Springer Tracts in Mechanical Engineering

Zhengzhi Han
Xiushan Cai
Jun Huang

Theory of Control Systems Described by Differential Inclusions

上海交通大學出版社
Shanghai Jiao Tong University Press

Springer

Springer Tracts in Mechanical Engineering

About this Series

Springer Tracts in Mechanical Engineering (STME) publishes the latest developments in Mechanical Engineering - quickly, informally and with high quality. The intent is to cover all the main branches of mechanical engineering, both theoretical and applied, including:

- Engineering Design
- Machinery and Machine Elements
- Mechanical structures and stress analysis
- Automotive Engineering
- Engine Technology
- Aerospace Technology and Astronautics
- Nanotechnology and Microengineering
- Control, Robotics, Mechatronics
- MEMS
- Theoretical and Applied Mechanics
- Dynamical Systems, Control
- Fluids mechanics
- Engineering Thermodynamics, Heat and Mass Transfer
- Manufacturing
- Precision engineering, Instrumentation, Measurement
- Materials Engineering
- Tribology and surface technology

Within the scopes of the series are monographs, professional books or graduate textbooks, edited volumes as well as outstanding Ph.D. theses and books purposely devoted to support education in mechanical engineering at graduate and post-graduate levels.

More information about this series at http://www.springer.com/series/11693

Zhengzhi Han • Xiushan Cai • Jun Huang

Theory of Control Systems Described by Differential Inclusions

上海交通大學出版社
Shanghai Jiao Tong University Press

Springer

Zhengzhi Han
Shanghai Jiao Tong University
Shanghai, China

Xiushan Cai
Zhejiang Normal University
Jinhua, China

Jun Huang
Soochow University
Suzhou, China

ISSN 2195-9862 ISSN 2195-9870 (electronic)
Springer Tracts in Mechanical Engineering
ISBN 978-3-662-49243-7 ISBN 978-3-662-49245-1 (eBook)
DOI 10.1007/978-3-662-49245-1

Library of Congress Control Number: 2015960942

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Berlin Heidelberg

Preface

When I give lecture to my students, I prefer to introduce the history of inventions in engineering development rather than to state only theoretically significance. It seems that my students very like to hear these stories which make them be interested in the teaching materials presented in the classes and understand the engineering background. Hence, I spent lots of time to read and collect these stories; of course, most of them were written by Chinese.

I think the development of control theory may be divided into three phases.

From the book of Joseph T. M. Needham, south-pointed cart designed and manufactured by Chinese people before Christ was an automatic mechanic device. On the cart, there was a puppet made of wood, one arm of the puppet always pointed to the south no matter how to turn the cart. Needham said it might be the first automatic device in the world.

However, the south-pointed cart was a legend. Nobody has seen the design or sketch drawing and heard any information from archaeological experts. A great invention of automatic device was the centrifugal governor for the steam machine. James Watt held three inventions for steam machines; the centrifugal governor was the last one by which the Industrial Revolution solved the problem of power and then happened. Many researchers agreed that the centrifugal governor was the first automatic control device and used until now. Until the invention of centrifugal governor by Watt, there had been several controlled devices, but no control theory.

After the invention, lots of physicians investigated the reason why the small device could make the speed stable for a huge steam machine. It was said that J. Maxwell also considered the problem, but I did not find any valuable results about his investigation in this area. The problem was solved by mathematicians who worked in the field of differential equations; they established the theory of stability and found the conditions for the stability. Using their words, the governor made the plant stable. Among these mathematicians, two people, Jules H. Poincaré and Aleksandr M. Lyapunov, have to be mentioned. They independently presented the stability theory for differential equations and gave vital effects for the subsequent investigation of the stability of differential equations.

In all fairness, the research of stability of differential equations is a part of control theory. The governor is a controller, the controlled steam machine is a control system, and the controller stabilizes the plant by the words of the twentieth century. Till now, every classical textbook of control theory has to state the stability of control systems. However, at the years of Poincaré and Lyapunov, there was no such a subject of control theory. On the other hand, the theory of differential equation liked the sun at noon so that its bright would make others lose their colors. Hence, I called the phase the *Pregnancy Period of Control Theory*. There existed controller design and the researches of stability and stabilization, but there was no terminology of control theory. It is remarkable that the mathematical model used in this pregnancy phase was differential equations.

It was commonly recognized the birth of control theory was in World War II. By 1938, two excellent experts disappeared from the public field of vision. Herdrik W. Bode and Norbert Weiner were recruited by the army to improve the control of the radar and anti-aircraft fire. From the film of *Eagles over London*, one may understand the contribution of the control of anti-aircraft systems. After the war, MIT opened a course of *Principle of Servo Motors*. The course revealed the theory used in anti-aircraft system control. It was also remarkable that the research used the transfer function as mathematical model to analyze and design controllers. The control theory established at that time was called by *Classical Control Theory* in China. It also was called as 2P theory since the design could be completed by using Pencil + Paper. 2P has great significance because it solved the control system design in the computerless era.

Most people love the new and hate the old. When a new theory was born, people were happy and celebrated its birth, and then they found lots shortcomings of the theory, and then started to complain. The classical control theory can handle single-input-single-output (SISO) systems very effectively, but it is hard to treat the multi-input and multi-output (MIMO) systems. By 1960, Rudolph E. Kalman applied state space description to research the properties of control systems, so that the MIMO systems could be treated as the same as SISO systems. Two important concepts, controllability and observability were defined. When Kalman proposed his paper, his invention was not attached enough importance. About 2 years later, the Kalman filter was successfully applied in aerospace technology, people then looked the state space method with new eyes and called it *Modern Control Theory*. In 1982, Kalman visited China, after his presentation, one person asked a question about the difference between the Kalman filter and Wiener filter. After thinking a while, Kalman answered the question with a new question, he said: Do you know the reason why the Soviet Union launched manned spacecraft before USA, but USA arrived in the Moon first? With one more minutes silence, Kalman answered the question himself, “It was because the Soviet Union people did not understand Kalman filter.” As an evidence, I knew in the aerospace engineering Chinese scientists are applying Kalman filter to process the information sent back by Chang’e. It is no doubt, the modern control theory uses state space model in its investigation.

What can we conclude from the history of control theory? I draw two conclusions: The mathematical model is the mark for the developing phases of control theory; The birth of new theory needs an engineering background.

Since 1980s, I have heard many times that the third-generation control theory has been born, such as the theory of large-scale systems, the theory of intelligent control systems, the theory for complex networks, etc. Except the mathematical models have been renewed, the birth of a new theory was not accepted by most people.

In the turn of this century, the word of *uncertainty* appeared on the desks of many scientists who were working on different research areas. It is very usual that the same action will lead to different results. For example, there are ten resistances made in a production line at almost the same time. They are all labeled 1 K, and set on ten baseboards. After working 10 h, their resistance values may be similar, after 10 days, difference starts to happen, after 10 months, it is certain that these resistances may have more than seven values. It is necessary for us to use a set to describe these resistance values. The value of a simple resistance is a variable what can be said for a natural phenomenon? When we deal with the design of control system, the differential equation $\dot{x} = f(x, u)$ is applied as the description of control systems, where x is the state and u the control. The above discussion asserts, for a determined control u, you cannot expect the state is determined although the initial condition is fixed. The inner parameters, the environment conditions and other unknown causes lead to the uncertainty of function $f(x, u)$, and the change shows irregularly. Therefore, it is much suitable to treat $f(x, u)$ as a set-valued mapping, and the dynamic model becomes $\dot{x} \in f(x, u)$. This is a differential inclusion, and is a new mathematical model of control systems. Does the introduction of differential inclusion model result in a new era of the development of control theory? My answer is "no, I do not know." The present research results cannot support such a conclusion, but it probably leads to a new generation if there is a vital social necessary.

Before 1980, several authors applied differential inclusion to deal with control systems. At this period, they tried to extend the maximum principle to the differential inclusion systems. Similar conclusion has been established. By Year 2000, lots authors were interested in the controllability of such systems. In the twenty-first century, researchers applied differential inclusions to deal with Luré system and polytope system. This is the simple history of differential inclusion control theory. In the past almost four decades, the development of differential inclusion control theory is very slow. There are two excuses for the researchers working in the area. One is the development of differential inclusion theory is quite slow related to other fields of mathematics. The another reason may be the lack of engineering needs. In China, I could only find two books for the theory. One is *Differential Inclusions – Set-Valued Maps and Viability Theory* written by J-P Aubin and A. Cellina, and the another is *Introduction of the Theory of differential Inclusions* by G. V. Smirnov. In Amazon, I did not find other more. These facts motivated authors to spend almost 5 years to write this book, we really wish it can generalize the investigation of differential inclusion control systems.

The arrangement of the book is as follows. The first chapter provides preliminaries, and focuses on convex analysis which is elemental for further investigation. Chapter 2 introduces set-valued mappings and differential inclusions. The authors required themselves that the chapter can contain basic knowledge of differential inclusions so that readers can carry out farther research with a massy base. To understand the conclusions presented in this chapter needs knowledge of functional analysis and differential equations. The third chapter deals with convex processes which can be treated as an extension of linear mappings. The fourth chapter considers polytope control systems which can be looked as a convex hull of several linear systems. The last chapter is about Luré differential inclusion systems. Most of the chapter deals with the design of observers for the systems.

Both Ms. Xiushan Cai and Mr. Jun Huang were my students who studied under my supervision for their Ph. D. degrees. When they studied in Shanghai Jiao Tong University, they made many studies for the control of differential inclusion systems. Now they joined Zhejiang Normal University and Soochow University, respectively, and still are interested in research of such systems. My students Wei Zhang, Leipo Liu, Junfeng Zhang, Hai Wu, Shaojie Shen and Peiquan Wang joined seminars and discussions when they studied in my lab. We are grateful to their contributions.

Shanghai, China Zhengzhi Han
18 August 2015

Contents

Chapter 1
Convex Sets and Convex Functions

The first chapter introduces the fundamental concepts and conclusions of functional analysis so that readers can have a foundation for going on reading this book successfully and can also understand notations used in the book. The arrangement of this chapter is as follows: The first section deals with normed linear spaces and inner product spaces which both provide a platform for further investigation; the second section introduces convex sets, and the third section considers convex functions; the last section of this chapter introduces semi-continuous functions. These are all necessary for research of set-valued mappings and differential inclusions which are two key concepts in this book. The most materials given in this chapter are referred to Conway (1985) which has been widely used in Chinese universities.

1.1 Normed Spaces and Inner Product Spaces

This section introduces normed spaces and inner product spaces which are fundamental objectives of functional analysis. The normed space sometimes is called normed linear space or normed vector space. An inner space is also a normed space but it holds an inner product.

1.1.1 Sets and Mappings

The section starts with very basic concept of set. In this book, capital letters $A, B, \ldots, X, Y$ are applied to express sets and lower case letters $a, b, \ldots, x, y$ to express the elements of sets. Sometimes, a compound notation may be applied

Z. Han et al., *Theory of Control Systems Described by Differential Inclusions*,
Springer Tracts in Mechanical Engineering, DOI 10.1007/978-3-662-49245-1_1

to express a set whenever we need more information, for example, the notation $C\left([a,b],\mathbb{R}\right)$ is used to express the set of continuous real functions which are defined on the closed interval $[a,b]$ and take their values on $\mathbb{R}$. We will illustrate the meaning when a compound notation appears at its first time.

There are two special sets. One is the empty denoted by $\varnothing$, and another is the full set denoted by Ω which is the largest set in consideration.

Let A and B be two sets. Then $A\cup B$, $A\cap B$ and $A\backslash B$ are the union, the intersection, and the difference of A and B, respectively. $A\times B$ denotes the Cartesian product of A and B, i.e., $A\times B=\{(a,b)\,;a\in A,b\in B\}$, where (a,b) is an ordered pair of a and b. A^{c} is the complement of A, and $\mathscr{P}A$ is the power set of A. The de Morgan laws are the following two equations

$$\cup A^{\mathrm{c}}=(\cap A)^{\mathrm{c}};\quad \cap A^{\mathrm{c}}=(\cup A)^{\mathrm{c}}.$$

The most common sets used in this book are sets of numbers. $\mathbb{Z}$ is the set of integers, $\mathbb{Z}^{+}$ or $\mathbb{N}$ is the set of natural numbers. In this book, the set of natural numbers is defined by $\{1,2,3,\ \ldots\}$. Let $\mathbb{R}$ be the set of real numbers, and $\mathbb{C}$ the set of complex numbers. In this book, the set $\mathbb{C}$ is seldom applied. Let $\mathbb{R}^{+}$ and $\mathbb{R}^{-}$ be the sets of positive and negative real numbers, respectively. Sometimes, we apply $\mathbb{R}\,(>0)$ to denote $\mathbb{R}^{+}$, and $\mathbb{R}\,(<0)$ for $\mathbb{R}^{-}$. By such a usage, $\mathbb{R}\,(\geq 0)$ and $\mathbb{Z}\,(\geq 0)$ are the sets of nonnegative real numbers and nonnegative integers, respectively. The closure of $\mathbb{R}$ is denoted by $\mathscr{R}$, i.e., $\mathscr{R}=\mathbb{R}\cup\{\infty,-\infty\}$. We also denote $\mathbb{R}\,(\infty)$ and $\mathbb{R}\,(-\infty)$ for the sets of $\mathbb{R}\cup\{\infty\}$ and $\mathbb{R}\cup\{-\infty\}$, respectively. On the set $\mathscr{R}$, we can operate arithmetic, such as addition, subtraction, multiplication, and division. But the indefinite operations, such as $\infty-\infty$ and $0\cdot\infty$, are not allowed. Let $\mathbb{R}^{n}$ be the n-dimensional real space, i.e., $\mathbb{R}^{n}=\overbrace{\mathbb{R}\times\mathbb{R}\times\cdots\times\mathbb{R}}^{n}$. $\mathbb{R}^{n}$ is always treated as a linear space over $\mathbb{R}$.

A relation from A to B is a subset of $A\times B$. If $A=\{a_1,a_2,a_3,a_4\}$ and $B=\{b_1,b_2\}$, then $M=\{(a_1,b_1)\,,(a_1,b_2)\,,(a_3,b_1)\,,(a_4,b_1)\}$ is a relation from set A to set B. Two sets

$$\operatorname{dom}M=\{a;\ a\in A,\exists b\in B,b\neq\pm\infty,\text{such that}\ \ (a,b)\in M\}$$

and

$$\operatorname{rang}M=\{b;\ b\in B,\exists a\in A,\ \ \text{such that}\ (a,b)\in M\}$$

are called the domain and range of the relation M, respectively. For $a\in\operatorname{dom}M$, the image of a is defined by $M(a)=\{b;b\in B,(a,b)\in M\}$. Correspondingly, if $b\in\operatorname{rang}M$, $M^{-1}(b)=\{a;a\in A,(a,b)\in M\}$ is the inverse of b. Usually, $M(a)$ and $M^{-1}(b)$ may have more than one elements for a relation M. At the example given above, $M\left(a_1\right)=\{b_1,b_2\}$ and $M^{-1}\left(b_1\right)=\{a_1,a_3,a_4\}$.

The concepts of image and inverse can be extended from an element to set. For example, if $S_1 \subset \mathrm{dom}\, M$, then the image of S_1, denoted by $M(S_1)$, is the set of

$$M\left(S_1\right) = \left\{b \in B,\ \text{ there exists an } a \in S_1 \text{ such that } \ (a, b) \in M\right\}.$$

Similarly, we can define $M^{-1}\left(S_2\right)$ for a set $S_2 \subset \mathrm{rang}\, M$, the detail is omitted.

A relation f is said to be a mapping from A to B, if for every $a \in \mathrm{dom}\, f, f(a)$ has only one element. But for $b \in \mathrm{rang}\, f, f^{-1}(b)$ may hold more than one elements although f is a mapping.

When the domain and range of a mapping are subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, then the mapping is called by function. When only the range of a mapping is a subset of $\mathbb{R}^m$, the mapping is then called by functional. If $\mathbb{R}$ is extended to $\mathscr{R}$, then two functions $f : A \to \mathscr{R}$ and $f : A \to \mathbb{R}$ may have different domains. Thus, we define respectively

$$\mathrm{Dom}\, f = \{x; x \in \mathscr{R} \ \text{ such that } f(x) \in \mathscr{R}\}$$

and

$$\mathrm{dom}\, f = \{x; x \in \mathbb{R} \ \text{ such that } f(x) \in \mathbb{R}\}.$$

Obviously, they are different. $\mathrm{dom}\, f$ is called the effective domain of f. When $\mathbb{R}$ is extended to $\mathscr{R}$, the effective domain is invariant for a function. For example, for the function $y = \ln\ x$, we have $\mathrm{Dom}\ f = [0, \infty]$ and $\mathrm{dom}\ f = (0, \infty)$. And for the function $y = \left|x^{-1}\right|$, if it is considered on $\mathscr{R}$, its domain is $\mathscr{R}$, and moreover it is a continuous function.

1.1.2 Normed Spaces

Let A be a set. If there exists a functional $d : A \times A \to \mathbb{R}\ (\geq 0)$ such that

(1) $d\,(x, y) = 0$, if and only if $x = y$;
(2) $d\,(x, y) = d\,(y, x)$ for all $x, y \in A$;
(3) $d\,(x, y) \leq d\,(x, z) + d\,(z, y)$ for all $x, y, z \in A$.

Then d is a distance on A. (A, d) is called by a metric space or distance space.

Usually, the inequality in Condition (3) is called by triangular inequality.

Let X be a vector space on $\mathbb{R}$, $\rho : X \to \mathbb{R}\ (\geq 0)$ be a functional. If ρ satisfies the following conditions, then ρ is called by a norm on X, and (X, ρ) is a normed space.

(1) $\rho(x) = 0$ if and only if $x = 0$;
(2) $\rho(ax) = |a|\, \rho(x)$, for every $a \in \mathbb{R}$ where $|a|$ is the absolute value of a;
(3) $\rho\,(x + y) \leq \rho(x) + \rho(y)$ for all $x, y \in X$.

The equation in Condition (2) is called positive homogeneity.

An element of a normed space X is called a vector or a point for simplicity.

For sake of convenience, the norm $\rho(x)$ is denoted by $\|x\|$. Moreover, it can be verified that for $x, y \in X$, $\|x-y\|$ satisfies all conditions of distance; hence, a vector space has to be a metric space.

Now we present several normed spaces which will be used frequently in this book.

$\mathbb{R}^n$ is the n-dimensional real space. On $\mathbb{R}^n$, there are lots norms. For example, let $x = [x_1 x_2 \ldots x_n]^T \in \mathbb{R}^n$, where the superscript "T" is denoted the transpose. Here a vector in $\mathbb{R}^n$ is always taken the form of column vector. For $\alpha \geq 1$, $\|x\|_\alpha = \left(|x_1|^\alpha + |x_2|^\alpha + \cdots + |x_n|^\alpha\right)^{\frac{1}{\alpha}}$ is a norm of $\mathbb{R}^n$. The proof that $\|\cdot\|_\alpha$ satisfies the three conditions of a norm can be found in every textbook of functional analysis, and is omitted. $\|\cdot\|_\alpha$ is called by α-norm of $\mathbb{R}^n$. $\|x\|_\infty = \max\limits_i \left(|x_i|, i = 1, 2, \ldots, n\right)$ can be verified to be another norm of $\mathbb{R}^n$. $\|x\|_\infty$ is called infinite norm of $\mathbb{R}^n$. We also point out that when $\alpha \to \infty$, $\|x\|_\alpha \to \|x\|_\infty$ for a given $x \in \mathbb{R}^n$. It is the reason why we call $\|x\|_\infty$ to be the infinite norm. The frequently used norm is $\|x\|_2$, which is called by Euclidian norm. For convenience, the subscript 2 is often omitted and denote $\|x\|$ for $\|x\|_2$.

The above discussion for $\mathbb{R}^n$ illustrates that a vector space can be equipped with different norms. In general, different norms may lead to different topological properties. But for a space with finite dimension, the topological properties are almost similar. We now give the following definition.

