix, we introduce basic concepts and theorems used in this thesis.

7.1 Basic Concepts

Definition 7.1.1. *A **polyhedron**, denoted by $\mathcal{P} \subseteq \mathbb{R}^n$, can be defined as the intersection of a finite number of half-spaces:*

$$\mathcal{P} = \{x \in \mathbb{R}^n \,|\, Gx \leq b\}, \tag{7.1}$$

where $G \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$

An equivalently definition of a polyhedral set is given as follows:

Definition 7.1.2. *A **polyhedron**, denoted by $\mathcal{P} \subseteq \mathbb{R}^n$, can be defined the set of points that can be given as the sum of the convex combination of a finite number of points and the conic combination of a finite number of vectors:*

$$\mathcal{P} = \Big\{x \in \mathbb{R}^n \,|\, x = \sum_{i=1}^{\ell_1} \theta_i x^i + \sum_{j=1}^{\ell_2} \hat{\theta}_j \hat{x}^j,\ \sum_{i=1}^{\ell_1} \theta_i = 1, \theta_i \geq 0, \hat{\theta}_j \geq 0\Big\}, \tag{7.2}$$

*where $x^1, ..., x^{\ell_1}, \hat{x}^1, ..., \hat{x}^{\ell_2} \in \mathbb{R}^n$. The **vertices** of $\mathcal{P}$ form a subset of $x^i, i \in \mathcal{I}(\ell_1)$, and the **extreme rays** of $\mathcal{P}$ are represented as $x^i + \alpha \hat{x}^j, \alpha > 0$, for some $i \in \mathcal{I}(\ell_1)$ and $j \in \mathcal{I}(\ell_2)$.*

We highlight that a bounded polyhedron, i.e., $\ell_2 = 0$ in (7.2), is called a *polytope*.

Definition 7.1.3. *A **polyhedral cone**, denoted by $\mathcal{C}_{\mathcal{P}} \subseteq \mathbb{R}^n$, can be also considered as a special class of polyhedra, and it can be defined as:*

$$\mathcal{C}_{\mathcal{P}} = \{x \in \mathbb{R}^n \,|\, Gx \leq 0\}, \tag{7.3}$$

or equivalently,

$$\mathcal{C}_{\mathcal{P}} = \Big\{x \in \mathbb{R}^n \,|\, x = \sum_{j=1}^{\ell} \hat{\theta}_j \hat{x}^j,\ \hat{\theta}_j \geq 0\Big\}, \tag{7.4}$$

*where $G \in \mathbb{R}^{m \times n}$ and $\hat{x}^1, ..., \hat{x}^\ell \in \mathbb{R}^n$. The **extreme rays** of $\mathcal{C}_{\mathcal{P}}$ are $\hat{x}^j, j > 0$.*

An ellipsoid is defined as follows:

Definition 7.1.4. *An **ellipsoid**, denoted by $\mathcal{E} \subseteq \mathbb{R}^n$, centered at the origin, is defined as:*

$$\mathcal{E} = \{x \in \mathbb{R}^n \,|\, x^T Q x \leq 1\}, \tag{7.5}$$

where $Q \in \mathbb{R}^{n \times n}$ and $Q \succ 0$.

We note that any ellipsoid with nonzero center can be transformed to an ellipsoid centered at the origin.

Definition 7.1.5. *A **Lorenz cone**[1], denoted by $\mathcal{C}_{\mathcal{L}} \subseteq \mathbb{R}^n$, with vertex at the origin, is defined as:*

$$\mathcal{C}_{\mathcal{L}} = \{x \in \mathbb{R}^n \,|\, x^T Q x \leq 0,\ x^T u_n \geq 0\}, \tag{7.6}$$

where $Q \in \mathbb{R}^{n \times n}$ is a symmetric nonsingular matrix with exactly one negative eigenvalue λ_n, i.e., $\text{inertia}\{Q\} = \{n-1, 0, 1\}$, and u_n is the eigenvector corresponding to the only negative eigenvalue λ_n.