Definition 1.1.1 Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two norms of a normed space X. $\|\cdot\|_a$ is topologically equivalent to $\|\cdot\|_b$, if there are two constants C_1, $C_2 \in \mathbb{R}^+$ such that for every $x \in X$, the inequalities $C_1\|x\|_a \leq \|x\|_b \leq C_2\|x\|_a$ hold. □

It is obvious that if $C_1\|x\|_a \leq \|x\|_b \leq C_2\|x\|_a$, then there are two constants D_1, $D_2 \in \mathbb{R}^+$ such that $D_1\|x\|_b \leq \|x\|_a \leq D_2\|x\|_b$. The fact supports that the topological equivalence of two norms is an equivalent relation (Problem 1 of this section).

We have the following theorem whose proof can be found in a textbook for functional analysis and is omitted.

Theorem 1.1.1 Two norms in a normed space with finite dimension are topologically equivalent. □

Theorem 1.1.1 illustrates that when we deal with topological properties such as convergence, continuity, and compactness for a normed space with finite dimension, we can select a norm arbitrary. Moreover, we can change the norm flexibly in the verification of questions.

Let $C([a,b], \mathbb{R})$[1] be the set of continuous functions defined on a closed interval $[a,b]$. The norm of $C([a,b], \mathbb{R})$ can be defined as $\|x(t)\| = \max\limits_{t\in[a,b]} |x(t)|$. There is no difficulty to prove that the definition satisfies three conditions of norm. Thus, $C([a,b], \mathbb{R})$ is a normed space. Moreover, let $C([a,b], \mathbb{R}^n)$ be the set of

[1] At the notation $C([a,b], \mathbb{R})$, C means continuous, $[a,b]$ is the domain and $\mathbb{R}$ is the range. $C([a,b], \mathbb{R}^n)$ has a similar meaning.

n-dimensional continuous functions defined on a closed interval $[a, b]$, i.e., for each $x(t) \in C([a, b], \mathbb{R}^n)$ and $t \in [a, b]$, $x(t) = [x_1(t)x_2(t)\cdots x_n(t)]^T \in \mathbb{R}^n$. It is easy to see that $C([a, b], \mathbb{R}^n) = (C([a, b], \mathbb{R}))^n$. Thus we can define a norm of $x(t)$ by

$$\|x(t)\|_\alpha = \left(||x_1(t)||^\alpha + ||x_2(t)||^\alpha + \cdots + ||x_n(t)||^\alpha\right)^{\frac{1}{\alpha}}$$

with $\alpha \geq 1$, where $\|x_i(t)\|$ is the norm of $x_i(t)$ on $C([a, b], \mathbb{R})$. We can prove that the definition meets the requirements of norm. $\|x(t)\|_\alpha$ is called as the α-norm of $x(t)$. When $\alpha = 2$, we call $\|x(t)\|_2$ is the Euclidian norm of $x(t)$ and is simplified as $\|x(t)\|$. The notation of $\|x(t)\|$ may lead to ambiguity. By the definition of norm, $\|x(t)\|$ is a real number. But if we treat $x(t)$ as a time-varying vector, then for every fixed t_0, $x(t_0) \in \mathbb{R}^n$, we can define its Euclidian norm $\|x(t_0)\| = \sqrt{x_1^2(t_0) + x_2^2(t_0) + \cdots + x_n^2(t_0)}$. When t varies in $[a, b]$, then $\|x(t)\|$ is a function of t. To distinguish from the norm of $C([a, b], \mathbb{R}^n)$. Hence, we use $\|x(t)\|_{\mathbb{R}}$ to denote its Euclidian function norm, i.e., $\|x(t)\|_{\mathbb{R}} = \sqrt{x_1^2(t) + x_2^2(t) + \cdots + x_n^2(t)}$ which is a function of t.

The above procedure provides a normal method to extend a norm from scalar case to its vector mode with finite dimension.

Let $L_1([a, b], \mathbb{R})$ be the set of functions which are absolutely Lebesgue integrable on the interval $[a, b]$, i.e., if $x(t) \in L_1([a, b], \mathbb{R})$, then $\int\limits_a^b |x(t)|\, dt < \infty$.[2] Then $\|x(t)\|_1 = \int\limits_a^b |x(t)|\, dt$ is qualified as a norm of $L_1([a, b], \mathbb{R})$.

Similarly, for $p \geq 1$, we apply $L_p([a, b], \mathbb{R})$ to express the set of functions which are absolutely p-Lebesgue integrable on the interval $[a, b]$, i.e., if $x(t) \in L_p([a, b], \mathbb{R})$, then $\int\limits_a^b |x(t)|^p dt < \infty$. The norm of $x(t)$ is defined as $\|x(t)\|_p = \left(\int\limits_a^b |x(t)|^p dt\right)^{\frac{1}{p}}$. Thus $L_p([a, b], \mathbb{R})$ is a normed space.

Let $L_\infty([a, b], \mathbb{R})$ be the set of essential bounded functions on the interval $[a, b]$, i.e., if $x(t) \in L_\infty([a, b], \mathbb{R})$, then there exists a constant M such that $|x(t)| \leq M, \quad t \in [a, b] \backslash E$, where $E \subset [a, b]$ is a set whose measure is zero. For $x(t) \in L_\infty([a, b], \mathbb{R})$, the norm of $x(t)$ can be defined as

$$\|x(t)\|_\infty = \inf_{E \subset [a,b], m(E)=0} \Big(\sup_{t \in [a,b] \backslash E} |f(t)| \Big)$$

[2] The integration is always in the meaning of Lebesgue integration if we do not give an illustration.

where $m(E)$ is the Lebesgue measure of set E. By the definition of the norm, $L_\infty([a,b],\mathbb{R})$ is a normed space.

It can be verified that if $1 \leq \alpha \leq \beta$, then

$$L_1([a,b],\mathbb{R}) \supset L_\alpha([a,b],\mathbb{R}) \supset L_\beta([a,b],\mathbb{R}) \supset L_\infty([a,b],\mathbb{R}).$$

By the normal procedure of extending the norm of $C([a,b],\mathbb{R})$ to $C([a,b],\mathbb{R}^n)$, we can define the normed space $L_p([a,b],\mathbb{R}^n)$ for every $p \in [1,\infty]$, and the detailed statement is omitted.

$AC([a,b],\mathbb{R})$ is the set of all functions which are absolutely continuous on the interval $[a,b]$. The absolute continuity is an important concept which will be widely used, hence, we give a normal definition below.

Definition 1.1.2 Let $f : [a,b] \to \mathbb{R}$ be a function. f is said to be absolutely continuous on $[a, b]$, if for every $\varepsilon > 0$, there is a $\delta = \delta(\varepsilon) > 0$ such that for arbitrary non-overlapping intervals $(a_1,b_1), (a_2,b_2), \ldots, (a_n,b_n) \subset [a,b]$ which satisfy $\sum\limits_{i=1}^{n}(b_i - a_i) < \delta$, then

$$\sum_{i=1}^{n} |f(b_i) - f(a_i)| < \varepsilon. \qquad \square$$

When $n = 1$, the definition of absolute continuity leads to the uniform continuity. Hence, a function is absolutely continuous, then it is uniformly continuous, so continuous. The reversed statement is not true. A continuous function may not be absolutely continuous. The following conclusion is necessary in this book.

Theorem 1.1.2 $f : [a,b] \to \mathbb{R}$ is an absolutely continuous function if and only if it is derivable almost everywhere on $[a,b]$. Furthermore,

$$f(x) = f(a) + \int_a^x f'(t)dt \qquad \square$$

Theorem 1.1.2 implies that an absolutely continuous function can be decomposed into two parts, one is a constant and another is a function with variable on the upper limit of integration. Theorem 1.1.2 also suggests that the norm of $AC([a,b],\mathbb{R})$ can be defined as follows

$$\|f(x)\| = |f(a)| + \int_a^b \left|f'(t)\right| dt. \tag{1.1.1}$$

It is easy to see the definition of Eq. (1.1.1) meets the requirements of norm. From Eq. (1.1.1), one can conclude that $AC([a,b],\mathbb{R})$ is isomorphic to a subspace of

$\mathbb{R} \times L_1([a,b],\mathbb{R})$ (it is also to say that $AC([a,b],\mathbb{R})$ can be embedded into $\mathbb{R} \times L_1([a,b],\mathbb{R})$). The above investigation can be extended to $AC([a,b],\mathbb{R}^n)$ by the normal way used for $C([a,b],\mathbb{R})$, and the detailed processing is omitted.

1.1.3 Elementary Topology

In this subsection, we deal with elementary topology which is an important foundation for investigation of set-valued mappings and differential inclusions.

Let $\overline{B} = \{x; \|x\| \le 1\}$ be the closed unit ball in the normed space X, and $B = \{x; \|x\| < 1\}$ be the inner part of $\overline{B}$, sometimes, B is called by the open unit ball. $\overline{B} \backslash B = \{x; \|x\| = 1\}$ is denoted the sphere of $\overline{B}$ or B, i.e., the boundary of $\overline{B}$ or B.

Let $A \subset X$ be a set in the normed space X. $x \in A$ is called an inner point of A, if there exists an $\varepsilon > 0$ such that $B(x,\varepsilon) = x + \varepsilon B \subset A$, where $x + \varepsilon B$ is the set of $\{y; y = x + \varepsilon b, b \in B\}$ and $B(x,\varepsilon)$ is called the ε-neighborhood of x. In the following, we will use such notations frequently and without illustration. $\mathrm{int}A$ denotes the set of all inner points of A and called interior of A. $x \in X$ is an accumulation point or a cluster point of A, if in the set $B(x,\varepsilon) \cap A$ there exists a point which is different form x for arbitrarily $\varepsilon > 0$. Let A' denote the set of all accumulation points of A; sometimes, A' is called the derived set of A. $\mathrm{cl}\,A = A \cup A'$ is the closure of A. The boundary of A is the set of $\mathrm{cl}A/\mathrm{int}A$ which is denoted by $\mathrm{bd}A$. The set $\mathrm{bd}A$ holds such a property that if $x \in \mathrm{bd}A$, then for each neighborhood $B(x,\varepsilon)$, the joints $B(x,\varepsilon) \cap A$ and $B(x,\varepsilon) \cap A^c$ are nonempty sets, where A^c is the complement of A.

Let X be a normed space. A set $O \subset X$ is said to be open if each point in O is an inner point, i.e., $O = \mathrm{int}\,O$. A set $C \subset X$ is said to be closed if its complement C^c is an open set. By the definitions, it is obvious that $\mathrm{int}A$ is always an open set and $\mathrm{cl}\,A$ is always a closed set for every set A in X. It can be proved that A is closed if and only if $A \supset \mathrm{cl}\,A$, or equivalently, $A \supset A'$.

A set A is bounded if there is a constant $r \in \mathbb{R}\,(> 0)$ such that $A \subset r\overline{B}$.

Let X be a normed space. $x \in X$ is a point (vector) and $A \subset X$ is a set, $d(x,A) = \inf\limits_{a \in A} \|x - a\|$ is defined as the distance between x and A. If A is closed, then there is a $y \in A$ such that $\|x - y\| = d(x,y) = d(x,A)$. In general, the point y is not necessary to be unique. Let $A_1, A_2 \subset X$ be two sets. The distance of A_1 and A_2 is defined as $d(A_1,A_2) = \inf\limits_{x \in A_1, y \in A_2} \|x - y\|$. When A_1 and A_2 are all closed, there exist $x \in A_1, y \in A_2$ such that $\|x - y\| = d(x,y) = d(A_1,A_2)$.

Let $A \subset X$ be a set. If for every $a \in A$, there is an open set O_a such that $a \in O_a$, then the set of open sets $\{O_a\}$ is qualified as an open covering of A.

A set $A \subset X$ is said to be compact if for an open covering $\{O_a\}$, there exists a finite set $\{O_1, O_2, \ldots, O_n\} \subset \{O_a\}$ such that $O_1 \cup O_2 \cup \cdots \cup O_n \supset A$. Compactness is a very useful property for the study of mappings, hence there are lots researches to investigate the compactness in different normed spaces. It is well-known that a set in $\mathbb{R}^n$ is compact if and only if it is bounded and closed. In the space $C([a,b],\mathbb{R})$, a set is compact if it is bounded and uniformly continuous, the fact is known as Arzela-Ascoli theorem. Generally speaking, finding a character for the compact set in a normed space is a hard job. Fortunately, the compact sets considered in this book are involved mostly in $\mathbb{R}^n$ or $C([a,b],\mathbb{R})$.

$\{x_n; n = 1, 2, \dots\} \subset X$ is a sequence. The sequence $\{x_n; n = 1, 2, \dots\}$ is convergent to a point $x_0 \in X$, if for every $\varepsilon > 0$, there is an $N \in \mathbb{N}$, such that for each $n > N$, $\|x_n - x_0\| < \varepsilon$. It is obvious that $\|x_n - x_0\| < \varepsilon$ is equivalent to $x_n \in B(x_0, \varepsilon)$. The fact is denoted by $\lim\limits_{n\to\infty} x_n = x_0$, or $x_n \to x_0, (n \to \infty)$. For the sake of convenience, $\{x_n; n = 1, 2, \dots\}$ is often simplified as $\{x_n\}$ and $x_n \to x_0, (n \to \infty)$ as $x_n \to x_0$, whenever no confusion happens. If $x_n \to x_0$, then $\{x_n\}$ is a convergent sequence and x_0 is the limitation of $\{x_n\}$. By the definition of accumulation, x_0 is also an accumulation point of $\{x_n\}$. For a convergent sequence, its limitation or accumulation point is unique.

For the compactness in $\mathbb{R}^n$, we have the Weierstrass theorem: $A \subset \mathbb{R}^n$ is compact if for a sequence $\{x_n\} \subset A$, $\{x_n\}$ contains a convergent subsequence $\{x_{n_k}, k = 1, 2, \dots\}$; moreover, the limitation of $\{x_{n_k}, k = 1, 2, \dots\}$ belongs to A.

A sequence $\{x_n\} \subset X$ is said to be a Cauchy sequence if for every $\varepsilon > 0$, there is an $N \in \mathbb{N}$, such that $\|x_n - x_m\| < \varepsilon$ provided that $n, m > N$. Let $A \subset X$ be a set, A is complete if for every Cauchy sequence $\{x_n\}$ in A, $\{x_n\}$ has a limitation x_0, and $x_0 \in A$. If $A = X$, then the normed space is complete. A complete normed space is usually called Banach space.

These spaces $\mathbb{R}^n$, $C([a,b],\mathbb{R}^n)$, $L_p([a,b],\mathbb{R}^n)$, and $AC([a,b],\mathbb{R}^n)$ are all Banach spaces with their norms respectively pre-defined.

There is a wonderful theorem called completion that a normed space can be densely embedded into a Banach space, i.e., the normed space is isomorphic to a densely set of a Banach space and the isomorphism is equidistance. Hence when we talk with a normed space, we always assume that it is complete.

Let X and Y be two normed spaces, and $f : X \to Y$ be a mapping. f is continuous at $x_0 \in X$ if for every $\varepsilon > 0$, there is a $\delta > 0$, where the δ may depend on ε, such that if $x \in B(x_0, \delta)$, then $\|f(x) - f(x_0)\| < \varepsilon$. Equivalently, $f(B(x_0, \delta)) \subset B(f(x_0), \varepsilon)$. Moreover, f is continuous on X if f is continuous at every $x \in X$. Such a property is called by continuity.

There are many criteria for the continuity of a mapping. For example, if $f(x_n) \to f(x_0)$ for every sequence $\{x_n\}$ with $x_n \to x_0$, then f is continuous at $x_0 \in X$. The conclusion is known as Heine theorem. It is meaningful that the two arrows may mean two different norms. Another conclusion we will use is that if for every open set $O \subset Y$, $f^{-1}(O)$ is an open set in X, then f is continuous on X. Section 1.4 will deal with the semi-continuity of a mapping. The continuity is a base for further discussion there. Hence, readers should be familiar with the properties of continuous mappings.

Let X and Y be two normed spaces. $f : X \to Y$ is said to be a linear mapping, if f satisfies the following conditions:

(1) For two vectors $x_1, x_2 \in X$, the equation $f(x_1 + x_2) = f(x_1) + f(x_2)$ holds
(2) For $x \in X$ and $a \in \mathbb{R}$, then $f(ax) = af(x)$

The first condition is usually called as additivity, and the second one is called homogeneity since $f(a^n x) = a^n f(x)$ for an integer n.

A mapping $f : X \to Y$ is said to be bounded if there is a positive constant M, such that $\|f(x)\| \leq M\|x\|$ is true for every $x \in X$. If $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, then a linear mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ is always bounded. But for the general case, the conclusion does not hold. A well-known conclusion is that a linear mapping $f : X \to Y$ is bounded if and only if it is continuous.

For a linear mapping $f : X \to Y$, we can define its operator norm as follows

$$\|f\| = \sup_{x\in X, x\neq 0} \frac{\|f(x)\|}{\|x\|}. \tag{1.1.2}$$

It is direct to verify that the above definition meets with the all requirements of a norm. And, it is also clear that f is bounded if and only if $\|f\| < \infty$ since Eq. (1.1.2) implies $\|f(x)\| \leq \|f\|\,\|x\|$ for every $x \in X$. Because f is linear, we can prove that $\|f\|$ is also equal to

$$\|f\| = \sup_{\|x\|=1} \|f(x)\| = \sup_{\|x\|\leq 1} \|f(x)\| = \sup_{\|x\|\leq 1, x\neq 0} \frac{\|f(x)\|}{\|x\|}.$$

Obviously, different definitions of $\|f(x)\|$ and $\|x\|$ may lead to different value of $\|f\|$. Consequently, sometimes, Eq. (1.1.2) is rewritten by

$$\|f\|_{\beta\alpha} = \sup_{x\in X, x\neq 0} \frac{\|f(x)\|_\beta}{\|x\|_\alpha}$$

to emphasize the dependency, where $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are α-norm of space X and β-norm of space Y, respectively. $\|f\|_{\beta\alpha}$ is then called by reduced operator norm. If there is another $\|f\|_\gamma$ which is not a reduced operator norm form α-norm and β-norm, but it also satisfies $\|f(x)\|_\beta \leq \|f\|_\gamma \|x\|_\alpha$. Then the $\|f\|_\gamma$ is called a compatible norm with $\|x\|_\alpha$ and $\|y\|_\beta$.

By the operator norm, all linear bounded mappings from X to Y form a normed linear space which is denoted by $L(X \to Y)$. Especially, if $Y = \mathbb{R}$, the space $L(X \to \mathbb{R})$ is simplified as X^* and called by conjugate space of X. No matter whether X is complete, the space X^* is always complete.

1.1.4 Limitation Theorems

In this subsection, we introduce several limitation theorems. It is known from the theory of real functions that the introduction of Lebesgue integration can make the Newton-Leibniz formula is true for all integrable functions; moreover, it provides much convenience for the application of limitation theorems. Let us start with definitions of convergence.

Let $x_n(t),\ n = 1, 2, \ldots$ be measurable functions defined on a measurable set $E \subset \mathbb{R}$, and let $\{x_n(t)\}$ denote the sequence of $x_n(t)$. The sequence $\{x_n(t)\}$ is convergent to $x_0(t)$ with respect to the measure $m(\cdot)$ (in this book we only consider the Lebesgue measure), if for every $\varepsilon > 0$, the limitation $\lim\limits_{n\to\infty} m\{t; ||x_n(t) - x_0(t)|| > \varepsilon\} = 0$ holds. The fact is denoted by $x_n(t) \overset{m}{\to} x_0(t)$ for simplification.

The sequence $\{x_n(t)\}$ is convergent to $x_0(t)$ almost everywhere if $m\Big(E\backslash\Big\{t; \lim\limits_{n\to\infty} x_n(t) = x_0(t)\Big\}\Big) = 0$, i.e., the measure of the set on which $x_n(t)$ is not convergent to $x_0(t)$ is zero. The fact is denoted as $x_n(t) \overset{a.e.}{\to} x_0(t)$.

The sequence $\{x_n(t)\}$ is convergent to $x_0(t)$ at every point in its domain if the limitation $\lim\limits_{n\to\infty} x_n(t) = x_0(t)$ is valid for every $t \in E$ where E is the common domain of $x_n(t),\ n = 1, 2, \ldots$ and $x_0(t)$. The fact is denoted as $x_n(t) \to x_0(t)$.

The sequence $\{x_n(t)\}$ is uniformly convergent to $x_0(t)$ if $\lim\limits_{n\to\infty} \sup\limits_{t\in E} ||x_n(t) - x_0(t)|| = 0$. The fact is denoted as $x_n(t) \overset{u}{\to} x_0(t)$.

From the above definitions, one can obtain directly that

$$x_n(t) \overset{u}{\to} x_0(t) \Rightarrow x_n(t) \to x_0(t) \Rightarrow x_n(t) \overset{a.e.}{\to} x_0(t) \Rightarrow x_n(t) \overset{m}{\to} x_0(t),$$

where the last implication requires that $m(E) < \infty$, i.e., the common domain holds a finite measure. The following theorem is known as Riesz theorem which will be needed in further discussion.

Theorem 1.1.3 (Riesz theorem) If $x_n(t) \overset{m}{\to} x_0(t)$, then there is a subsequence $\{x_{n_k}(t)\}$ of $\{x_n(t)\}$ such that $x_{n_k}(t) \overset{a.e.}{\to} x_0(t)$. □

The theorem gives an opposite result for the conclusion that $x_n(t) \overset{a.e.}{\to} x_0(t) \Rightarrow x_n(t) \overset{m}{\to} x_0(t)$.

It is meaningful to note that all of these conclusions are still valid if $\mathbb{R}$ is replaced by $\mathbb{R}^n$, and the norms in $\mathbb{R}^n$ can be defined arbitrarily.

Recall those spaces mentioned before. In $C([a, b], \ \mathbb{R}^n)$, the convergence with respect to the norm is equivalent to the uniform convergence. In $L_p([a, b], \mathbb{R}^n)$, the convergence with respect to the norm is equivalent to convergence with respect to measure. Because $AC([a, b], \mathbb{R}^n)$ can be embedded into $\mathbb{R}^n \times L_1([a, b], \mathbb{R}^n)$, its convergence is equivalent to that the initial values $\{x_n(a)\}$ is convergent and $\{x_n(t)\}$ is convergence with respect to its measure on the interval $[a, b]$.

We now turn to the limitation theorems.

Theorem 1.1.4 Let $\{x_n(t)\}$ be a sequence of measurable functions defined on a measurable set E, and $x_n(t) \overset{m}{\to} x_0(t)$; Let $F(t)$ be an integrable function on E, such that $|x_n(t)| \le F(t)$ for $t \in E$ and $n \in \mathbb{N}$, then $x_0(t)$ is integrable on E and

$$\lim_{n\to\infty} \int_E x_n(t)dt = \int_E x_0(t)dt. \qquad \square$$

Theorem 1.1.4 is known as control convergence theorem presented by Lebesgue. $F(t)$ is a control function.

Theorem 1.1.5 Let $\{x_n(t)\}$ be a sequence of nonnegative functions on E, and $x_n(t) \le x_{n+1}(t)$ for each n and $t \in E$, then

$$\lim_{n\to\infty}\int_E x_n(t)dt = \int_E \lim_{n\to\infty} x_n(t)dt. \qquad \square$$

Theorem 1.1.5 is known as monotonic convergence theorem established by Levi. If $x_n(t) \to \infty$ for some $t \in E$, then Theorem 1.1.5 is still true.

The next theorem is called Fatou lemma.

Theorem 1.1.6 (Fatou lemma) Let $\{x_n(t)\}$ is a sequence of nonnegative functions defined on a measurable set E, then

$$\underline{\lim}_{n\to\infty}\int_E x_n(t)dt \ge \int_E \underline{\lim}_{n\to\infty} x_n(t)dt. \qquad \square$$

It is worth to note that there is a sequence such that the inequality really holds.

To end this subsection, we deal with some concepts and conclusions about "weakness".

Let X be a normed space. A set $A \subset X$ is said to be bounded if there is a constant M, such that for every $x \in A$, $\|x\| \le M$. The set A is said to be weakly bounded if for every $f \in X^*$, there is a constant $M = M(f)$, i.e., M may depend on the functional f, such that $|f(x)| \le M$ for each $x \in A$. It is obvious that a set A is bounded then it is weakly bounded. the opposite statement may not be true. Let A^* be a set in X^*, i.e., $A^* \subset X^*$. A^* is said to be weakly $*$-bounded if for every $x \in X$, there is a constant $M = M(x)$ such that $|f(x)| \le M$ for each $f \in A^*$.

X is a normed space. A set $A \subset X$ is said to be compact if any open covering can reduce to a finite open covering. The set A is said to be weakly compact if for every $f \in X^*$, the set $f(A) = \{f(x); x \in A\}$ is compact. Let A^* be a set in X^*, i.e., $A^* \subset X^*$, A^* is said to be weakly $*$-compact if for every $x \in X$, the set $\{f(x); f \in A^*\}$ is a compact set of $\mathbb{R}$.

The following theorem was devoted by Alaoglu, its proof is referred to de Bruim et al. (2009).

Theorem 1.1.7 (Alaoglu theorem) Let X be a Banach space. Then the closed unit ball in X^* is weakly $*$-compact. □

By Theorem 1.1.7, every closed bounded set in X^* is weakly $*$-compact. It is meaningful to comparing with the finite dimensional spaces where a closed bounded set is compact, but in infinite dimensional spaces it is only weakly $*$-compact.

1.1.5 Inner Product Spaces

Inner product will be the major operation in the following investigation. Hence, we spend some space to deal with the inner product space. In this section, we only give preliminary conclusions, and others will be given as they are needed.

Let X be a linear space. $\rho : X \times X \to \mathbb{R}$[3] is a functional. If ρ satisfies the following considerations.

(1) $\rho(x, x) \geq 0$ for every $x \in X$, and $\rho(x, x) = 0$ if and only if $x = 0$
(2) $\rho(x, y) = \rho(y, x)$, for $x, y \in X$
(3) $\rho(ax + by, z) = a\rho(x, z) + b\rho(y, z)$, for $x, y, z \in X$ and $a, b \in \mathbb{R}$

Then the ρ is called as an inner product on X, and X is called as an inner product space.

The inner product $\rho(x, y)$ is usually denoted by $\langle x, y \rangle$ for simplicity. A complete inner product space is called by Hilbert space.

It is easy to verify that $\sqrt{\langle x, x \rangle}$ is qualified as a norm on X. The norm is then called the reduced norm obtained from inner product. Therefore, a Hilbert space is always a Banach space. The opposite statement may fail. Let $\| \cdot \|$ be a reduced norm. Then the following equation holds

$$\|x - y\|^2 + \|x + y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right). \tag{1.1.3}$$

Equation (1.1.3) is usually called parallelogram law. If a norm in a normed space satisfies the parallelogram law, then the norm can reduce an inner product

$$\langle x, y \rangle = \frac{1}{4}\left(\|x + y\|^2 - \|x - y\|^2\right).$$

We have mentioned $\mathbb{R}^n$, $C([a, b], \mathbb{R}^n)$, $L_p([a, b], \mathbb{R}^n)$ and $AC([a, b], \mathbb{R}^n)$ are all Banach spaces. Among these spaces, $(\mathbb{R}^n, \|x\|_2)$ and $L_2([a, b], \mathbb{R}^n)$ are only Hilbert spaces, and their inner products are

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i, \quad x, y \in \mathbb{R}^n,$$

and

$$\langle x(t), y(t) \rangle = \int_a^b x^T(t) y(t) dt = \int_a^b \sum_{i=1}^{n} x_i(t) y_i(t) \mathrm{d}t, \quad x(t), y(t) \in L_2([a, b], \mathbb{R}^n),$$

respectively.

[3] The inner product can be defined as $X \times X \to \mathbb{C}$. But in this book, we mainly consider the inner product in $\mathbb{R}$.

Schwarz inequality is very fundamental in inner product space, it is

$$\langle x, y\rangle \leq \|x\| \, \|y\| , \tag{1.1.4}$$

where $\|\cdot\|$ is the reduced norm. By Inequality (1.1.4), we can define an angle included by x and y as follows

$$\cos\theta = \frac{\langle x, y\rangle}{\|x\| \, \|y\|}$$

where the value of θ is restricted in the interval $[0, \pi)$.

If $\theta = \pi/2$, then $\langle x, y\rangle = 0$. We say that x is orthogonal to y and denote the fact by $x \perp y$. If x has a unit length, i.e., $\|x\| = 1$, then $\langle x, y\rangle = \|y\| \cos\theta$. It is the projection of y at the line where x is located. This fact is frequently cited later. The following theorem is known as Reisz theorem.

Theorem 1.1.8 (Reisz theorem) Let X be a Hilbert space. Then for every $f \in X^*$, there is a $y \in X$ such that $f(x) = \langle x, y\rangle$. Moreover, $\|f\| = \|y\|$. □

The theorem illustrates every linear bounded functional in a Hilbert space is only the inner product.

Let $M \subset X$ be a set in X. Then all vectors which are orthogonal to M form a subspace and denoted by $M^\perp$. $M^\perp$ is called by the orthogonal complement of M. No matter whether M is closed, $M^\perp$ is always a closed subspace. Furthermore, if M is a subspace then M and $M^\perp$ form an orthogonal decomposition of X, i.e., $M \cap M^\perp = \{0\}$, and $M + M^\perp = X$.

M is a subspace of X, and $y \notin M$. If there is a $x_0 \in M$ such that $(y - x_0) \perp M$, then the x_0 is called by the projection of y on M. The x_0 holds property that

$$\|y - x_0\| = d(y, x_0) = d(y, M) = \inf_{x \in M} d(y, x) = \inf_{x \in M} \|y - x\| .$$

Usually, we cannot guarantee the existence of x_0. But when X is complete and M is closed there is a unique x_0 to meet the requirement. In the next section, we shall extend the property to convex sets by using the above equations.

Problems

1. Let M be a relation on $A \times A$. M is called as an equivalent relation if M satisfies:

   (1) For every $a \in A$, $(a, a) \in M$
   (2) If $(a, b) \in M$, then $(b, a) \in M$
   (3) If $(a, b) \in M$ and $(b, c) \in M$, then $(a, c) \in M$

   Prove that all equivalent norms in a normed space form an equivalent relation.

2. Let $O_\alpha, \alpha \in I$ be open sets where I is a set of indices which may be infinite and even uncountable. Then $\underset{\alpha}{\cup} O_\alpha$ is an open set. Give an example to show that $\underset{\alpha}{\cap} O_\alpha$ fails to be open. But if I is a finite set, then $\underset{\alpha}{\cap} O_\alpha$ is always open.

3. Let $C_\alpha, \alpha \in I$ be closed sets where I is a set of indices which may be infinite and even uncountable. Then $\underset{\alpha}{\cap} C_\alpha$ is a closed set. Give an example to show that $\underset{\alpha}{\cup} C_\alpha$ may fail to be closed. But if I is a finite set, then $\underset{\alpha}{\cup} C_\alpha$ is closed.
4. Let A_1 and A_2 be two open sets. Is their Cartesian production $A_1 \times A_2$ an open set? Is their linear sum $A_1 + A_2$ an open set? For two closed sets A_1 and A_2, discuss similar problems.
5. Let A_1 and A_2 be two compact sets. Is their Cartesian production $A_1 \times A_2$ a compact set? Is their linear sum $A_1 + A_2$ a compact set?
6. Let bd$A$ denote the boundary of set A. Prove the following conclusions:

   (1) $x \in \mathrm{bd}A$ if and only if every neighborhood $B(x, \varepsilon)$ contains vectors in A and A^c, respectively.

   (2) $\mathrm{bd}A = \mathrm{cl}A \cap \mathrm{cl}A^c$.

7. Let I be a closed interval in $\mathbb{R}$, and for each $k \in \mathbb{N}$, $x_k(t)$: $I \to \mathbb{R}^n$ be an absolutely continuous function. If for every $t \in I$, the set $\{x_k(t)\}$ is a compact set in $\mathbb{R}^n$. Moreover, there is a function $c(t) : I \to \mathbb{R}$ such that $\left\|\dot{x}_k(t)\right\| \le c(t)$. Then there is a subsequence $\{x_{k_n}(t)\}$ of $\{x_k(t)\}$, $\{x_{k_n}(t)\}$ satisfies the following properties:

   (1) There is an absolutely continuous function $x(t)$: $I \to \mathbb{R}^n$, such that $\{x_{k_n}(t)\}$ is convergent uniformly to $x(t)$ on the interval I.

   (2) $\left\{\dot{x}_{k_n}(t)\right\}$ is convergent weakly to $\dot{x}(t)$.

   (Hint: to apply Alaoglu theorem and Arzela-Ascoli theorem)
8. Let X be a Hilbert space, and M be a closed subspace of X. Then for every $y \in X$, there is a unique $y_M \in M$ such that $(y - y_M) \perp y_M$. We can then define a mapping $P_M : X \to M$ such that $P_M(y) = y_M$. Verify that the P_M is a linear mapping and $\|P_M\| = 1$.
9. Prove that every subspace of a normed space with finite dimension is closed.
10. A matrix $A \in \mathbb{R}^{m \times n}$ can be treated as a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$.

    (1) Prove that for every matrix norm $\|A\|$ there are two norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, such that $\|A\|$ is compatible to $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$

    (2) If $n = m$, find the reduced norm of A, if $\|\cdot\|_\alpha = \|\cdot\|_\beta = \|\cdot\|_2$

    (3) If $n = m$ and $\|\cdot\|_\alpha = \|\cdot\|_\beta = \|\cdot\|_2$, then $\|A\|_2 = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}$ is a compatible norm. $\|A\|_2$ is called Euclidean norm of matrix A.

1.2 Convex Sets

The convex analysis is a special subject in the theory of functional analysis and foundation of both optimization and differential inclusions. The theory of convex analysis contains mainly two contents. One is the convex set which provides a stage and another is the convex function which is the role playing on the stage. This section introduces basic properties of convex sets, and the next section turns to convex functions. These materials are used frequently in the further investigation of this book.

1.2.1 Convex Sets and Their Properties

X is always to express a normed space in this subsection. If $x_1, x_2 \in X$ and $x_1 \neq x_2$, then equation $x = x_1 + \lambda\,(x_2 - x_1)$, $\lambda \in \mathbb{R}$, is a line passed x_1 and x_2.[4] The two points x_1 and x_2 divide the line into three parts. If $\lambda \in [0, 1]$, then the equation $x = x_1 + \lambda\,(x_2 - x_1)$ expresses a line segment which starts at x_1 and ends at x_2. The segment is called the segment connecting x_1 and x_2 and denoted by $[x_1, x_2]$. Similarly, we can define inductively $(x_1, x_2]$, (x_1, x_2), and so on. It is worth noting that the signal $[x_1, x_2]$ expresses an interval, but does not imply $x_1 < x_2$. When $\lambda \in [1, \infty)$, the equation $x = x_1 + \lambda\,(x_2 - x_1)$ stands for a radial line starting at x_2 and radiating to infinite. The radial line is denoted by $[x_2, \infty)$. We can define similarly $(-\infty, x_1]$ for the radial line starting at x_1 and ending in infinite with opposite direction of $[x_2, \infty)$. For every $\lambda \in \mathbb{R}$, $x_\lambda = x_1 + \lambda\,(x_2 - x_1) = (1 - \lambda)\,x_1 + \lambda x_2$ is denoted for a point in the line for the specialized λ.

Definition 1.2.1 $A \subset X$ is a convex set in X, if for any two points $x_1, x_2 \in A$, the segment $[x_1, x_2] \subset A$. □

An alternative statement of $[x_1, x_2] \subset A$ is that $x_\lambda = (1 - \lambda)\,x_1 + \lambda x_2 \in A$ for every $\lambda \in [0, 1]$.

By the definition, the closed unit ball $\overline{B}$ of a normed space X is convex. Because if $x_1, x_2 \in \overline{B}$, then $\|x_i\| \le 1$, $i = 1, 2$. For $\lambda \in [0, 1]$, $x_\lambda = (1 - \lambda)\,x_1 + \lambda x_2$, we have

$$\|x_\lambda\| = \|(1 - \lambda)\,x_1 + \lambda x_2\| \le \|(1 - \lambda)\,x_1\| + \|\lambda x_2\| = (1 - \lambda)\,\|x_1\| + \lambda\,\|x_2\| \le 1.$$

Hence $x_\lambda \in \overline{B}$, i.e., $[x_1, x_2] \subset \overline{B}$.

By a similar way, it can be verified that the open unit ball B is also a convex set.

Preliminary properties of convex sets are listed as follows.

(1) The empty $\varnothing$ is convex and X is also convex.
(2) A subspace in X is convex.
(3) The intersection $\underset{i}{\cap} A_i$ is convex, if $A_i \subset X$, $\quad i \in I$ are all convex, where I is an index set;
(4) $x \in X$ is a vector, then $x + A$ is convex if A is a convex set;
(5) For $\alpha \in \mathbb{R}$, αA is convex if A is a convex set.

[4] If $x_1 = x_2$, $x = x_1 + \lambda\,(x_2 - x_1) \equiv x_1$. The line degenerates to a point. In this book we do not consider the special case.

(6) $A_1 \times A_2$ is a convex set if both A_1 and A_2 are convex.
(7) $A_1 + A_2$ is a convex set if both A_1 and A_2 are convex.

It is valuable to point out that the properties (6) and (7) can be extended to the case where $A_1, A_2, \ldots, A_n$ are n sets, but we are not sure it is true for infinite sets, since a vector in $A_1 \times A_2 \times \cdots$ or $A_1 + A_2 + \cdots$ may fail to be normed.

In general, the union of A_1 and A_2, $A_1 \cup A_2$ is not to be a convex set.[5]

Theorem 1.2.1 If A is a convex set, then $\operatorname{int} A$ and $\operatorname{cl} A$ are all convex sets.

Proof If $\operatorname{int} A = \varnothing$, then by the property (1) listed above, the $\operatorname{int} A$ is convex. We now turn to the case that $\operatorname{int} A \neq \varnothing$. If $x_1, x_2 \in \operatorname{int} A$, and $x \in [x_1, x_2]$, then there is a $\lambda \in [0,\ 1]$ such that $x = (1-\lambda)x_1 + \lambda x_2$. Define a subset of interval [0,1] as follows,

$$\Lambda = \{\lambda; \forall \sigma \in [0, \lambda],\ (1-\sigma)x_1 + \sigma x_2 \in \operatorname{int} A\}.$$

$\Lambda \neq \varnothing$ since $x_1 \in \operatorname{int} A$, then $\delta \in \Lambda$ for a small positive δ since $\operatorname{int} A$ is open. Denote $\lambda_0 = \sup \Lambda$, then $\lambda_0 > 0$. If $\lambda_0 = 1$, then $[x_1, x_2] \subset A$. The first conclusion of theorem is verified. We now prove the fact by contradiction.