Similar to ellipsoids, any Lorenz cone with nonzero vertex can be transformed to a Lorenz cone with vertex at the origin. For every Lorenz cone, given as in (7.6), there exists an orthonormal basis $\{u_1, u_2, ..., u_n\}$, i.e., $u_i^T u_j = \delta_{ij}$, where u_i is the eigenvector corresponding to the eigenvalue, λ_i, of Q, and δ_{ij} is the Kronecker delta function, such

[1] A Lorenz cone is sometimes also called an ice cream cone, a second order cone, or an ellipsoidal cone, see, e.g., [58].

that $Q = U\Lambda^{\frac{1}{2}}\tilde{I}\Lambda^{\frac{1}{2}}U^T$, where $U = [u_1, u_2, ..., u_n], \Lambda^{\frac{1}{2}} = \text{diag}\{\sqrt{\lambda_1}, ..., \sqrt{\lambda_{n-1}}, \sqrt{-\lambda_n}\}$ and $\tilde{I} = \text{diag}\{1, ..., 1, -1\}$.

Definition 7.1.6. *A **standard Lorenz cone**, denoted by $\mathcal{C}^*_{\mathcal{L}} \subseteq \mathbb{R}^n$, with vertex at the origin is defined as:*

$$\mathcal{C}^*_{\mathcal{L}} = \{x \in \mathbb{R}^n \,|\, x^T\tilde{I}x \leq 0, x^T e_n \geq 0\}, \tag{7.7}$$

where $\tilde{I} = \text{diag}\{1, ..., 1, -1\}$ and $e_n = (0, ..., 0, 1)^T$.

7.2 Basic Theorems

In this section, basic theorems related to, or used as tools to study, invariant sets for dynamical systems are introduced.

The Farkas lemma [60] and the S-lemma [56, 79], both of which are also called the Theorem of Alternatives, are fundamental tools to derive invariance conditions for discrete systems. The S-lemma, first proved by Yakubovich [79], is somewhat analogous to a special case of the nonlinear Farkas lemma, see Pólik and Terlaky [56].

Theorem 7.2.1. (Farkas lemma [60]) *Let $P \in \mathbb{R}^{m\times n}$, $d \in \mathbb{R}^m, c \in \mathbb{R}^n$, and $\beta \in \mathbb{R}$. Then the following two statements are equivalent:*

1. *There is no $y \in \mathbb{R}^m$, such that $P^T y \leq c$ and $d^T y > \beta$;*
2. *There exists a vector $z \in \mathbb{R}^n$, such that $z \geq 0, Pz = d$, and $c^T z \leq \beta$.*

Theorem 7.2.2. (S-lemma [56, 79]) *Let $g_1(y), g_2(y) : \mathbb{R}^n \to \mathbb{R}$ be quadratic functions, and suppose that there is a $\hat{y} \in \mathbb{R}^n$ such that $g_2(\hat{y}) < 0$. Then the following two statements are equivalent:*

1. *There exists no $y \in \mathbb{R}^n$, such that $g_1(y) < 0, g_2(y) \leq 0$.*
2. *There exists a scalar $\rho \geq 0$, such that $g_1(y) + \rho g_2(y) \geq 0$, for all $y \in \mathbb{R}^n$.*

Theorem 7.2.3. (Convex Farkas lemma [56, 60]) *Let $h(y), g_1(y), g_2(y), ..., g_m(y) : \mathbb{R}^n \to \mathbb{R}$ be convex functions. Assume the Slater condition[2] is satisfied. Then the following*

[2]The Slater condition means that there exists a $\hat{y} \in \mathbb{R}^n$, such that $g_j(\hat{y}) \leq 0$ for all j when $g_j(x)$ is linear and $g_j(\hat{y}) < 0$ for all j when $g_j(x)$ is nonlinear.

two statements are equivalent:

1. *The inequality systems* $h(y) < 0,\ g_j(y) \leq 0, j = 1, 2, ..., m$ *have no solution.*

2. *There exist* $\beta_1, \beta_2, ..., \beta_m \geq 0$*, such that* $h(y) + \sum_{j=1}^{m} \beta_j g_j(y) \geq 0$ *for all* $y \in \mathbb{R}^n$.