If $\lambda_0 < 1$, denote $x_{\lambda_0} = (1-\lambda_0)x_1 + \lambda_0 x_2$. Because $x_2 \in \operatorname{int} A$, there is an $\varepsilon > 0$ such that $B(x_2, \varepsilon) \subset \operatorname{int} A$. We now show that $B(x_{\lambda_0}, \lambda_0\varepsilon) \subset \operatorname{int} A$. Suppose $x_3 \in B(x_{\lambda_0}, \lambda_0\varepsilon)$, i.e., $\|x_3 - x_{\lambda_0}\| < \lambda_0\varepsilon$. Now consider

$$x_4 = x_1 + \frac{x_3 - x_1}{\lambda_0}, \tag{1.2.1}$$

we have

$$\|x_4 - x_2\| = \left\| x_1 + \frac{x_3 - x_1}{\lambda_0} - x_2 \right\| = \frac{1}{\lambda_0}\|x_3 - [(1-\lambda_0)x_1 + \lambda_0 x_2]\| = \frac{1}{\lambda_0}\|x_3 - x_{\lambda_0}\| < \varepsilon.$$

It follows that $x_4 \in B(x_2, \varepsilon) \subset \operatorname{int} A$. By Eq. (1.2.1), we obtain $x_3 = (1-\lambda_0)x_1 + \lambda_0 x_4$. It implies that $x_3 \in [x_1, x_4] \subset A$ since A is convex. Because x_3 is selected arbitrarily in $B(x_{\lambda_0}, \lambda_0\varepsilon)$, x_{λ_0} is an inner point, and $B(x_{\lambda_0}, \lambda_0\varepsilon) \subset \operatorname{int} A$. Thus $\lambda_0 + \lambda_0\varepsilon \in \Lambda$, it then conflicts with the fact that $\lambda_0 = \sup \Lambda$. Therefore, $\sup \Lambda = 1$.

We now turn to cl$A$. Let $x_1, x_2 \in \operatorname{cl} A$. Then there are two sequences $\{x_n^1\} \subset A$ and $\{x_n^2\} \subset A$[6] such that $x_n^1 \to x_1$ and $x_n^2 \to x_2$. Thus, $(1-\lambda)x_n^1 + \lambda x_n^2 \to (1-\lambda)x_1 + \lambda x_2$ for every $\lambda \in [0, 1]$. Because A is convex, $(1-\lambda)x_n^1 + \lambda x_n^2 \in A$. It follows $(1-\lambda)x_1 + \lambda x_2 \in \operatorname{cl} A$, i.e., $[x_1, x_2] \subset \operatorname{cl} A$. □

The following two theorems extend the results established in Theorem 1.2.1.

[5] The reader should divide the notation $A_1 \cup A_2$ from $A_1 + A_2$.

[6] It is possible that $\{x_n^1\}$ and $\{x_n^2\}$ are constant sequences.

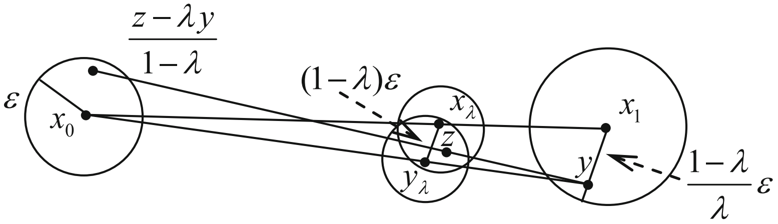


Fig. 1.1 Illustration of the proof of Theorem 1.2.2

Theorem 1.2.2 If A is a convex set, $x_0 \in \operatorname{int} A$ and $x_1 \in \operatorname{cl}A$, then $[x_0, x_1) \subset \operatorname{int} A$.

The proof of Theorem 1.2.2 is similar to that of Theorem 1.2.1, but is more precise.

Proof If $x_1 \in \operatorname{int} A$, then by Theorem 1.2.1int$A$ is also convex, and then $[x_0, x_1) \subset [x_0, x_1] \subset \operatorname{int} A$. Hence, it is sufficient to consider the case that $x_1 \notin \operatorname{int} A$, i.e., x_1 is on the boundary.

$x_0 \in \operatorname{int} A$, hence there is an $\varepsilon > 0$ such that $B(x_0, \varepsilon) \subset A$. To verify the theorem, it is sufficient to prove that $x_\lambda = (1-\lambda)x_0 + \lambda x_1 \in [x_0, x_1) \in \operatorname{int} A$ for every $\lambda \in (0, 1)$. Because $x_1 \in \operatorname{bd}A$, there is a $y \in B\left(x_1, \frac{1-\lambda}{\lambda}\varepsilon\right) \cap A$. A is convex, the segment $[x_0, y] \subset A$.

Let $y_\lambda = (1-\lambda)x_0 + \lambda y \in [x_0, y]$. Then $\|x_\lambda - y_\lambda\| = \lambda\|x_1 - y\| < (1-\lambda)\varepsilon$. For every $z \in B(y_\lambda, (1-\lambda)\varepsilon)$, we have

$$(1-\lambda)\varepsilon > \|z - y_\lambda\| = \|z - (1-\lambda)x_0 - \lambda y\| = (1-\lambda)\left\|\frac{z-\lambda y}{1-\lambda} - x_0\right\|. \quad (1.2.2)$$

Inequality (1.2.2) implies $\frac{z-\lambda y}{1-\lambda} \in B(x_0, \varepsilon)$. Furthermore, $z = (1-\lambda)\frac{z-\lambda y}{1-\lambda} + \lambda y \in A$. Because z is selected arbitrarily form $B(y_\lambda, (1-\lambda)\varepsilon)$, $B(y_\lambda, (1-\lambda)\varepsilon) \subset A$. We have verified that $\|x_\lambda - y_\lambda\| = \lambda\|x_1 - y\| < (1-\lambda)\varepsilon$, i.e., $x_\lambda \in B(y_\lambda, (1-\lambda)\varepsilon)$. Consequently, $x_\lambda \in \operatorname{int}A$ (Fig. 1.1). □

Theorem 1.2.2 presents an important feature of the convex set. A segment which connects an inner point and a point on the boundary is located in the interior except the endpoint of boundary. The following theorem can be verified by a similar way, hence we leave it to readers as an exercise.

Theorem 1.2.3 Suppose A is a convex set, and $x_0 \in \operatorname{int} A$ and $x_1 \in \operatorname{cl}A$. $x = (1-\lambda)x_0 + \lambda x_1$ is the line passed x_0 and x_1. If $\lambda > 1$, then $x_\lambda \notin \operatorname{cl}A$, i.e., $(x_1, \infty) \cap \operatorname{cl}A = \varnothing$. □

1.2.2 Construction of Convex Sets

This subsection deals with the construction of convex sets in a normed space.

Let X be a normed space and $x_1, x_2 \in X$. The set $\{x; x = (1-\lambda)\,x_1 + \lambda x_2, \lambda \in [0,1]\}$ is called the segment connecting x_1 and x_2, and denoted by $[x_1, x_2]$ at the last subsection. It is obvious that the segment $[x_1, x_2]$ is a convex set and is the minimal convex set which contains x_1 and x_2. Hence, in this subsection we also call it the convex combination of two points x_1 and x_2. We now extend the concept to more than two points.

Let $x_1, x_2, x_3 \in X$. Then the following set

$$\{x; x = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3, \lambda_1, \lambda_2, \lambda_3 \in [0,1]\,, \lambda_1 + \lambda_2 + \lambda_3 = 1\}$$

is defined as the convex combination of x_1, x_2, and x_3, and denoted by conx(x_1, x_2, x_3). When x_1, x_2, and x_3 are linear independent, i.e., they are not in a line,[7] then the set conx(x_1, x_2, x_3) is a closed triangle with vertexes x_1, x_2, and x_3. The common definition of convex combination is as follows.

Definition 1.2.2 Let $A = \{x_n, n = 1, 2, \ldots\}$ be a set in X, and $\{x_{n_i}, i = 1, 2, \ldots, m\} \subset A$ is a finite subset of A, then $x = \sum_{i=1}^{m} \lambda_i x_{n_i}, \lambda_i > 0, \sum_{i=1}^{m} \lambda_i = 1$ is called a finite convex combination of A. The set of all finite convex combinations of A is called the convex combination of A and denoted by conx $(x_n; n = 1, 2, \ldots)$, i.e.,

$$\operatorname{conx}\,(x_n, n = 1, 2, \ldots) = \left\{x; x = \sum_{i=1}^{m} \lambda_i x_{n_i}, m \in \mathbb{N}, \lambda_i > 0, \sum_{i=1}^{m} \lambda_i = 1, x_{n_i} \in \{x_n\}\right\}. \tag{1.2.3}$$

□

An alternative statement of Eq. (1.2.3) is

$$\operatorname{conx}\,(x_n, n = 1, 2, \ldots) = \left\{x; x = \sum_{n=1}^{\infty} \lambda_i x_n, \lambda_i \geq 0, \sum_{i=1}^{m} \lambda_i = 1, \right. \\ \left. \text{but only finite } \lambda_i \neq 0\right\}.$$

Usually, conx $(x_n;\ n = 1, 2, \ldots)$ is also called the polytope of $\{x_n; n = 1, 2, \ldots\}$, and $x_n, n = 1, 2, \ldots$ are its vertexes. If $x_i - x_1, i = 2, 3, \ldots$ are linear independent, then conx $(x_n; n = 1, 2, \ldots)$ is a simplex.

Theorem 1.2.4 conx $(x_n, n = 1, 2, \ldots)$ is a convex set.

[7] The requirement is equivalent to $x_3 \notin \{x; x = (1-\lambda)\,x_1 + \lambda x_2, \lambda \in \mathbb{R}\}$.

Proof Let $x, y \in \operatorname{conx}(x_n, n = 1, 2, \dots)$. Then by the definition, we have

$$x = \sum_{i=1}^{m} \nu_i x_{n_i}, \quad \nu_i \geq 0, \quad \sum_{i=1}^{m} \nu_i = 1,$$

and[8]

$$y = \sum_{i=1}^{m} \mu_i x_{n_i}, \quad \mu_i \geq 0, \quad \sum_{i=1}^{m} \mu_i = 1.$$

Then for every $\lambda \in [0, 1]$

$$(1 - \lambda)\, x + \lambda y = (1 - \lambda) \sum_{i=1}^{m} \nu_i x_{n_i} + \lambda \sum_{i=1}^{m} \mu_i x_{n_i} = \sum_{i=1}^{m} [(1 - \lambda)\, \nu_i + \lambda \mu_i]\, x_{n_i}.$$

It is obvious that $(1 - \lambda)\, \nu_i + \lambda \mu_i \geq 0$, and

$$\sum_{i=1}^{m} [(1 - \lambda)\, \nu_i + \lambda \mu_i] = (1 - \lambda) \sum_{i=1}^{m} \nu_i + \lambda \sum_{i=1}^{m} \mu_i = 1.$$

The above discussion implies that $(1 - \lambda)\, x + \lambda y \in \operatorname{conx}(x_n, n = 1, 2, \dots)$. □

Theorem 1.2.4 also illustrates that if $x \in \operatorname{conx}(x_n, n = 1, 2, \dots)$, then there exist a $x_{i_0} \in \{x_n, n = 1, 2, \dots\}$ and a $y \in \operatorname{conx}\{\{x_n, n = 1, 2, \dots\} / \{x_{i_0}\}\}$ such that $x = (1 - \lambda)\, x_{i_0} + \lambda y, \ \ \lambda \in [0, 1]$. The detailed proof of the conclusion is left to readers.

From the proof, we can find that the conclusion is still valid in a linear space, because we have no use of the norms of vectors.

Definition 1.2.3 $A \subset X$ is a set. The convex hull of A is the minimal convex set which contains A. The convex hull of set A is denoted by co$A$. □

Let $\mathscr{P}A$ be the power set of A. It is well-known that by the operation of containing, $\mathscr{P}A$ is only a partial order set. Fortunately, we can prove $\operatorname{co}A = \bigcap_{\Lambda \supset A, \Lambda \text{ is convex}} \Lambda$, the right side is unique, so is the co$A$. Because $\operatorname{conx}(x_i, x_i \in A)$ is convex and contains A. Hence, it can be proved that $\operatorname{co}A = \operatorname{conx}(x_i, x_i \in A)$.

If A is a set in a finite dimensional space X. The following theorem given by Caratheodory reveals the construction of co A.

Theorem 1.2.5 Let X be an n-dimensional linear space, $A \subset X$ be a set of X. For every $x \in$ co A there exist m vectors $x_i \in A, \quad i = 1, 2, \dots, m$, such that $x = \sum_{i=1}^{m} \lambda_i x_i$ where $\lambda_i > 0$ and $\sum_{i=1}^{m} \lambda_i = 1$. Particularly, we can make $m \leq n + 1$.

[8] By adding zeros, we can require that x and y are yielded by the same x_{n_i}'s.

Proof If the vector x belongs to A, then the conclusion is obvious. Hence the key issue is to prove the conclusion for the vector in coA/A. By Theorem 1.2.4, every vector in co$A$ is a finite combination $\sum_{i=1}^{m} \lambda_i x_i$ where $x_i \in A$ and $\lambda_i > 0$, $\sum_{i=1}^{m} \lambda_i = 1$, $i = 1, 2, \ldots, m$. It is sufficient to verify that $m \leq n + 1$.

The conclusion is verified by the following way. If $m > n + 1$, then there is another convex combination which also yields x but only contains $m - 1$ x_i's.

Suppose now $x = \sum_{i=1}^{m} \lambda_i x_i$ with $m > n + 1$, all $\lambda_i > 0$ and $x_i \in A \subset X$, $i = 1, 2, \ldots, m$. X is an n-dimensional linear space, every x_i can be expressed by $x_i = \left[x_i^{(1)}\ x_i^{(2)} \cdots x_i^{(n)}\right]^T$ where $x_i^{(j)}$ is the jth component of x_i. Now we expand x_i to an $(n+1)$-dimensional vector $\tilde{x}_i = \left[x_i^{(1)}\ x_i^{(2)} \cdots x_i^{(n)}\ 1\right]^T$. The m vectors $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m$ have to be linearly dependent. It implies there are m real numbers $\alpha_1, \alpha_2, \ldots, \alpha_m$ which are not all equal to zero such that $\alpha_1 \tilde{x}_1 + \alpha_2 \tilde{x}_2 + \cdots + \alpha_m \tilde{x}_m = 0$, i.e.,

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m = 0, \tag{1.2.4a}$$

$$\alpha_1 + \alpha_2 + \cdots + \alpha_m = 0. \tag{1.2.4b}$$

Without loss of generality, we assume that $\alpha_m \neq 0$ and $\alpha_m > 0$ (if $\alpha_m < 0$, we can multiply -1 to Eqs. (1.2.4a) and (1.2.4b).

Now we consider the set $\Lambda = \left\{\frac{\lambda_j}{\alpha_j}; \alpha_j > 0,\ j = 1, 2, \ldots, m\right\}$. Because λ_m/α_m belongs to Λ, Λ is a nonempty and finite set. Let $\beta = \lambda_{j_0}/\alpha_{j_0} = \min\left\{\frac{\lambda_j}{\alpha_j}; \alpha_j > 0,\ j = 1, 2, \ldots, m\right\}$, then $\beta > 0$. Let us denote $\mu_i = \lambda_i - \beta\alpha_i$ for all $i = 1, 2, \ldots, m$. Then $\mu_i \geq 0$, and $\mu_{j_0} = 0$. Now we compute

$$\sum_{i=1}^{m} \mu_i = \sum_{i=1}^{m} (\lambda_i - \beta\alpha_i) = \sum_{i=1}^{m} \lambda_i - \beta \sum_{i=1}^{m} \alpha_i = 1,$$

and

$$\sum_{i=1}^{m} \mu_i x_i = \sum_{i=1}^{m} (\lambda_i - \beta\alpha_i) x_i = \sum_{i=1}^{m} \lambda_i x_i - \beta \sum_{i=1}^{m} \alpha_i x_i = x.$$

Because $\mu_{j_0} = 0$, in the equation $x = \sum_{i=1}^{m} \mu_i x_i$, there are at most $m - 1$ nonzero coefficients. Thus, we complete the proof of the theorem. □

In an n-dimensional space, a set of basis holds n vectors, and every vector in the space can be described uniquely by a linear combination of the n vectors. Theorem 1.2.5 shows that a vector in co$A$ can be described by a convex combination of elements in A, but it may need $n + 1$ vectors in an n-dimensional space. Moreover, even if it can be described by a convex combination of n or less vectors, these vectors still may be linear dependent. The reason is that all coefficients in a convex combination have to be in the interval [0,1].

As a useful corollary we state the following result.

Corollary 1.2.1 If A is a compact set in an n-dimensional space, then co$A$ is also compact. □

The corollary is easy to be verified by using the fact that a compact set in a finite dimensional space is bounded and closed. We omit its proof and leave it to readers.

When we deal with the topological properties of convex sets, the concepts of open set, interior and boundary of a set are very frequently mentioned. However when a convex is located in a subspace of a normed space, it has no inner point. The fact results in much inconvenience. Hence we yield an idea to introduce the related topology for these convex sets.

Definition 1.2.4 Let X be a normed space, $A \subset X$ is a set. If there exists a vector $x_0 \in X$ and a subspace $H \subset X$ such that $x_0 + A \subset H$, then the x_0 is called by supporting vector, and H is carrying subspace, and the fact is denoted by $H = \text{car}\, A$. □

If x is a supporting vector and H is a carrying subspace of A, then for two vectors $a, y \in A$, $x + a \in H$ and $x + y \in H$. It follows $a - y \in H$, or $-y + A \subset H$ for every $y \in A$. The argument illustrates that if A holds a carrying subspace and $y \in A$, then $-y$ is qualified as a supporting vector. Hence, the supporting vector is not unique, but we can prove that the minimal carrying subspace for a set $A \subset X$ is unique. In order to prove the conclusion, we give a lemma and omit its proof since it is direct.

Lemma 1.2.1 Suppose $x_1, x_2, \ldots, x_n$ form a basis of an n-dimensional linear space X. Then the set $\{x; x = a_1x_1 + a_2x_2 + \cdots + a_nx_n,\ a_i \in (-\varepsilon, \varepsilon)\}$ is an open set containing the origin. □

Lemma 1.2.1 is valid for a finite dimensional normed space X. For $\mathbb{R}^n$, we have a more precise illustration. Let x_i be the i-th unit vector of $\mathbb{R}^n$, i.e., its i-th component is 1 and others are all zero. Then the open set defined by Lemma 1.2.1 is $\left\{x; \max\limits_i |x_i| < \varepsilon\right\}$. When $\mathbb{R}^n$ adopts the maximal norm, then the open set is just an open ball whose center is the origin O and radius ε.

Theorem 1.2.6 Let X be an n-dimensional linear space, and $A \subset X$ be a convex set. If int$A = \varnothing$, then there is an m-dimensional ($m < n$) subspace H of X such that $H = \text{car}\, A$.

Proof If $A = \varnothing$, then every subspace is qualified as its carrying subspace. Hence, we can assume that $A \neq \varnothing$. Let us fix a vector $x_0 \in A$, and find as many as possible x_i in A, $i = 1, 2, \ldots, r$, such that $x_i - x_0$ are linear independent. We then consider conx$(x_0, x_1, \ldots, x_r)$. conx $(x_0, x_1, \ldots, x_r) \subset A$ because of the convexity of A. To prove the theorem, we deal with two cases.

(1) If $r = n$, then $x_i - x_0$, $i = 1, 2, \ldots, r$, form a basis of X. Define

$$\overline{x} = \frac{x_0 + x_1 + x_2 + \cdots + x_n}{n+1}, \tag{1.2.5}$$

then $\overline{x} \in$ conx $(x_0, x_1, \ldots, x_n) \subset A$. Equation (1.2.5) can be rewritten as

$$\overline{x} - x_0 = \frac{1}{n+1}(x_1 - x_0) + \frac{1}{n+1}(x_2 - x_0) + \cdots + \frac{1}{n+1}(x_n - x_0). \tag{1.2.6}$$

Define a vector $x(\eta)$ as follows

$$\begin{aligned} x(\eta) - x_0 = &\left(\frac{1}{n+1} + \eta_1\right)(x_1 - x_0) + \left(\frac{1}{n+1} + \eta_2\right)(x_2 - x_0) \\ &+ \cdots + \left(\frac{1}{n+1} + \eta_n\right)(x_n - x_0). \end{aligned} \tag{1.2.7}$$

where η_i, $i = 1, 2, \ldots, n$, all vary in the interval $\left[-1/(n+1)^2, 1/(n+1)^2\right]$. Subtracting Eq. (1.2.6) from Eq. (1.2.7), we have

$$x(\eta) - \overline{x} = \eta_1(x_1 - x_0) + \eta_2(x_2 - x_0) + \cdots + \eta_n(x_n - x_0).$$

By Lemma 1.2.1, $\left\{x(\eta),\ \eta_i \in \left[-1/(n+1)^2, 1/(n+1)^2\right],\quad i = 1, 2, \ldots, n\right\}$ is an open set with center $\overline{x}$.

On the other hand, $x(\eta)$ is equal to

$$\begin{aligned} x(\eta) = x_0 &+ \left(\frac{1}{n+1} + \eta_1\right)(x_1 - x_0) + \left(\frac{1}{n+1} + \eta_2\right)(x_2 - x_0) + \cdots \\ &+ \left(\frac{1}{n+1} + \eta_n\right)(x_n - x_0) \\ = &\left(1 - \frac{n}{n+1} - \sum_{i=1}^{n} \eta_i\right)x_0 + \left(\frac{1}{n+1} + \eta_1\right)x_1 + \left(\frac{1}{n+1} + \eta_2\right)x_2 + \cdots \\ &+ \left(\frac{1}{n+1} + \eta_n\right)x_n. \end{aligned}$$

Because $\eta_i \in \left(-1/(n+1)^2, 1/(n+1)^2\right)$,

$$0 < \frac{n}{(n+1)^2} = \frac{1}{n+1} - \frac{1}{(n+1)^2} \le \frac{1}{n+1} + \eta_i \le \frac{1}{n+1} + \frac{1}{(n+1)^2} = \frac{n+2}{(n+1)^2} < 1,$$

and

$$0 < \frac{1}{(n+1)^2} = \frac{1}{n+1} - \frac{n}{(n+1)^2} \leq 1 - \frac{n}{n+1} - \sum_{i=1}^{n} \eta_i \leq \frac{1}{n+1} + \frac{n}{(n+1)^2}$$
$$= \frac{2n+1}{(n+1)^2} < 1.$$

The above two inequalities are all valid for $n \in \mathbb{N}$. Furthermore,

$$\left(1 - \frac{n}{n+1} - \sum_{i=1}^{n} \eta_i\right) + \left(\frac{1}{n+1} + \eta_1\right) + \left(\frac{1}{n+1} + \eta_2\right) + \cdots + \left(\frac{1}{n+1} + \eta_n\right) = 1. \tag{1.2.8}$$

We conclude that $x(\eta) \in \text{conx}\,(x_0, x_1, \ldots, x_n) \subset A$ for all $\eta_i \in (-1/(n+1)^2, 1/(n+1)^2)$. Thus, $\bar{x} \in \text{int} A$ which contradicts to $\text{int} A = \varnothing$.

(2) From (1), we have $r < n$, then $A - x_0 \subset \text{span}\,(x_1 - x_0, x_2 - x_0, \ldots, x_r - x_0)$. By Definition 1.2.4, $-x_0$ is a supporting vector and $\text{span}\,(x_1 - x_0, x_2 - x_0, \ldots, x_r - x_0)$ is the carrying subspace whose dimension is $r\,(< n)$. □

Using the concept of carrying subspace, we can define the following concepts.

Definition 1.2.5 Let $\dim\ (\text{car}\, A) < n$. Then $x \in A$ is called as a relative inner point if there is an $\varepsilon > 0$, such that $x + (\varepsilon B \cap \text{car} A) \subset A$. The set of relative inner points is called relative interior, and denoted by $\text{re}\,\text{int} A$; $\text{cl}\, A/\text{re}\,\text{int} A$ is called the relative boundary of A and denoted by $\text{re bd} A$. □

By using Definition 1.2.5, Theorems 1.2.2 and 1.2.3 can be extended to the relative interior. For example, the extension of Theorem 1.2.2 is given as follows, and its proof is omitted.

Corollary 1.2.2 If A is a convex set, $x_0 \in \text{re int}\ A$ and $x_1 \in \text{cl}\ A$, then $[x_0, x_1) \subset \text{re}\,\text{int}\ A$. □

1.2.3 Separation Theorems

This subsection deals with the separation theorems of convex sets. Separation theorems are also basic conclusions for the optimization theory. This subsection starts with the properties of projection on convex sets, and then separation theorems are presented and proved. In this subsection, X is always an inner product space whose dimension is finite.

Suppose set $A \subset X$, and $x \in X$. The distance from x to A is $d\,(x, A) = \inf\limits_{a \in A} \|x - a\|$. Define a set $\pi(x, A)$ as follows:

$$\pi\,(x, A) = \{y; y \in A, \|x - y\| = d\,(x, A)\}\,.$$

$\pi(x, A)$ is called by the projection of x on A. For a given $x \in X$, usually, $\pi(x, A)$ is only a set. It is because that if A is not closed $\pi(x, A)$ may be empty; but sometimes, $\pi(x, A)$ may have more than one elements. When A is a closed and convex set, the conclusion is determined. The following theorem shows that $\pi(x, A)$ is a mapping provided that A is a closed and convex.

Theorem 1.2.7 If A is a closed and convex set of X, then for every $x \in X$, $\pi(x, A)$ has one and only one element.

Proof Because A is closed, $\pi\,(x, A) \neq \varnothing$ for every $x \in X$. We now verify that it has only one element.

If there are $a, b \in A$ and $a \neq b$ such that $\|x-a\| = \|x-b\| = d\,(x, A)$, then by the parallelogram law of inner product (Eq. 1.1.3), we have

$$2\left(\|x-a\|^2 + \|x-b\|^2\right) = \|a-b\|^2 + \|2x-a-b\|^2.$$

Because $a \neq b$, $\|a-b\| > 0$. The above equation implies $d(x, A)^2 = \|x-a\|^2 > \left\|x - \frac{a+b}{2}\right\|^2$. A is convex, consequently, $(a+b)/2 \in A$. The inequality contradicts to the definition of distance. We then conclude $\pi(x, A)$ has only one element. □

Differing from the case of linear subspace, usually, it is not true for a convex set in an inner space that $(x - \pi\,(x, A)) \perp A$. An alternative conclusion will be given below.

We have mentioned at the last section that if $\|x_0\| = 1$, then $\langle x, x_0\rangle$ is the projection of x at the line where the vector x_0 is located. The fact is used to deal with the separation of two convex sets.

Let A_1 and A_2 be two convex sets of an n-dimensional space X, and $A_1 \cap A_2 = \varnothing$. If there is an ($n$-1)-dimensional hyperplane $p^T x = c$, where p is the normal vector of the hyperplane, such that A_1 and A_2 locate different sides of the hyperplane, then we say that A_1 and A_2 are separated by the hyperplane $p^T x = c$. Figure 1.2 presents an illustration of the separation.

If A_1 and A_2 are separated by a hyperplane $p^T x = c$, then for one set, to say A_1, $\langle p, x_1\rangle \leq c$ for every $x_1 \in A_1$; and $\langle p, x_2\rangle \geq c$ for every $x_2 \in A_2$. Therefore, separating two convex sets by a hyperplane is equivalent to look for a vector p such that $\langle p, x_1\rangle \leq \langle p, x_2\rangle$ for every $x_1 \in A_1$ and every $x_2 \in A_2$.

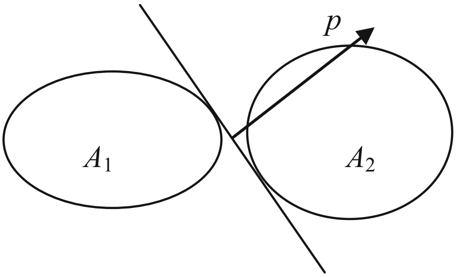


Fig. 1.2 Two convex sets separated by a hyperplane

Lemma 1.2.2 Suppose $A \subset X$ is a convex set, and $x_0 \notin A$. Then there is a $p \in X$ and $p \neq 0$ such that $\langle p, x_0 \rangle \geq \langle p, x_A \rangle$ for every $x_A \in A$.

Proof The proof of the lemma consists of two cases.

(1) $x_0 \notin \text{cl } A$. By Theorem 1.2.6, there is a $y \in \text{cl } A$ such that $y = \pi(x_0, \text{cl}A)$. It follows that $\|x_A - x_0\| \geq \|y - x_0\|$ for every $x_A \in \text{cl}A$. Since A is convex, $\lambda x_A + (1-\lambda) y \in A$. We have

$$\|y - x_0\|^2 \leq \|\lambda x_A + (1-\lambda) y - x_0\|^2$$

$$= \lambda^2 \|x_A - y\|^2 + 2\lambda \langle x_A - y, y - x_0 \rangle + \|y - x_0\|^2.$$

Hence, for every $x_A \in \text{cl}A$ and $\lambda \in [0, 1]$,

$$\lambda^2 \|x_A - y\|^2 + 2\lambda \langle x_A - y, y - x_0 \rangle \geq 0. \tag{1.2.9}$$

From Inequality (1.2.9), we can conclude that for every $x_A \in \text{cl}A$

$$\langle x_A - y, y - x_0 \rangle \geq 0. \tag{1.2.10}$$

If Inequality (1.2.10) is not true, i.e., there is a $x_{A_0} \in \text{cl}A$ such that $\langle x_{A_0} - y, y - x_0 \rangle < 0$, then $\lambda^2 \|x_{A_0} - y\|^2 + 2\lambda \langle x_{A_0} - y, y - x_0 \rangle < 0$ for some small λ.

Now, let $p = x_0 - y$. Then $p \neq 0$. Inequality (1.2.10) becomes $\langle x_A - y, p \rangle \leq 0$, i.e.,

$$\langle x_A, p \rangle \leq \langle y, p \rangle = \langle x_0 - p, p \rangle = \langle x_0, p \rangle - \|p\|^2. \tag{1.2.11}$$

It implies $\langle p, x_A \rangle < \langle p, x_0 \rangle$.

(2) If $x_0 \in \text{cl}A/A$, then x_0 is on the boundary. There is a sequence $\{x_k\} \subset A^c$ such that $x_k \to x_0$. By the proof above, there is a p_k for each k such that

$$\langle p_k, x_A \rangle \leq \langle p_k, x_0 \rangle - \|p_k\|^2.$$

Define $\widehat{p}_k = p_k / \|p_k\|$, then $\|\widehat{p}_k\| = 1$. Substituting p_k by $\widehat{p}_k \|p_k\|$, the above inequality yields

$$\langle \widehat{p}_k, x_A \rangle \leq \langle \widehat{p}_k, x_0 \rangle - \|p_k\|. \tag{1.2.12}$$

The shell of the unit ball is a compact set in a finite dimensional space, there is a subsequence $\{\widehat{p}_{k_i}\}$ of $\{\widehat{p}_k\}$, such that $\widehat{p}_{k_i} \to p_0$. Because $\|\widehat{p}_{k_i}\| = 1$, $\|p_0\| \neq 0$. Hence Inequality (1.2.12) leads to $\langle \widehat{p}_0, x_A \rangle \leq \langle \widehat{p}_0, x_0 \rangle$.

The lemma is proved. □

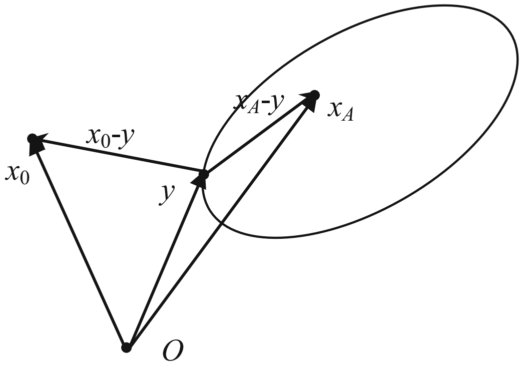


Fig. 1.3 An illustration of Inequality (1.2.13)

Lemma 1.2.2. illustrates that if $x_0 \notin A$, then there is a hyperplane which separates x_0 from A. If the equality sign in $\langle p_0, x_A\rangle \leq \langle p_0, x_0\rangle$ cannot be removed, then $x_0 \in$ cl A and the hyperplane is tangent to set A and the point x_0 is on the hyperplane. Lemma 1.2.2 is a useful conclusion; hence, more remarks are given below

Remark 1 Because $y = \pi\,(x_0, A)$, Inequality (1.2.10) is equivalent to

$$\langle x_A - \pi\,(x_0, A)\,, x_0 - \pi\,(x_0, A)\rangle \leq 0, \tag{1.2.13}$$

where $x_0 \notin A$ and A is convex. Inequality (1.2.13) is also true if $x_0 \in A$, because it implies $x_0 = \pi\,(x_0, A)$. Figure 1.3 presents an illustration of Inequality (1.2.13), the angle included by $x_0 - \pi\,(x_0, A)$ and $x_A - \pi\,(x_0, A)$ is always an obtuse angle. □

Remark 2 In Sect. 1.1, we have mentioned when A is a subspace, then $(x_0 - \pi\,(x_0, A)) \in M^{\perp}$, i.e., $\langle x_A - \pi\,(x_0, A)\,, x_0 - \pi\,(x_0, A)\rangle = 0$ for every $x_A \in A$. In this sense, Inequality (1.2.13) is an extension from the subspace case. □

Remark 3 We can prove an inverse result that for a closed and convex set A, $x_0 \notin A$, if $\langle x_A - y, x_0 - y\rangle \leq 0$ holds for every $x_A \in A$, then $y = \pi\,(x_0, A)$. □

Remark 4 Inequality (1.2.11) is a stronger conclusion than Inequality (1.2.10). It shows if $x_0 \notin$ cl A, then there is a constant $\beta > 0$ and a vector $p \in X$, such that $\langle x_0,\ p\rangle > \beta > \langle x_A, p\rangle$. The fact also can be explained as that if $x_0 \notin$ cl A, then there is a hyperplane which separates x_0 from A; moreover, neither x_0 nor A touches the hyperplane. □

Remark 5 The conclusion can be extended to the case where X is not finitely dimensional by using Alaoglu theorem mentioned in Sect. 1.1.4. □

Lemma 1.2.2 is now used to verify the main conclusion of this subsection.

Theorem 1.2.8 If A_1 and A_2 are two convex sets in a finite dimensional space X, and $A_1 \cap A_2 = \varnothing$, then there is $p \in X$ such that $\langle x_1, p\rangle \leq \langle x_2, p\rangle$ for every $x_1 \in A_1$ and every $x_2 \in A_2$.

Proof Define $A = A_1 - A_2$, then A is convex (to see Problem 6 of this section) and $0 \notin A$. By Lemma 1.2.2, there is a vector p such that $\langle x_A, p\rangle \leq \langle 0, p\rangle = 0$. Note that the vector x_A takes the form of $x_A = x_1 - x_2$, with $x_1 \in A_1$ and $x_2 \in A_2$. Hence $\langle x_1 - x_2, p\rangle \leq 0$, i.e., $\langle x_1, p\rangle \leq \langle x_2, p\rangle$. □

Theorem 1.2.8 is also known as variation theorem which plays a primary role in the theory of variation. Theorem 1.2.8 has a corollary as follows.

Corollary 1.2.3 If A_1 and A_2 are two closed and convex sets in a finite dimensional space X, A_1 is compact and $A_1 \cap A_2 = \varnothing$, then there is $p \in X$ and $\varepsilon > 0$ such that $\langle x_1, p\rangle \leq \langle x_2, p\rangle - \varepsilon$ for every $x_1 \in A_1$ and every $x_2 \in A_2$. □

By the property of compact set, the proof of Corollary 1.2.3 is direct, we leave it to readers as an exercise.

Problems

1. A is a set in a linear space. If $x \in A$, and $\alpha x \in A$, for every $\alpha \in \mathbb{R}\,(\geq 0)$, then A is a cone. Prove that A is a convex cone if and only if for $x, y \in A$, $\alpha, \beta \in \mathbb{R}\,(\geq 0)$, $\alpha x + \beta y \in A$.
2. Let $A_i, i = 1, 2, \ldots$ be convex sets in a normed space X. Define $X^\infty = X \times X \times \cdots$ is the set of $\{x = [x_1, x_2, \ldots], x_i \in X\}$ in which only finite components x_i are nonzero. X^∞ is called as the minimal infinite extension of X. Prove that

   (1) X^∞ is a normed space if X is.
   (2) If $A_1 \times A_2 \times \cdots \subset X^\infty$ then it is a convex set in X^∞ (It requires the origin belongs to almost all A_i).

3. If $A \subset X$ is a bounded set and X is an n-dimensional linear space, then cl co$A$ is also a compact set.
4. Prove Theorem 1.2.3.
5. Prove Corollary 1.2.1.
6. If A_1 and A_2 are both convex sets, $A_1 \cap A_2 = \varnothing$, then $A_1 - A_2$ is a convex. Give an example to illustrate that the conclusion may fail if $A_1 \cap A_2 \neq \varnothing$.
7. Let X be an inner product space, and $A \subset X$ be a convex set. Then the projection $\pi(x, A)$ has the following properties.

   (1) $\forall x \in A, \ \pi(x, A) = x$.
   (2) $\pi(x, A)$ is a Lipschitz mapping and its Lipschitz constant can be 1.
   (3) $\pi(x, A)$ is monotonous, i.e., $\langle \pi(x, A) - \pi(y, A), x - y\rangle \geq 0$ for $x, y \in X$.

8. If $A \subset X$ is a convex cone, then the projection $\pi(x, A)$ has the following properties.

   (1) For $x \in X, y \in A$, $\langle x - \pi(x, A), y\rangle \leq 0$
   (2) For $x \in X$, $\langle x - \pi(x, A), \pi(x, A)\rangle = 0$
   (3) For $x \in X$, $\|x\|^2 = \|\pi(x, \mathrm{cl}A)\|^2 + \|x - \pi(x, \mathrm{cl}A)\|^2$

   The above conclusions illustrate that a convex cone is more similar to a linear subspace because they hold many similar properties.
9. Prove Corollary 1.2.3.

1.3 Convex Functions

At the beginning of the last section, we pointed out that convex analysis contains two fundamental concepts: convex sets and convex functions. The convex sets have been considered. This section turns to convex functions. The concept of convex functions is given first, then elemental properties of convex functions are listed, and several useful convex functions are introduced; at last, the continuity and derivative are discussed for convex functions.

1.3.1 Convex Functions and Their Elementary Properties

The space considered in this section is $\mathbb{R}^n$ which is treated as an inner product space. $f : \mathbb{R}^n \to \mathscr{R}$ is a function, $A \subset \mathbb{R}^n$ is a set.

Definition 1.3.1 $f : A \to \mathscr{R}$ is a convex function on a set A if A is a convex set and for $x_1, x_2 \in A$ and $\lambda \in [0, 1]$

$$f\left((1-\lambda)x_1 + \lambda x_2\right) \leq (1-\lambda) f(x_1) + \lambda f(x_2)\,. \tag{1.3.1}$$

f is a strictly convex function if the signal "$\leq$" in Inequality (1.3.1) can be replaced by "<". □

Remark The set $\{x = (1-\lambda)x_1 + \lambda x_2, \quad \lambda \in [0, 1]\}$ has been denoted by a segment $[x_1, x_2]$. Therefore, f is a convex function on A if and only if it is convex on every segment $[x_1, x_2] \subset A$. □

The concave function is defined by using the convex function. $f : A \to \mathscr{R}$ is said to be a concave function if $-f$ is a convex function. It is worth to note that the domain of a concave function is a convex set. This book does not consider concave functions except it has to be mentioned.

An equivalent statement of Definition 1.3.1 is as follows.

f is a convex function on the set A, if A is convex and for n vectors $x_i \in A, \quad i = 1, 2, \ldots, n$, the following inequality holds

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i), \tag{1.3.2}$$

where $\lambda_i \in [0, 1]\,, \quad \sum_{i=1}^{n} \lambda_i = 1.$

Inequality (1.3.1) really implies Inequality (1.3.2). The proof is left as an exercise to readers. Inequality (1.3.2) is known as Jensen Inequality.

The properties of convex functions are listed below and all proofs are omitted. The readers are suggested to refer to a textbook of convex analysis for these proofs.