The concept of *tangent cone* plays an important role in the analysis for invariance conditions for continuous systems.

Definition 7.2.4. *Let* $\mathcal{S} \subseteq \mathbb{R}^n$ *be a closed convex set, and* $x \in \mathcal{S}$*. The* ***tangent cone*** *of* $\mathcal{S}$ *at* x*, denoted by* $\mathcal{T}_{\mathcal{S}}(x)$*, is given as*

$$\mathcal{T}_{\mathcal{S}}(x) = \Big\{ y \in \mathbb{R}^n \ \Big|\ \liminf_{t \to 0^+} \frac{\text{dist}(x + ty, \mathcal{S})}{t} = 0 \Big\}, \tag{7.8}$$

where $\text{dist}(x, \mathcal{S}) = \inf_{s \in \mathcal{S}} \|x - s\|$.

The following classic result proposed by Nagumo [52] provides a general criterion to determine whether a closed convex set is an invariant set for a continuous system. This theorem, however, is not valid for discrete systems, for which one can find a counterexample in [13].

Theorem 7.2.5. (Nagumo [13, 52]) *Let* $\mathcal{S} \subseteq \mathbb{R}^n$ *be a closed convex set, and assume that the system* $\dot{x}(t) = f(x(t))$*, where* $f : \mathbb{R}^n \to \mathbb{R}^m$ *is a continuous mapping, admits a globally unique solution for every initial point* $x(0) \in \mathcal{S}$*. Then* $\mathcal{S}$ *is an invariant set for this system if and only if*

$$f(x) \in \mathcal{T}_{\mathcal{S}}(x),\ \textit{for all } x \in \partial\mathcal{S}, \tag{7.9}$$

where $\mathcal{T}_{\mathcal{S}}(x)$ *is the tangent cone of* $\mathcal{S}$ *at* x.

Nagumo's Theorem 7.2.5 has a geometrical interpretation as follows: for any trajectory that starts in $\mathcal{S}$, it has to go through $\partial\mathcal{S}$ if it will go out of $\mathcal{S}$. Then one needs only to consider the property of this trajectory on $\partial\mathcal{S}$. Note that $f(x)$ is the derivative of the trajectory, thus (7.9) ensures that the trajectory will point inside $\mathcal{S}$ on the boundary, which means $\mathcal{S}$ is an invariant set. The disadvantage of Theorem 7.2.5, however, is that it may be difficult to verify whether (7.9) holds for all points on the boundary of a given set.

Bibliography

[1] A.M. Aliluiko and O.H. Mazko. Invariant cones and stability of linear dynamical systems. *Ukrainian Mathematical Journal*, 58(11):1635–1655, 2006.

[2] G. Baker, Jr. and P. Graves-Morris. *Padé Approximants*. Cambridge University Press, New York, NY, 1996.

[3] A. Barvinok. *A Course in Convexity*. American Mathematical Society, 2002.

[4] R. Bellman. *Introduction to Matrix Analysis*. SIAM Studies in Applied Mathematics, Philadelphia, second edition, 1987.

[5] A. Berman and R. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. SIAM Studies in Applied Mathematics, SIAM, Philadelphia, PA, 1994.

[6] D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, Nashua, 1998.

[7] G. Birkhoff. Linear transformations with invariant cones. *The American Mathematical Monthly*, 74(3):274–276, 1967.

[8] G. Bitsoris. On the positive invariance of polyhedral sets for discrete-time systems. *System and Control Letters*, 11(3):243–248, 1988.

[9] G. Bitsoris. Positively invariant polyhedral sets of discrete-time linear systems. *International Journal of Control*, 47(6):1713–1726, 1988.

[10] F. Blanchini. Constrained control for uncertain linear systems. *Journal of Optimization Theory and Applications*, 71(3):465–484, 1991.

[11] F. Blanchini. Constrained control for uncertain linear systems. *International Journal of Optimization Theory and Applications*, 71(3):465–484, 1991.