In the following, f and f_i, $i = 1, 2, \ldots$ are convex functions. They hold the same domain A.

(1) $g = f_1 + f_2$ and $g = af_1$ are convex function if $a > 0$.
(2) Let $M \in \mathbb{R}^{m \times n}$ be a real matrix, and $b \in \mathbb{R}^m$ be a vector. Then $f(Mx + b)$ is a convex function.
(3) If $g : \mathbb{R} \to \mathscr{R}$ is a monotonously increasing and convex function, then $g(f(x))$ is a convex function by the definitions that $g(\infty) = \sup\limits_{x \in \mathbb{R}} g(x)$, $g(-\infty) = \inf\limits_{x \in \mathbb{R}} g(x)$.
(4) $g(x) = \sup\limits_{i} f_i(x)$ is a convex function.
(5) $g(x) = \lim\limits_{i \to \infty} \sup f_i(x)$ is a convex function where $\lim\limits_{i \to \infty} \sup f_i(x) = \lim\limits_{i \to \infty} \sup\limits_{k > i} (f_k(x))$.

The last two conclusions are also valid if there are infinite f_i's.

As exercises, the readers are suggested to deal with the corresponding properties for strictly convex functions and concave functions, respectively.

Definition 1.3.2 Let $f : A \to \mathbb{R}$ be a function. The epigraph of f is defined as follows

$$\operatorname{epi} f = \{(x, \alpha)\,;\; x \in A,\ \alpha \in \mathbb{R},\ \text{ and } f(x) \le \alpha\}. \qquad \square$$

Theorem 1.3.1 $f : A \to \mathbb{R}$ is a convex function if and only if its epigraph epi$f$ is a convex set in $\mathbb{R}^n \times \mathbb{R}$

Proof Suppose that (x_1, α_1), $(x_2, \alpha_2) \in$ epi f. By the definition of epigraph, we have $f(x_i) \le \alpha_i$, $i = 1, 2$.

A point in $\mathbb{R}^n \times \mathbb{R}$ can be expressed by the form of $\begin{bmatrix} x \\ \alpha \end{bmatrix}$. Using the notation, we have $\begin{bmatrix} x_i \\ f(x_i) \end{bmatrix} \in$ epi f, $\quad i = 1, 2$. If epi$f$ is a convex set, then

$$(1-\lambda)\begin{bmatrix} x_1 \\ f(x_1) \end{bmatrix} + \lambda \begin{bmatrix} x_2 \\ f(x_2) \end{bmatrix} = \begin{bmatrix} (1-\lambda)x_1 + \lambda x_2 \\ (1-\lambda)f(x_1) + \lambda f(x_2) \end{bmatrix} \in \operatorname{epi} f,$$

because A is convex, $(1-\lambda)x_1 + \lambda x_2 \in A$ for every $\lambda \in [0,\ 1]$. By the definition of epigraph, it is true that $f((1-\lambda)x_1 + \lambda x_2) \le (1-\lambda)f(x_1) + \lambda f(x_2)$. The sufficiency is verified.

Now we turn to the proof of necessity. If $\begin{bmatrix} x_i \\ \alpha_i \end{bmatrix} \in$ epi f, $i = 1, 2$, then $x_i \in A$ and $f(x_i) \le \alpha_i$. $(1-\lambda)x_1 + \lambda x_2 \in A$ because A is a convex set. Moreover, since f is a convex function, $f((1-\lambda)x_1 + \lambda x_2) \le (1-\lambda)f(x_1) + \lambda f(x_2)$. It implies

$$f((1-\lambda)x_1 + \lambda x_2) \le (1-\lambda)f(x_1) + \lambda f(x_2) \le (1-\lambda)\alpha x_1 + \lambda \alpha_2.$$

By the definition of epigraph,

$$(1-\lambda)\begin{bmatrix} x_1 \\ \alpha_1 \end{bmatrix} + \lambda\begin{bmatrix} x_2 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} (1-\lambda)\,x_1 + \lambda x_2 \\ (1-\lambda)\,\alpha_1 + \lambda\alpha_2 \end{bmatrix} \in \operatorname{epi} f.$$

It implies epi$f$ is convex. □

Define a set as follows

$$\text{lev }\ (f \le \alpha) = \{x \in A; f(x) \le \alpha\}\,.$$

The set lev $(f \le \alpha)$ is called the level set of$f$ related to real number α. Similarly, we can define the set lev $(f < \alpha)$. It can be verified that lev $(f \le \alpha)$ and lev $(f < \alpha)$ are both convex if$f$ is a convex function. In the next chapter and Chap. 4, the concept of level set will be applied.

To end this subsection, we deal with the Lipschitzian property for the convex function.

Lemma 1.3.1 Let$f : \mathbb{R}^n \to \mathscr{R}$ be a convex function and $x_0 \in \mathbb{R}^n$. Suppose that$f$ holds upper bounded on the neighborhood$B(x_0, \varepsilon)$, then there is a $\delta > 0$ such that$f$ is a Lipschitzian function on the neighborhood$B(x_0, \delta)$.

Proof By the assumption of the lemma, there is a constant $M > 0$, such that for every $x \in B\,(x_0, \varepsilon), f(x) \le M$. Now we prove that$f$ is also lower bounded. Let $z \in B\,(x_0, \varepsilon)$ then there is a $z' \in B\,(x_0, \varepsilon)$ such that $x_0 = (z + z')\,/2$. f is a convex function; hence we have$f\,(x_0) \le \frac{1}{2}f(z) + \frac{1}{2}f\,(z')$. $f\,(z') \le M$, thus,$f(z) \ge 2f\,(x_0) - M$.

Denote $c = \max\{|2f\,(x_0) - M|\,, M\}$, the above discussion implies $|f(z)| \le c$, for every $z \in B\,(x_0, \varepsilon)$. Take $\delta = \varepsilon/2$, for $x_1, x_2 \in B\,(x_0, \delta)$, we denote $\Delta = \|x_1 - x_2\| < 2\delta$ and define $x_3 = x_2 + (\delta/\Delta)\,(x_2 - x_1)$. $x_3 \in B\,(x_0, \varepsilon)$ because $\|x_3 - x_0\| \le \|x_2 - x_0\| + (\delta/\Delta)\,\|x_2 - x_1\| < 2\delta = \varepsilon$. From the equation $x_3 = x_2 + (\delta/\Delta)\,(x_2 - x_1)$, we have $x_2 = \frac{\delta}{\Delta+\delta}x_1 + \frac{\Delta}{\Delta+\delta}x_3$, hence

$$f\,(x_2) \le \frac{\delta}{\Delta + \delta}f\,(x_1) + \frac{\Delta}{\Delta + \delta}f\,(x_3)\,.$$

Subtracting$f(x_1)$ from both sides, it yields

$$f\,(x_2) - f\,(x_1) \le \frac{\Delta}{\Delta + \delta}\,(f\,(x_3) - f\,(x_1)) \le \frac{\Delta}{3\delta}\,|f\,(x_3) - f\,(x_1)| \le \frac{2c}{3\delta}\,||x_1 - x_2||\,.$$

If we define $x_4 = x_1 + (\delta/\Delta)\,(x_1 - x_2)$ and repeat the above processing, we can obtain$f\,(x_1) - f\,(x_2) \le \dfrac{2c}{3\delta}\,||x_1 - x_2||$. The two inequalities together imply

$$|f\,(x_1) - f\,(x_2)| \le \frac{2c}{3\delta}\,||x_1 - x_2||\,.$$ □

Lemma 1.3.1 leads to the following theorem directly.

Theorem 1.3.2 If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and dom $f = \mathbb{R}^n$, then f is a local Lipschitzian function on $\mathbb{R}^n$.

Proof Let $x \in \mathbb{R}^n$. Then there is a simplex $\Sigma = \mathrm{co}\,(x_1, x_2, \ldots, x_{n+1})$ which is convex by Definition 1.2.2, such that $x \in \mathrm{int}\,\Sigma$.

Because $x = \sum\limits_{i=1}^{n+1} \lambda_i x_i, \quad \lambda_i \geq 0, \quad \sum\limits_{i=1}^{n+1} \lambda_i = 1$

$$f(x) \leq \sum_{i=1}^{n+1} \lambda_i f(x_i) \leq \max_i |f(x_i)| .$$

Because the effective domain of f is $\mathbb{R}^n$, f is upper bounded. By Lemma 1.3.1, the theorem is verified. □

Remark From the proofs of Lemma 1.3.1 and Theorem 1.3.2, we obtain two properties for the convex functions.

(1) Let f be a convex function. If f is locally upper bounded, then it is also locally lower bounded; furthermore, f is locally bounded.
(2) Let f be a convex function. If f is bounded at a neighborhood of x_0, then f is continuous at x_0. Furthermore, a bounded convex function is continuous. □

1.3.2 Examples of Convex Functions

This subsection provides several convex functions which will be used frequently later.

In a normed space X, the norm is a mapping from X to $\mathbb{R}\,(\geq 0)$. Because

$$\|(1-\lambda)\,x_1 + \lambda x_2\| \leq (1-\lambda)\,\|x_1\| + \lambda\,\|x_2\| ,$$

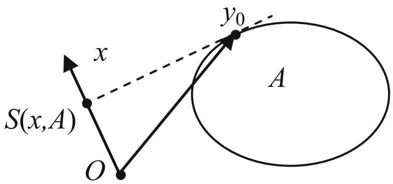
the norm is then a convex function on X.

If $V \in \mathbb{R}^{n\times n}$ is positive definite matrix, then we can define a norm for $\mathbb{R}^n$ by using V, i.e., $\|x\| = \sqrt{x^T V x}$. It follows from the above discussion that $\sqrt{x^T V x}$ is a convex function on $\mathbb{R}^n$. Moreover, $g(x) = x^2$ is a monotonously increasing and convex function when $x \geq 0$. We conclude, by the property (3) of convex function (Sect. 1.3), that a positive definite quadratic function $V(x) = x^T V x$ is convex. The fact can also be seen from the following differential criterion.

Supporting function $S(x, A)$. Let $A \subset \mathbb{R}^n$ be a set. A function $S(x, A) : \mathbb{R}^n \to \mathscr{R}$ is defined as follows:

$$S(x, A) = \sup_{x_a \in A} \langle x_a, x \rangle .$$

Fig. 1.4 The supporting function $S(x, A)$



$S(x, A)$ is called by the supporting function of set A. By the property (4) (Sect. 1.3) and the convexity of inner product, we then conclude that $S(x, A)$ is convex. It can verified that $S(x, A)$ is also a homogenous function.

Figure 1.4 gives a geometrical illustration for the supporting function $S(x, A)$. From the figure, $S(x, A)$ is the maximal projection among the points of cl A on the vector x. In other words, for a given $x \in \mathbb{R}^n$, we can define an $(n-1)$-dimensional hyperplane $\langle x, y\rangle = S(x, A)$, where x is the normal vector of the hyperplane such that the hyperplane is tangent with the closed set cl A. When A is a convex set, there is a unique $y_0 \in \text{cl } A$ such that $\langle x, y_0\rangle = S(x, A)$.

For a specific set A, the supporting function $S(x, A)$ can hold a specific meaning. For example, $A = \{y_0\}$, where $y_0 \in \mathbb{R}^n$, i.e., A only has one element, then $S(x, A) = \langle x, y_0\rangle$. $S(x, A)$ is a linear bounded functional. Moreover, when $\|y_0\| = 1$, then $S(x, A)$ is exact the projection of x on y_0. One more example is that if A is a closed unit ball of $\mathbb{R}^n$, then $S(x, A) = \|x\|$, i.e., the norm of vector x. The supporting function can be used to describe the property of a set; hence it is a useful tool in the following investigation. The readers can learn more specific supporting functions from the problems of this section.

Minkowski function $\mu(x)$. Let $A \subset \mathbb{R}^n$ be a convex set, and $0 \in \text{int} A$. The Minkowski function $\mu(x, A) : \mathbb{R}^n \to \mathbb{R}$ is defined as follows.

$$\mu(x, A) = \inf\left\{a > 0;\ \ a^{-1}x \in A\right\}.$$

Sometimes, we just denote $\mu(x, A)$ by $\mu(x)$ if A is unambiguous. We can prove that, for every $x \in \mathbb{R}^n$, $x \notin A$, we have $\infty > \mu(x) \geq 1$, for $x \in A$, $\mu(x) \leq 1$. It is also easy to verify that the $\mu(x)$ is positive homogeneous. Hence, to prove that $\mu(x)$ is a convex function, it is sufficient to show that $\mu(x + y) \leq \mu(x) + \mu(y)$.

Suppose that $a = \mu(x),\ \ b = \mu(y)$. For a given $\varepsilon > 0$, denote $c = a + b + \varepsilon$. Because $a + (\varepsilon/2) > \mu(x)$, we have $(a + (\varepsilon/2))^{-1}x \in A$. By the same reason, $(b + (\varepsilon/2))^{-1}y \in A$. Thus,

$$c^{-1}(x + y) = \frac{a + (\varepsilon/2)}{c}(a + (\varepsilon/2))^{-1}x + \frac{b + (\varepsilon/2)}{c}(b + (\varepsilon/2))^{-1}y \in A$$

since A is convex. From the definition of Minkowski function, we conclude

$$\mu(x + y) \leq c = a + b + \varepsilon = \mu(x) + \mu(y) + \varepsilon.$$

The inequality holds for all $\varepsilon > 0$, hence, $\mu(x + y) \leq \mu(x) + \mu(y)$.

1.3.3 Derivative of Convex Functions

We are already aware that a bounded and convex function is continuous. This subsection will deal with its differential property. We start with a criterion for the convexity of differentiable functions, then we make a deep discussion about the derivate of convex functions.

Let $A \subset X$ be a convex set and f: $\mathbb{R} \to \mathbb{R}$ be a differentiable function. Then the convexity of f can be determined by its first-order derivative or its second-order derivative. We present some equivalent conditions of convex functions with single variable. The proofs for these conclusions can be found in a textbook for convex analysis and are omitted.

1. f is a convex function on A.
2. $f'(x),\ x \in A$ is a monotone increasing function.
3. $f(x_2) \geq f(x_1) + f'(x_1)(x_2 - x_1)$ for $x_1, x_2 \in A$.
4. $f''(x) \geq 0$ for $x \in A$.

We now extend these conclusions to multivariable functions.

Theorem 1.3.3 Let $A \subset \mathbb{R}^n$ be a convex set, and $f : A \to \mathbb{R}$ be a function. Then the following statements are equivalent.

(1) f is a convex function on A.

(2) For two vectors $x_1, x_2 \in A$, $\langle x_2 - x_1,\ \nabla f(x_2) - \nabla f(x_1)\rangle \geq 0$, where ∇f is the gradient of f, i.e., $\nabla f^T = \begin{bmatrix} \dfrac{\partial f}{\partial x^1} & \dfrac{\partial f}{\partial x^2} & \cdots & \dfrac{\partial f}{\partial x^n} \end{bmatrix}$[9] where x^i is the i-th component of x.

(3) $f(x_2) \geq f(x_1) + \langle \nabla f(x_1),\ x_2 - x_1\rangle$ for $x_1, x_2 \in A$.

(4) $\nabla^2 f(x) \geq 0$,[10] for every $x \in A$, where $\nabla^2 f = \left[\partial^2 f / \partial x^i \partial x^j\right]$ is the Hessian matrix of f.

Proof By the remark given after Definition 1.3.1, it is sufficient to verify that f is convex on the segment $[x_1, x_2]$.

We consider now the function $\bar{f}(\lambda) = f(x_1 + \lambda(x_2 - x_1))$, where x_1 and x_2 are fixed, but $\lambda \in [0,\ 1]$ is a variable. Hence, $\bar{f}(\lambda)$ is a scalar function. By the derivate of compound function, we have

$$\begin{aligned} \frac{d\bar{f}(\lambda)}{d\lambda} &= \left\langle \left(\left.\frac{\partial f(x)}{\partial x}\right|_{x = x_1 + \lambda(x_2 - x_1)} \right), \frac{d(x_1 + \lambda(x_2 - x_1))}{d\lambda} \right\rangle \\ &= \langle \nabla f(x_1 + \lambda(x_2 - x_1)), (x_2 - x_1)\rangle . \end{aligned} \tag{1.3.3}$$

By the property (2) listed before Theorem 1.3.3, $\bar{f}(\lambda)$ is convex if and only if $d\bar{f}/d\lambda$ is monotonous increasing in the interval $\lambda \in [0, 1]$.

[9] In general, the gradient is a row vector, but we consider ∇f to be a column vector for unification.

[10] A matrix $M > 0$ means that the matrix is symmetric and positive definite. $M \geq 0$ means it is semi-positive definite.

(1)⇒(2) $f(x)$ is a convex function, so is $\bar{f}(\lambda)$. By property (2), $d\bar{f}/d\lambda$ is increasing with the increase of λ, then

$$\left.\frac{d\bar{f}}{d\lambda}\right|_{\lambda=1} = \nabla f(x_2)\cdot(x_2-x_1) \geq \nabla f(x_1)\cdot(x_2-x_1) = \left.\frac{d\bar{f}}{d\lambda}\right|_{\lambda=0}.$$

(2)⇒(1) If (2) is true, replacing x_1 by $x_1+\lambda_1(x_2-x_1)$ and x_2 by $x_1+\lambda_2(x_2-x_1)$ we have

$$\begin{aligned}0 &\leq \langle[\nabla f(x_1+\lambda_2(x_2-x_1))-\nabla f(x_1+\lambda_1(x_2-x_1))],(\lambda_2-\lambda_1)(x_2-x_1)\rangle\\ &=(\lambda_2-\lambda_1)\langle[\nabla f(x_1+\lambda_2(x_2-x_1))-\nabla f(x_1+\lambda_1(x_2-x_1))],(x_2-x_1)\rangle\\ &=(\lambda_2-\lambda_1)\left[\left.\frac{d\bar{f}(\lambda)}{d\lambda}\right|_{\lambda=\lambda_2}-\left.\frac{d\bar{f}(\lambda)}{d\lambda}\right|_{\lambda=\lambda_1}\right].\end{aligned}$$

The last equation comes from Eq. (1.3.4). Thus, $d\bar{f}/d\lambda$ is monotonously increasing, $\bar{f}(\lambda)$ is convex, so is $f(x)$.

(1) $\Longleftrightarrow$ (3) The equivalence is verified by using property (iii) of the scalar case. $\bar{f}(\lambda)$ is convex if and only if

$$\begin{aligned}f(x_1+\lambda_2(x_2-x_1)) &\geq f(x_1+\lambda_1(x_2-x_1))+\frac{d\bar{f}(x_1+\lambda_1(x_2-x_1))}{d\lambda}(\lambda_2-\lambda_1)\\ &=f(x_1+\lambda_1(x_2-x_1))+(\lambda_2-\lambda_1)\\ &\quad\times\Big\langle\nabla f(x_1+\lambda_1(x_2-x_1)),(x_2-x_1)\Big)\Big\rangle\\ &=f(x_1+\lambda_1(x_2-x_1))\\ &\quad+\Big\langle\nabla f(x_1+\lambda_1(x_2-x_1)),(\lambda_2-\lambda_1)(x_2-x_1)\Big)\Big\rangle.\end{aligned} \tag{1.3.4}$$

Let $\lambda_1=0,\ \lambda_2=1$. Then Inequality (1.3.5) leads to conclusion (3). If we define $\tilde{x}_1=x_1+\lambda_1(x_2-x_1)$ and $\tilde{x}_2=x_1+\lambda_2(x_2-x_1)$, and replace $x_i,\ i=1,2$ by $\tilde{x}_i,\ i=1,2$, respectively, in conclusion (3), then it is exactly equal to Inequality (1.3.5). The equivalence of (3) is verified.

(1) $\Longleftrightarrow$ (4) At last, derivating Eq. (1.3.4) both sides related to λ, by using the property (iv), the equivalence of conclusion (4) can be obtained. □

Because the Hessian matrix of $x^T Vx$ is $2V$, the quadratic function $x^T Vx$ is a convex function if and only if V is positive definite. Moreover, the quadratic function $x^T Vx+b^T x+c$ is also convex provided that V is positive definite and $b\in\mathbb{R}^n$ and $c\in\mathbb{R}$ are selected arbitrarily.

Definition 1.3.3 Suppose $f:\mathbb{R}^n\to\mathbb{R}(\infty)$, $\upsilon\in\mathbb{R}^n$ and $\lambda\in\mathbb{R}$. If the following limitation exists

$$\lim_{\lambda\downarrow 0}\frac{f(x_0+\lambda\upsilon)-f(x_0)}{\lambda}$$

(the limitation is allowed to be ±∞), where $\lambda \downarrow 0$ means λ decreases monotonously to zero, then the limitation is called by directional derivative of f at x_0 and along with the direction υ, and the directional derivative is denoted by $Df(x_0)(\upsilon)$. □

The notation $Df(x_0)(\upsilon)$ can be understood as a function, i.e., $\upsilon \in \mathbb{R}^n$ is the argument and x_0 is a parameter. The parameterized mapping $Df(x_0)$ maps υ into $\mathscr{R}$.

It was known that in calculus, there is also a definition of directional derivative, but the definition is somewhat different from Definition 1.3.3 where λ is convergent monotonously from right side.

From calculus we have known that if f is continuously differentiable at x_0, then $Df(x_0)(\upsilon) = \langle \nabla f(x_0), \upsilon \rangle$. For the convex functions, we have the following conclusion.

Theorem 1.3.4 Let $f : \mathbb{R}^n \to \mathbb{R}(\infty)$ be a convex function. Then for $x_0, \upsilon \in \mathbb{R}^n$ the directional derivative $Df(x_0)(\upsilon)$ exists. Moreover, for a given x_0, $Df(x_0)(\upsilon)$ is convex and positively homogeneous.

Proof Before proving the theorem, we verify a preliminary conclusion.

Suppose $f : \mathbb{R} \to \mathbb{R}$ is a convex function, and $x_0 < x_1 < x_2$, then

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} \le \frac{f(x_2) - f(x_0)}{x_2 - x_0} \le \frac{f(x_2) - f(x_1)}{x_2 - x_1}. \tag{1.3.5}$$

The meaning of Inequality (1.3.6) can be illustrated by Fig. 1.5 as follows.

Denote $\alpha = (x_2 - x_1)/(x_2 - x_0)$. If $x_0 < x_1 < x_2$, then $\alpha \in [0, 1]$, and $x_1 = \alpha x_0 + (1 - \alpha) x_2 . f(x)$ is a convex function, hence,

$$f(x_1) \le \frac{x_2 - x_1}{x_2 - x_0} f(x_0) + \frac{x_1 - x_0}{x_2 - x_0} f(x_2).$$

It is easy to show that the inequality is equivalent to $\frac{f(x_1) - f(x_0)}{x_1 - x_0} \le \frac{f(x_2) - f(x_0)}{x_2 - x_0}$. The another inequality in (1.3.6) can be verified by a similar way and is omitted.

When $f : \mathbb{R}^n \to \mathbb{R}(\infty)$ is restricted on the segment $[x_0, x_0 + \upsilon]$, $f : [x_0, x_0 + \upsilon] \to \mathbb{R}(\infty)$ is a scalar function, it can be rewritten as $\overline{f}(\lambda) = f(x_0 + \lambda\upsilon)$ where $\overline{f} : [0, 1] \to \mathbb{R}$ and $\overline{f}(\lambda)$ is convex. For $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$, with $1 \ge \lambda_2 > \lambda_1 >$

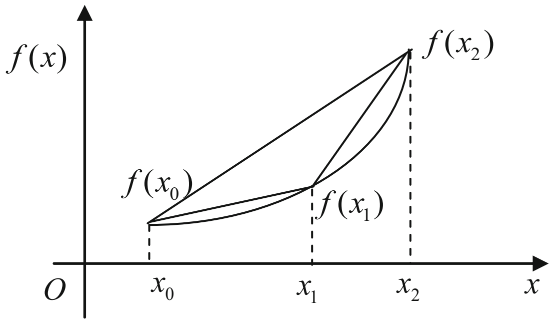


Fig. 1.5 An illustration of Inequality (1.3.5)

$\lambda_0 \geq 0$, Inequality (1.3.6) implies that $\frac{\bar{f}(\lambda_1)-\bar{f}(\lambda_0)}{\lambda_1-\lambda_0} \leq \frac{\bar{f}(\lambda_2)-\bar{f}(\lambda_0)}{\lambda_2-\lambda_0}$. If we fix $\lambda_0 = 0$, then the inequality leads to

$$\frac{f(x_0+\lambda_1 \upsilon)-f(x_0)}{\lambda_1} \leq \frac{f(x_0+\lambda_2 \upsilon)-f(x_0)}{\lambda_2}.$$

It implies that $\frac{f(x_0+\lambda \upsilon)-f(x_0)}{\lambda}$ is monotonously decreasing by $\lambda \downarrow 0$. Hence the limitation $\lim\limits_{\lambda \downarrow 0} \frac{f(x_0+\lambda \upsilon)-f(x_0)}{\lambda}$ exists (it may be $-\infty$), i.e., $Df(x_0)(\upsilon)$ exists.

If $a \in \mathbb{R}\,(> 0)$, then

$$Df(x_0)(a\upsilon) = \lim_{\lambda \downarrow 0} \frac{f(x_0+\lambda a\upsilon)-f(x_0)}{\lambda} = a \lim_{\lambda \downarrow 0} \frac{f(x_0+\lambda a\upsilon)-f(x_0)}{a\lambda} = aDf(x_0)(\upsilon),$$

i.e., $Df(x_0)$ is positively homogeneous.

At last, we prove it is convex. For $\mu \in [0, 1]$, we have

$$\begin{aligned} &\frac{f\Big(x_0+\lambda(\mu\upsilon_1+(1-\mu)\upsilon_2)\Big)-f(x_0)}{\lambda} \\ &= \frac{f(\mu(x_0+\lambda\upsilon_1)+(1-\mu)(x_0+\lambda\upsilon_2))-f(x_0)}{\lambda} \\ &\leq \frac{\mu\,[f((x_0+\lambda\upsilon_1)-f(x_0)\Big]}{\lambda} + \frac{(1-\mu)\,[f(x_0+\lambda\upsilon_2)-f(x_0)]}{\lambda}. \end{aligned}$$

Let $\lambda \downarrow 0$, the inequality results in

$$Df(x_0)(\mu\upsilon_1+(1-\mu)\upsilon_2) \leq \mu Df(x_0)(\upsilon_1)+(1-\mu)Df(x_0)(\upsilon_2).$$

The theorem is now completely verified. □

Theorem 1.3.4 reveals that convex functions hold similar property to continuously differentiable functions. Moreover, its directional derivative is convex. The following definition is about subdifferential, it can be treated as an extension of gradient defined for continuously differentiable functions.

Definition 1.3.4 Let $f : \mathbb{R}^n \to \mathbb{R}(\infty)$ be a convex function. Then the set

$$\partial f(x_0) = \{y; \forall x \in \mathbb{R}^n,\ f(x)-f(x_0) \geq \langle y, x-x_0\rangle\}$$

is defined as the subdifferential of f at x_0. If $\partial f(x_0) \neq \varnothing$, then f is subdifferentiable at x_0. □

It is worth to note that by the third conclusion of Theorem 1.3.3, if f is continuously differentiable at x_0, then its gradient $\nabla f(x_0) \in \partial f(x_0)$. Hence the subdifferential is an extension of gradient. Moreover, from the definition the subdifferential $\partial f(x_0)$ is a closed and convex set for every $x_0 \in \mathbb{R}^n$. The following theorem presents more property of subdifferential.

Theorem 1.3.5 Let $f : \mathbb{R}^n \to \mathbb{R}(\infty)$ be a convex function, $x_0 \in \mathbb{R}^n$ and $f(x_0) \neq \infty$. Then

$$\partial f(x_0) = \partial_\upsilon Df(x_0)(\upsilon)|_{\upsilon=0} = \partial_\upsilon Df(x_0)(0)$$

where ∂_υ is the subdifferential of $Df(x_0)(\upsilon)$ related to vector υ.[11]

Proof By the definition of directional derivative,

$$Df(x_0)(0) = \lim_{\lambda \downarrow 0} \frac{f(x_0 + \lambda 0) - f(x_0)}{\lambda} = 0.$$

And by the definition of subdifferential, for $\upsilon \in \mathbb{R}^n$, $\partial_\upsilon Df(x_0)(0) = \{y; Df(x_0)(\upsilon) \geq \langle y, \upsilon \rangle\}$.

If $y \in \partial f(x_0)$, then for any $\upsilon \in \mathbb{R}^n$, $f(x_0 + \lambda\upsilon) - f(x_0) \geq \langle y, \lambda\upsilon \rangle = \lambda \langle y, \upsilon \rangle$, i.e.,

$$\frac{f(x_0 + \lambda\upsilon) - f(x_0)}{\lambda} \geq \langle y, \upsilon \rangle .$$

Let $\lambda \downarrow 0$. It leads to $Df(x_0)(\upsilon) \geq \langle y, \upsilon \rangle$. Hence $y \in \partial_\upsilon Df(x_0)(0)$, or $\partial f(x_0) \subset \partial_\upsilon Df(x_0)(0)$. On the other hand, if $y \in \partial_\upsilon Df(x_0)(0)$, then $Df(x_0)(\upsilon) \geq \langle y, \upsilon \rangle$ by the definition of $\partial_\upsilon Df(x_0)(0)$. When $0 < \lambda < 1$, Inequality (1.3.6) implies that

$$\begin{aligned} f(x) - f(x_0) &= f(x_0 + (x - x_0)) - f(x_0) \\ &\geq \frac{f(x_0 + \lambda(x - x_0)) - f(x_0)}{\lambda} \\ &\geq Df(x_0)(x - x_0) \\ &\geq \langle y, x - x_0 \rangle . \end{aligned}$$

From Definition 1.3.4, we conclude $y \in \partial f(x_0)$, i.e., $\partial f(x_0) \supset \partial_\upsilon Df(x_0)(0)$. The proof is completed. □

Thus we can restate Definition 1.3.4 as follows.

$$\partial f(x_0) = \{y; \forall \upsilon \in \mathbb{R}^n, \ Df(x_0)(\upsilon) \geq \langle y, \upsilon \rangle\} . \tag{1.3.6}$$

Theorem 1.3.6 If $f, g : \mathbb{R}^n \to \mathbb{R}(\infty)$ are all convex functions, then the following statements are valid.

(1) $f(x_0) = \min\{f(x)\}$ if and only if $0 \in \partial f(x_0)$.
(2) $a \in \mathbb{R}(\geq 0)$, then $\partial a f(x_0) = a\partial f(x_0)$.

[11] The reader should understand the meaning of the notation $Df(x_0)(\upsilon)$. It is emphasized that in $Df(x_0)(\upsilon)$ $x_0 \in \mathbb{R}^n$ is treated as a parameter, $\upsilon \in \mathbb{R}^n$ is the argument, ∂_υ is the subdifferential related to the argument υ.

(3) If $x \in \operatorname{dom} f \cap \operatorname{dom} g$, then $\partial (f + g)(x) \supset \partial f(x) + \partial g(x)$, moreover, if there is an $x_0 \in \operatorname{dom} f \cap \operatorname{dom} g$ such that f or g is continuous at x_0, then

$$\partial (f + g)(x) = \partial f(x) + \partial g(x)$$

for all $x \in \operatorname{dom} f \cap \operatorname{dom} g$.

Proof

(1) By the definition of subdifferential, if $0 \in \partial f(x_0)$, then $f(x) - f(x_0) \geq \langle 0, x - x_0 \rangle = 0$, i.e., $f(x_0)$ reaches its minimum. On the other hand, if $f(x_0)$ is the minimum, then $f(x) \geq f(x_0)$, hence, $f(x) - f(x_0) \geq \langle 0,\ x - x_0 \rangle = 0$, or, $0 \in \partial f(x_0)$.

(2) If $a = 0$, then $af(x) \equiv 0$, $\partial af(x) \equiv 0$. The conclusion holds. Hence, we only consider the case that $a > 0$. If $y \in \partial af(x_0)$, then $af(x) - af(x_0) \geq \langle y, x - x_0 \rangle$. It follows $a^{-1}y \in \partial f(x_0)$, $y \in a\partial f(x_0)$. On the other hand, if $y \in a\partial f(x_0)$, then $a^{-1}y \in \partial f(x_0)$, hence, $f(x) - f(x_0) \geq \langle a^{-1}y, x - x_0 \rangle$, $y \in \partial af(x_0)$.

(3) If $x,\ \overline{x} \in \operatorname{dom} f \cap \operatorname{dom} g$ and $y_1 \in \partial f(x),\ y_2 \in \partial g(x)$ then by the definition of subdifferential, we have $f(\overline{x}) - f(x) \geq \langle y_1, \overline{x} - x \rangle$, and $g(\overline{x}) - g(x) \geq \langle y_2, \overline{x} - x \rangle$. Consequently, $f(\overline{x}) + g(\overline{x}) - (f(x) + g(x)) \geq \langle y_1 + y_2, \overline{x} - x \rangle$, i.e., $y_1 + y_2 \in \partial (f + g)(x)$, i.e., $\partial (f + g)(x) \supset \partial f(x) + \partial g(x)$.

To prove the second statement of the third conclusion, it is sufficient to show $\partial (f + g)(x) \subset \partial f(x) + \partial g(x)$. Without loss of generality, we suppose that f is continuous at x_0. Let $\overline{x} \in \operatorname{dom} f \cap \operatorname{dom} g$ and $\overline{y} \in \partial (f + g)(\overline{x})$. Then $f(x) + g(x) - (f(\overline{x}) + g(\overline{x})) \geq \langle \overline{y}, x - \overline{x} \rangle$, i.e.,

$$f(x) - f(\overline{x}) - \langle \overline{y}, x - \overline{x} \rangle \geq g(\overline{x}) - g(x). \tag{1.3.7}$$

Define now two relations as follows

$$\begin{aligned} F &= \{(x_F, r_F) \in \operatorname{dom} f \times \mathbb{R}, f(x_F) - f(\overline{x}) - \langle \overline{y}, x_F - \overline{x} \rangle < r_F\}, \\ G &= \{(x_G, r_G) \in \operatorname{dom} g \times \mathbb{R}, g(\overline{x}) - g(x_G) \geq r_G\}\ . \end{aligned} \tag{1.3.8}$$

If $x_F \in \operatorname{dom} f$, there is certainly a r_F such that $(x_F, r_F) \in F$, then $F \neq \varnothing$ and $\operatorname{dom} f = \operatorname{dom} F$. Moreover, for a fixed x_F, the r_F can go to infinite. Similarly, we can prove $G \neq \varnothing$ and $\operatorname{dom} g = \operatorname{dom} G$. And the r_G can trend to negative infinite.

From the fact that both $f(x)$ and $g(x)$ are convex functions, we can verify that F and G are all convex sets. Furthermore for $x \in \operatorname{dom} f \cap \operatorname{dom} g$, Inequality (1.3.8) implies there is no r such that $(x, r) \in F$ and $(x, r) \in G$, consequently, $F \cap G = \varnothing$. By using Theorem 1.2.3, we conclude that there is an $p \in \mathbb{R}^n,\ \alpha \in \mathbb{R}$, such that $\left(p^T\ \ \alpha\right) \neq 0$ and $(p^T\ \ \alpha)$ separates set F from set G, i.e., for every $(x_F, r_F) \in F$ and every $(x_G, r_G) \in G$, the following Inequality (1.3.10) holds

$$\langle x_F, p \rangle + \alpha r_F \leq \langle x_G, p \rangle + \alpha r_G. \tag{1.3.9}$$

As pointed out above, r_G can go to negative infinite, hence, $\alpha \leq 0$.

We now prove that $\alpha \neq 0$. Otherwise, if $\alpha = 0$, Inequality (1.3.10) leads to

$$\langle x_F,\ p\rangle \leq \langle x_G, p\rangle\,. \tag{1.3.10}$$

Because f is continuous at x_0. It means that there is a neighborhood $B\left(x_0,\varepsilon\right) \subset$ dom f. When $y \in B\left(0,\varepsilon\right),\ \ y \neq 0$, then $x_0 + y \in B\left(x_0,\varepsilon\right)$ and $x_0 - y \in B\left(x_0,\varepsilon\right)$. When $x_0 + y \in B\left(x_0,\varepsilon\right)$, Inequality (1.3.11) leads to $\langle x_0 + y,\ p\rangle \leq \langle x_0, p\rangle$, i.e., $\langle y,\ p\rangle \leq 0$. Similarly, by $x_0 - y \in B\left(x_0,\varepsilon\right)$, the same deduction leads to $\langle y,\ p\rangle \geq 0$. The both inequalities implies $\langle y,\ p\rangle = 0$. Since y can be selected arbitrarily from $B(0,\ \varepsilon)$, $p = 0$. Furthermore $\left(p^T\alpha\right) = 0$. It contradicts to $\left(p^T\ \ \alpha\right) \neq 0$. Therefore, $\alpha \neq 0$.

From Inequality (1.3.10), we obtain

$$\left\langle x_F, \frac{p}{|\alpha|}\right\rangle - r_F \leq \left\langle x_G, \frac{p}{|\alpha|}\right\rangle - r_G. \tag{1.3.11}$$

For every x_F and $\delta > 0, f\left(x_F\right) - f\left(\overline{x}\right) - \left\langle \overline{y}, x_F - \overline{x}\right\rangle + \delta$ are qualified to be the r_F. And for every x_G, $g\left(\overline{x}\right) - g\left(x_G\right)$ is qualified to be r_G. Thus, Inequality (1.3.12) results in

$$\left\langle x_F, \frac{p}{|\alpha|}\right\rangle - \left(f\left(x_F\right) - f\left(\overline{x}\right) - \left\langle \overline{y}, x_F - \overline{x}\right\rangle\right) + \delta \leq \left\langle x_G, \frac{p}{|\alpha|}\right\rangle - \left(g\left(\overline{x}\right) - g\left(x_G\right)\right). \tag{1.3.12}$$

Let $x_F = \overline{x}$. Then due to $\delta > 0$ Inequality (1.4.1) results in

$$g\left(x_G\right) - g\left(\overline{x}\right) \geq \left\langle x_G - \overline{x}, -\frac{p}{|\alpha|}\right\rangle,$$

for every $x_G \in$ dom G. Hence $-\frac{p}{|\alpha|} \in \partial g\left(\overline{x}\right)$. If $x_G = \overline{x}$, then Inequality (1.4.1) results in

$$f\left(x_F\right) - f\left(\overline{x}\right) \geq \left\langle x_F - \overline{x}, \overline{y} + \frac{p}{|\alpha|}\right\rangle,$$

for $x_F \in$ dom F. It implies $\overline{y} + (p/\ |\alpha|) \in \partial f\left(\overline{x}\right)$. Thus, $\overline{y} = \overline{y} + (p/\ |\alpha|) - (p/\ |\alpha|) \in \partial f\left(\overline{x}\right) + \partial g\left(\overline{x}\right)$. □

The second part of Conclusion 3 is also known as Moreau-Rockafellar theorem. For Theorem 1.3.6, we give two remarks.

Remark 1 The continuity of f at x_0 is only used to verify $\alpha \neq 0$. Hence, if we can find a vector $\left(p^T\ \ \alpha\right)$ with $\alpha \neq 0$ to separates F from G, then the condition of continuity can be removed. □

Remark 2 In general, we cannot guarantee that dom $f \cap$ dom g is convex. But we can guarantee the sets F and G defined in Eq. (1.3.9) are all convex. Defining the two sets is the critical issue in the proof. □

To end this subsection, we give a conclusion which reveals the relation of subdifferential $\partial f(x)$ and gradient $\nabla f(x)$.

Theorem 1.3.7 If f is differentiable at x_0, then $\partial f(x_0) = \{\nabla f(x_0)\}$.