[12] F. Blanchini. Nonquadratic Lyapunov functions for robust control. *Automatica*, 31(3):451–461, 1995.

[13] F. Blanchini. Set invariance in control. *Automatica*, 35(11):1747–1767, 1999.

[14] F. Blanchini and S. Miani. Constrained stabilization via smooth Lyapunov functions. *Systems and Control Letters*, 35(3):155–163, 1998.

[15] 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*, 5:82–86, 1999.

[16] C. Bolley and M. Crouzeix. Conservation de la positivite lors de la discretisation des problèmes d'évolution paraboliques. *RAIRO. Analyse Numérique*, 12:237–245, 1978.

[17] 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.

[18] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, 2004.

[19] E.B. Castelan and J.C. Hennet. On invariant polyhedra of continuous-time linear systems. *IEEE Transactions on Automatic Control*, 38(11):1680–1685, 1993.

[20] V. F. Demyanov, R. Fletcher, T. Terlaky, G. Di Pillo, and F. Schoen. *Nonlinear Optimization*. Springer-Verlag, Berlin Heidelberg, 2010.

[21] N.A. Derzko and A.M. Pfeffer. Bounds for the spectral radius of a matrix. *Mathematics of Computation*, 19(89):62–67, 1965.

[22] I.I. Dikin. Iterative solution of problems of linear and quadratic programming. *Doklady Akademii Nauk SSSR*, 174:747–748, 1967.

[23] C.E.T. Dórea and J.C. Hennet. (A, B)-invariance conditions of polyhedral domains for continuous-time systems. *European Journal of Control*, 5:70–81, 1999.

[24] L. Elsner. On matrices leaving invariant a nontrivial convex set. *Linear Algebra and its Applications*, 42:103–107, 1982.

[25] K. Feng. On difference schemes and symplectic geometry. *Proceedings of the 5th International Symposium on Differential Geometry and Differential Equations*, pages 42–58, 1985.

[26] S. Gottlieb, D. Ketcheson, and C.-W. Shu. *Strong Stability Preserving Runge–Kutta and Multistep Time Discretizations.* World Scientific, Hackensack, NJ, 2011.

[27] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Review*, 43(1):89–112, 2001.

[28] M. Grant and S. Boyd. CVX: MATLAB Software for Disciplined Convex Programming: CVX Research. `http://cvxr.com/cvx/`, Nov., 2012.

[29] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving ordinary differential equations I: Nonstiff problems.* Springer-Verlag, New York, NY, 1993.

[30] E. Haynsworth, M. Fiedler, and V. Pták. Extreme operators on polyhedral cones. *Linear Algebra and its Applications*, 13:163–172, 1976.

[31] J.-C. Hennet. Discrete-time constrained linear systems. *Control and Dynamical Systems*, 71:157–213, 1995.

[32] N. Higham. *Accuracy and Stability of Numerical Algorithms.* SIAM, Philadelphia, PA, 2002.

[33] N. Higham. *Functions of Matrices: Theory and Computation.* SIAM, Philadelphia, PA, 2008.

[34] F. Hillier and G. Lieberman. *Introduction to Operations Research.* Holden-Day, Inc., San Francisco, CA, 4th edition, 1986.

[35] J.B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I.* Springer-Verlag, New York, NY, 1993.

[36] R. Horn and C. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1990.

[37] Z. Horváth. On the positivity of matrix-vector products. *Linear Algebra and its Applications*, 393:253–258, 2004.

[38] Z. Horváth. Invariant cones and polyhedra for dynamical systems. *Proceeding of the International Conference in Memoriam Gyula Farkas*, pages 65–74, 2005.

[39] Z. Horváth. On the positivity step size threshold of Runge-Kutta methods. *Applied Numerical Mathematics*, 53:341–356, 2005.

[40] Z. Horváth, Y. Song, and T. Terlaky. Invariance preserving discretizations of dynamical systems. *Lehigh University Report*, 2013.

[41] Z. Horváth, Y. Song, and T. Terlaky. Steplength thresholds for invariance preserving of discretization methods of dynamical systems on a polyhedron. *Discrete and Continuous Dynamical Systems - Series A*, 35(7):2997–3013, 2015.

[42] Z. Horváth, Y. Song, and T. Terlaky. *Invariance Conditions for Nonlinear Dynamical Systems.* Optimization and Applications in Control and Data Science, Optimization and Its Applications. Springer, 2016.

[43] Z. Horváth, Y. Song, and T. Terlaky. A novel unified approach to invariance conditions for a linear dynamical system. *forthcoming to Applied Mathematics and Computation*, 2016.

[44] H.K. Khalil. *Nonlinear Systems.* Prentice Hall, 2001.

[45] K.I. Kouramas, S.V. Rakovic, E.C. Kerrigan, and J.C. Allwright. On the minimal robust positively invariant set for linear difference inclusions. pages 2296–2301, 2005.

[46] J.F.B.M. Kraaijevanger. Absolute monotonicity of polynomials occurring in the numerical solution of initial value problems. *Numerische Mathematik*, 48:303–322, 1986.

[47] D. Li, X. Wu, and J. Lu. Estimating the ultimate bound and positively invariant set for the hyperchaotic Lorenz–Haken system. *Chaos, Solitions & Fractals*, 39(15):1290–1296, 2009.

[48] Z. Lin, A. Saberi, and A. Stoorvogel. Semi-global stabilization of linear discrete-time systems subject to input saturation via linear feedback - an ARE-based approach. *IEEE Transactions on Automatic Control*, 41(8):1203–1207, 1996.

[49] R. Loewy and H. Schneider. Positive operators on the n-dimensional ice cream cone. *Journal of Mathematical Analysis and Applications*, 49(2):375–392, 1975.

[50] D.W. Markiewicz. Survey on symplectic integrators. Technical report, University of California at Berlekey, 1999.

[51] K. Meyer, G. Hall, and D. Offin. Symplectic transformations. *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, 90:133–145, 2009.

[52] M. Nagumo. Über die Lage der Integralkurven gewöhnlicher Differentialgleichungen. *Proceeding of the Physical-Mathematical Society, Japan*, 24(3):551–559, 1942.

[53] L. Perko. *Differential Equations and Dynamical Systems.* Texts in Applied Mathematics. Springer, 2006.

[54] B. Pluymers, J.A. Rossiter, J.A.K. Suykens, and B. De Moor. The efficient computation of polyhedral invariant sets for linear systems with polytopic uncertainty. volume 2, pages 804–809, 2005.

[55] A. Polanski. On infinity norm as Lyapunov functions for linear systems. *IEEE Transactions on Automatic Control*, 40(7):1270–1274, 1995.

[56] I. Pólik and T. Terlaky. A survey of the S-lemma. *SIAM Review*, 49(3):371–418, 2007.

[57] S.V. Rakovic, E.C. Kerrigan, K.I. Kouramas, and D.Q. Mayne. Invariant approximations of the minimal robust positively invariant set. *IEEE Transactions on Automatic Control*, 50(3):406–410, 2005.

[58] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

[59] L. Rodman, H. Seyalioglu, and I.M. Spitkovsky. On common invariant cones for families of matrices. *Linear Algebra and its Applications*, 432(1):911–926, 2010.

[60] C. Roos, T. Terlaky, and J.-Ph. Vial. *Interior Point Methods for Linear Optimization*. Springer Science, Boston, 2006.

[61] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Book Co., New York, NY, third edition, 1976.

[62] H. Schneider and M. Vidyasagar. Cross-positive matrices. *SIAM Journal on Numerical Analysis*, 7(4):508–519, 1970.

[63] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of Computational Physics*, 77:439–471, 1988.

[64] M.N. Spijker. Contractivity in the numerical solution of initial value problems. *Numerische Mathematik*, 42:271–290, 1983.

[65] P. Stein. Some general theorems on iterants. *Journal of Research of National Bureau of Standards*, 48(1):82–83, 1952.

[66] R. Stengel. *Optimal Control and Estimation*. Dover Publications, 1994.

[67] R. Stern. On strictly positively invariant cones. *Linear Algebra and its Applications*, 48:13–24, 1982.

[68] R. Stern and H. Wolkowicz. Exponential nonnegativity on the ice cream cone. *SIAM Journal on Matrix Analysis and Applications*, 12(1):160–165, 1991.

[69] J. Sturm, I. Pólik, and T. Terlaky. SeDuMi. `http://sedumi.ie.lehigh.edu`, Apr., 2010.

[70] B. Sturmfels. *Solving Systems of Polynomial Equations*. CBMS Lectures Series, American Mathematical Society, 2002.

[71] B.-S. Tam. *Extreme positive operators on convex cones*. Five Decades as a Mathematician and Educator: On the 80th Birthday of Prof. Yung-Chow Wong, (Eds. K.Y. Chan and M.C. Liu). World Scientific Publishing Company, 1995.

[72] B.-S. Tam. A cone-theoretic approach to the spectral theory of positive linear operators: the finite-dimensional case. *Taiwanes Journal of Mathematics*, 5(2):207–277, 2001.

[73] S. Tarbouriech and C. Burgat. Positively invariant sets for constrained continuous-time systems with cone properties. *IEEE Transactions on Automatic Control*, 39(2):401–405, 1994.

[74] A. Tiwari, J. Fung, R. Bhattacharya, and R. M. Murray. Polyhedral cone invariance applied to rendezvous of multiple agents. In *43rd IEEE Conference on Decision and Control*, pages 165–170, 2004.

[75] K.C. Toh, M. Todd, and R. Tütüncü. SDPT3 version 4.0 – a MATLAB software for semidefinite-quadratic-linear programming. `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`, Feb., 2009.

[76] M. E. Valcher and L. Farina. An algebraic approach to the construction of polyhedral invariant cones. *SIAM Journal on Matrix Analysis and Applications*, 22(2):453–471, 2000.

[77] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

[78] P. Wolfe. A duality theorem for non-linear programming. *Quarterly of Applied Mathematics*, 19:239–244, 1961.

[79] V. Yakubovich. S-procedure in nonlinear control theory. *Vestnik Leningrad University*, (1):62–77, 1971.

[80] P. Yu and X. Liao. Globally attractive and positive invariant set of the lorenz system. *International Journal of Bifurcation and Chaos*, 16:757–764, 2006.

[81] F. Zhang, Y. Shu, H. Yang, and X. Li. Estimating the ultimate bound and positively invariant set for a synchronous motor and its application in chaos synchronization. *Chaos, Solitions & Fractals*, 44:137–144, 2011.

[82] X. Zhang and C.-W. Shu. On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. *Journal of Computational Physics*, 229:8918–8934, 2010.

[83] X. Zhang and C.-W. Shu. Positivity-preserving high order discontinuous Galerkin schemes for compressible euler equations with source terms. *Journal of Computational Physics*, 230:1238–1248, 2011.

[84] X. Zhang and C.-W. Shu. Positivity-preserving high order finite difference WENO schemes for compressible euler equations. *Journal of Computational Physics*, 231:2245–2258, 2012.

[85] H. Zhao. Invariant set and attractor of nonautonomous functional differential systems. *Journal of Mathematical Analysis and Applications*, 282(15):437–443, 2003.

[86] K. Zhou, J. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice Hall, New Jersey, first edition, 1995.

Biography

Yunfei Song was born in Hebei, China. He obtained a bachelor degree in mathematics from Nankai University in China, and double master degrees in mathematics and finance in the State University of New York at Buffalo. He served as the president of INFORMS Chapter at Lehigh and the president of Chinese Students and Scholars Associations at Lehigh. His professional experience include internships at TX Investment Consulting as Quantitative Analyst working on portfolio construction and risk management, SAS as a technician working on statistical and optimization software development, Argonne National Laboratory as Givens Associate working on optimization algorithm design and implementation, and Mitsubishi Electric Research Laboratory (MERL) as Research Scientist working on advanced numerical methods to solve advection diffusion equation as well as the application in air conditioners.