Proof After Definition 1.3.3, we have mentioned that when f is differentiable at x_0, then $Df(x_0)(\upsilon) = \langle \nabla f(x_0), \upsilon \rangle$ for every $\upsilon \in \mathbb{R}^n$. By Eq. (1.3.7), if $y \in \partial f(x_0)$, then $\langle \nabla f(x_0), \upsilon \rangle \geq \langle y, \upsilon \rangle$. On the other hand, if υ is replaced by $-\upsilon$, then $\langle \nabla f(x_0), -\upsilon \rangle \geq \langle y, -\upsilon \rangle$, i.e., $\langle \nabla f(x_0), \upsilon \rangle \leq \langle y, \upsilon \rangle$. Consequently, $\langle \nabla f(x_0), \upsilon \rangle = \langle y, \upsilon \rangle$, i.e., $y = \nabla f(x_0)$. □

Problems

1. For a given $x \in \mathbb{R}^n$, try to construct a simplex $\Sigma = \mathrm{co}(x_1, x_2, \ldots, x_{n+1})$ such that $x \in \mathrm{int}\,\Sigma$.
2. Prove that the level set lev $(f < \alpha)$ is convex provided that f is a convex function.
3. Is the statement that "$f : \mathbb{R}^n \to \mathscr{R}$ is a convex function, if f is lower bounded at a neighborhood $B(x_0, \varepsilon)$ of $x_0 \in \mathbb{R}^n$, there is a $\delta > 0$ such that f is upper bounded at $B(x_0, \delta)$" true? Why?
4. Let $A \subset \mathbb{R}^n$ be closed and convex. Show that $a \in A$ if and only if $S(x, A) \geq \langle a, x \rangle$ for every $x \in \mathbb{R}^n$.
5. Prove that if a function $\mu(x)$ is positively homogeneous and $\mu(x+y) \leq \mu(x) + \mu(y)$, then $\mu(x)$ is a convex function.
6. Let $A \subset \mathbb{R}^n$ be a convex set, $x \in X$, $a = \pi(x, \mathrm{cl}A)$, denote $y = x - a/\|x - a\|$ show that $S(y, A) = \langle y, a \rangle$.
7. Let $A \subset \mathbb{R}^n$ be a compact set. Then $S(x, A)$ is a Lipschitzian function.
8. Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a convex function, G be a convex set in $\mathbb{R}^n \times \mathbb{R}^m$. Then the function $\phi(x)$ defined by $\phi(x) = \inf\limits_{y \in G(x)} f(x, y)$ is a convex function where $G(x) = \{y \in \mathbb{R}^m;\ (x, y) \in G\}$.
9. If A is the closed unit ball in $\mathbb{R}^n$, prove that $S(x, A) = \|x\|$. Moreover, prove that the conclusion is still valid if A is the open unit ball in $\mathbb{R}^n$.
10. Show that $\partial f(x_0)$ is a closed and convex set provided that f is a convex function.
11. Prove that if $f(x) = ||x||$, $x \in \mathbb{R}^n$, then $\partial f(x) = \begin{cases} B_n, & x = 0, \\ x/\,||x||, & x \neq 0. \end{cases}$
12. Prove that $\partial f(x) = \{x^*; x^* \in A, \langle x^*, x \rangle = f(x)\}$, where $f(x) = S(x, A)$, $A \subset \mathbb{R}^n$ and A is a convex set.
13. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a positively homogeneous and convex function. Then $f(x) = S(x, \partial f(0))$ provided that $f(x)$ is bounded for every x in $\mathbb{R}^n$.
14. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and is subdifferentiable. Then $\partial f(x)$ is a bounded set for every x in $\mathbb{R}^n$.
15. Prove that if $\partial f(x_0)$ holds only one element y, then f is differentiable at x_0 and $\nabla f(x_0) = y$.

1.4 Semi-continuous Functions

In the theory of convex analysis, the semi-continuous function is also a key issue and is deeply related to the convex function. In the next chapter, almost all set-valued mappings are semi-continuous. Hence, this section introduces semi-continuous functions for preparation. Readers are suggested to compare the semi-continuous function with those set-valued semi-continuous mappings, which will be defined in the next chapter, and to try find their similarity and difference. This section contains two parts: the definitions and properties of semi-continuous functions, and examples of semi-continuous functions which play useful roles in the following discussion.

1.4.1 Semi-continuous Functions and Their Elementary Properties

In this section, $\mathbb{R}^n$ is still treated as an inner space. $f : \mathbb{R}^n \to \mathscr{R}$ is a function and set $A \subset \mathbb{R}^n$.

Definition 1.4.1 $f : \mathbb{R}^n \to \mathscr{R}$ is a function, if $f(x_0) \geq \overline{\lim_{x\longrightarrow x_0}} f(x) = \lim_{x\longrightarrow x_0} \sup f(x)$,[12] then f is upper semi-continuous at x_0. If for every $x \in A$, f is upper semi-continuous at x, then f is an upper semi-continuous function on A.

$f : \mathbb{R}^n \to \mathscr{R}$ is a function, if $f(x_0) \leq \underline{\lim_{x\longrightarrow x_0}} f(x) = \lim_{x\longrightarrow x_0} \inf f(x)$, then f is lower semi-continuous at x_0. If for every $x \in A$, f is lower semi-continuous at x, then f is a lower semi-continuous function on A. □

Readers may find from Definition 1.4.1 that f is upper semi-continuous at x_0 if and only if -f is lower semi-continuous at x_0, and vice versa. It is also easy to verify from the definition that f is continuous at x_0 if and only if it is upper semi-continuous and lower semi-continuous at x_0, simultaneously.

The following equivalent statements are listed for upper semi-continuous functions, their proofs are omitted.

(1) f is upper semi-continuous at x_0.
(2) For every $\varepsilon > 0$, there is a $\delta > 0$ such that $f(x) \leq f(x_0) + \varepsilon$ provided that $x \in B(x_0, \delta) \cap A$.
(3) For every $a \in \mathbb{R}$, the level set lev $(f < a)$ is an open set.
(4) For every $a \in \mathbb{R}$, the level set lev $(f \geq a)$ is a closed set.

[12] The notation $\lim_{x\longrightarrow x_0} \sup f(x)$ should be understood as follows. Let δ be a small positive real number, and $a(\delta) = \sup_{x\in B(x_0,\delta)} f(x)$. Then $a(\delta)$ is a monotonously decreasing function with the decreasing of δ. Hence, the limitation $\lim_{\delta\downarrow 0} a(\delta)$ exists. Then $\lim_{x\longrightarrow x_0} \sup f(x) = \lim_{\delta\downarrow 0} a(\delta) = \inf_{\delta\downarrow 0} \sup_{x\in B(x_0,\delta)} f(x)$. The meaning of $\lim_{x\longrightarrow x_0} \inf f(x)$ is similar.

By the relation that f is upper semi-continuous if and only if $-f$ is lower semi-continuous. Hence, readers can give the corresponding properties for lower semi-continuous functions, similarly. We do not repeat them here.

Theorem 1.4.1 Suppose that the sequence $\{x_k\}$ satisfies the property that $x_k \to x_0,\ f(x_k) \to a$, i.e., the sequence and its corresponding function values are all convergent. Then $f : \mathbb{R}^n \to \mathbb{R}\,(\infty)$ is upper semi-continuous at $x_0 \in \mathbb{R}^n$ if and only if $a \le f(x_0)$ for all sequences $\{x_k\}$ which hold the property mentioned above.

Proof $x_k \to x_0$, hence there is a sequence $\{\delta_k\}$ with the property that $\delta_k \downarrow 0$ such that $x_k \in B(x_0,\ \delta_k)$. Consequently, $f(x_k) \le \sup\limits_{x\in B(x_0,\delta_k)} f(x)$. By Definition 1.4.1, $a = \lim\limits_{k\to\infty} f(x_k) \le \limsup\limits_{x_k\to x_0} f(x) \le f(x_0)$. The necessity is verified.

It is sufficient for the sufficiency to prove that there exists a sequence $\{x_k\}$ such that $x_k \to x_0$ and $f(x_k) \to \limsup\limits_{x\to x_0} f(x)$. If it is proved, then by the condition of Theorem 1.4.1$\limsup\limits_{x\to x_0} f(x) \le f(x_0)$, i.e., f is upper semi-continuous at $x_0 \in \mathbb{R}^n$. By the definition of supremum, for any $\delta_k > 0$ and $\varepsilon > 0$, there is a $x_k \in B(x_0, \delta_k)$, such that $\sup\limits_{x\in B(x_0,\delta_k)} f(x) - \varepsilon \le f(x_k) \le \sup\limits_{x\in B(x_0,\delta_k)} f(x)$. Let $\delta_k \downarrow 0$. Then $x_k \to x_0$ and

$$\lim_{k\to\infty} \sup_{x\in B(x_0,\delta_k)} f(x) - \varepsilon \le \lim_{k\to\infty} f(x_k) \le \lim_{k\to\infty} \sup_{x\in B(x_0,\delta_k)} f(x).$$

Because ε can be selected arbitrarily, $\lim\limits_{k\to\infty} f(x_k) = \lim\limits_{k\to\infty} \sup\limits_{x\in B(x_0,\delta_k)} f(x)$. The sufficiency is now verified. □

The corresponding conclusion of Theorem 1.4.1 for continuous function is known as Heine theorem in Calculus. The Heine theorem is very useful to prove that a function is not continuous. Thus, Theorem 1.4.1 is also useful in checking the semi-continuity.

Example 1.4.1 Consider the function defined by

$$f(x) = \begin{cases} \sin\frac{1}{x}, & x \ne 0, \\ a, & x = 0. \end{cases}$$

It has been analyzed in Calculus that the function is not continuous at $x = 0$ no matter the value defined for a. By using Theorem 1.4.1, we can see that if $a \ge 1$, then $f(x)$ is upper semi-continuous at $x = 0$; and if $a \le -1$, then $f(x)$ is lower semi-continuous at $x = 0$. It also shows that the $f(x)$ cannot be continuous at $x = 0$ by the selection of a. □

Theorem 1.4.2 A function $f : \mathbb{R}^n \to \mathbb{R}$ is lower semi-continuous if and only if its epigraph epi$f$ is a closed set of $\mathbb{R}^n \times \mathbb{R}$.

Proof Suppose $(x_k, \alpha_k) \in \mathrm{epi}\, f$ and $(x_k, \alpha_k) \to (x_0, \alpha_0)$. $(x_k, \alpha_k) \in \mathrm{epi}\, f, f(x_k) \le \alpha_k$ by the definition of epigraph. By Theorem 1.4.1, f is lower semi-continuous, hence

$$f(x_0) \le \lim_{x_k \to x_0} \inf f(x_k) \le \lim_{k \to \infty} \alpha_k = \alpha_0.$$

It follows that $(x_0, \alpha_0) \in \mathrm{epi}\, f$, i.e., epi$f$ is closed. The necessity is verified.

If epi$f$ is a closed set. We firstly prove that for any real number α, the level set $\mathrm{lev}\,(f \le \alpha)$ is closed. Let $\{x_k\} \subset \mathrm{lev}\,(f \le \alpha)$ be a convergent sequence, i.e., $x_k \to x_0$. Then $f(x_k) \le \alpha < \infty$. We can assume $f(x_k) \to c$,[13] then $c \le \alpha$. We deal with two cases.

(1) $c \ne -\infty$. We have $(x_k, f(x_k)) \to (x_0, c) \in \mathrm{epi}\ f\ (k \to \infty)$ since epi$f$ is closed. By the definition of epigraph, $f(x_0) \le c \le \alpha$, i.e., $x_0 \in \mathrm{lev}\,(f \le \alpha)$. $\mathrm{lev}\,(f \le \alpha)$ is closed. (2) $c = -\infty$. We then assert that $f(x_0) = -\infty$. If not, we have $f(x_0) = c_0 \ne -\infty$. Because $f(x_k) \to -\infty$, for sufficiently large $k, f(x_k) < c_0 - \varepsilon$ for some $\varepsilon > 0$. It implies $(x_k, c_0 - \varepsilon) \in \mathrm{epi}\ f$. $(x_k, c_0 - \varepsilon) \to (x_0, c_0 - \varepsilon)$, $(x_0, c_0 - \varepsilon) \in \mathrm{epi}\ f$ since epi$f$ is closed. By the definition of epigraph, $f(x_0) \le c_0 - \varepsilon$, i.e., $c_0 \le c_0 - \varepsilon$. A contradiction appears. It illustrates that $f(x_0) = -\infty$ and $x_0 \in \mathrm{lev}\,(f \le \alpha)$.

We then prove f is lower semi-continuous by using the conclusion that $\mathrm{lev}\,(f \le \alpha)$ is closed. Let $x_k \to x_0$ and $f(x_k) \to c$. Then for sufficiently large k, $c - \varepsilon < f(x_k) < c + \varepsilon$, and $x_k \in \mathrm{lev}\,(f \le c + \varepsilon)$. Because $\mathrm{lev}\,(f \le c + \varepsilon)$ is closed, $x_0 \in \mathrm{lev}\,(f \le c + \varepsilon)$, i.e., $f(x_0) \le c + \varepsilon$. It implies $\lim\limits_{x_k \to x_0} f(x_k) + \varepsilon \ge f(x_0)$. ε is selected arbitrarily, hence $\lim\limits_{x_k \to x_0} f(x_k) \ge f(x_0)$. Theorem 1.4.1 asserts f is lower semi-continuous. □

We are enforced to consider the case of lower semi-continuous functions in Theorem 1.4.2. Indeed, we can give a similar result for upper semi-continuous functions if we define the set $\{(x, c)\,; f(x) \ge c\}$ which is called hypograph of f. But we do not want to do so. As an exercise, readers can try to state and prove it.

In the proof of sufficiency, we exactly prove the properties 3 and 4 given for upper semi-continuous functions given before Theorem 1.4.1.

Theorem 1.4.3 Let $f : \mathbb{R}^n \to \mathbb{R}\,(-\infty)$ be an upper semi-continuous function. Then f has a finite upper boundary in a compact set $A \subset \mathbb{R}^n$.

Proof If f is boundless on A, then there is a sequence $\{x_k\} \subset A$, $f(x_k) \to \infty$. Because A is compact, $\{x_k\}$ has a convergent subsequence. Without loss of generality, we assume $\{x_k\}$ is convergent, i.e., $x_k \to x_0$ and $f(x_0) \ge \infty$ by Theorem 1.4.1. It contradicts to the condition of Theorem 1.4.3 that its range is $\mathbb{R}\,(-\infty)$. □

[13] If $f(x_k)$ is not convergent, then we can consider a subsequence $\{x_{k_j}\}$ such that $f(x_{k_j})$ is convergent. Hence for simplicity, we assume directly the sequence $f(x_k)$ is convergent. The skill will be used frequently and we shall not give any explanation below.

We know that a continuous function is bounded in a compact set. Theorem 1.4.3 can be treated as an extension of the conclusion. A upper semi-continuous function holds upper boundary in a compact set.

Theorem 1.4.4 Let $f : \mathbb{R}^n \to \mathbb{R}(-\infty)$ be a upper semi-continuous function. If f is bounded on the set $A \subset \mathbb{R}^n$, then there is a sequence of continuous functions $\{f_k(x),\ \ x \in A\}$ for every k such that $f_k(x) \downarrow f(x)$ for every $x \in A$.

Proof For every $k \in \mathbb{N}$, we construct a function $f_k(x)$ as follows

$$f_k(x) = \sup_{z \in A} \{f(z) - k \|z - x\|\}.$$

By the definition, it is clear that for every $x \in A, f_k(x) < \infty$ and

$$f_k(x) \geq f_{k+1}(x) \geq f(x). \tag{1.4.1}$$

Now we prove that $f_k(x)$ is continuous on A. For every $y \in A$, by the definition of $f_k(x)$, we have

$$f_k(x) \geq f(z) - k \|z - x\| \geq f(z) - k \|z - y\| - k \|y - x\|,$$

$$f_k(x) \geq \sup_{z \in A} \{f(z) - k \|z - y\|\} - k \|y - x\| = f_k(y) - k \|y - x\|.$$

By exchanging x and y, the above procedure leads to $f_k(y) \geq f_k(x) - k \|x - y\|$. Hence,

$$|f_k(x) - f_k(y)| \leq k \|y - x\|.$$

$f_k(x)$ is a Lipschitzian function; therefore, $f_k(x)$ is continuous on A.

Now we prove that for every $x_0 \in A,\ \ f_k(x_0) \downarrow f(x_0)$. The proof consists of two parts.

(1) $f(x_0) \neq -\infty$. f is upper semi-continuous on A, hence, for a given $\varepsilon > 0$, there exists a $\delta > 0$ such that if $x \in B(x_0, \delta)$, then $f(x) \leq f(x_0) + \varepsilon$. It follows that

$$f(x) - k \|x - x_0\| \leq f(x) \leq f(x_0) + \varepsilon. \tag{1.4.2}$$

We now prove that Inequality (1.4.3) also holds for $x \notin B(x_0, \delta)$. Let $M = \sup\limits_{x \in A} f(x)$. Then $M < \infty$ by the condition of Theorem 1.4.4. There exists an integer k_0, such that $f(x_0) > M - k_0\delta$. On the other hand, when $x \notin B(x_0, \delta), f(x) - k \|x - x_0\| \leq M - k\delta$. Hence, when $k > k_0, f(x) - k \|x - x_0\| \leq f(x_0)$.

We conclude for sufficient large $k, f(x) - k \|x - x_0\| \leq f(x_0) + \varepsilon$ for every $x \in A$. i.e., $f_k(x_0) \leq f(x_0) + \varepsilon$ for sufficient large k. Now combining with Inequality (1.4.2), $f_k(x_0) \downarrow f(x_0)$.

(2) $f(x_0) = -\infty$. There exists $\delta > 0$ such that for $x \in B(x_0, \delta), f(x) = f(x_0) = -\infty$ since f is upper semi-continuous. For every integer N, there exists k_0 such that $-N > M - k_0\delta$. Consequently, when $x \notin B(x_0, \delta)$ and $k > k_0, f(x) - k\|x - x_0\| \leq M - k\delta < -N$. We conclude that for sufficient large k, $f(x) - k\|x - x_0\| \leq -N$ for all $x \in A$. Therefore, $f_k(x_0) = \sup\limits_{x \in A}\{f(x) - k\|x - x_0\|\} \leq -N$, i.e., $f_k(x_0) \downarrow -\infty = f(x_0)$. □

We now list several operating properties for semi-continuous functions. We only deal with the upper semi-continuous functions and leave the discussion for lower semi-continuous functions to readers.

(1) If f_1 and f_2 are two upper semi-continuous functions on $\mathbb{R}^n$, then $f_1 + f_2$ and αf_1 are also upper semi-continuous functions provided that $\alpha \geq 0$;
(2) If f_1 and f_2 are two upper semi-continuous functions on $\mathbb{R}^n$, and $f_1 > 0, f_2 > 0$, then $f_1 \cdot f_2$ is also an upper semi-continuous function.
(3) If $F : \mathbb{R}^n \to \mathbb{R}^m$ is a continuous function and $g : \mathbb{R}^m \to \mathscr{R}$ is upper semi-continuous, then $g(F(x))$ is a upper semi-continuous function.
(4) If $F : \mathbb{R}^m \to \mathscr{R}$ is upper semi-continuous, and $g : \mathbb{R} \to \mathscr{R}$ is also upper semi-continuous. After extending the definition of g to $\tilde{g} : \mathscr{R} \to \mathscr{R}$ that $\tilde{g}(x) = g(x), \ x \in \mathbb{R}$ and $\tilde{g}(\infty) = \sup g(x), \tilde{g}(-\infty) = \inf g(x), \tilde{g}(F(x))$ is an upper semi-continuous function.
(5) $\inf\limits_i f_i$ is upper semi-continuous provided that $f_i, i = 1, 2, \ldots,$ are all upper semi-continuous.
(6) $\sup\limits_i f_i$ is upper semi-continuous provided that $f_i, i = 1, 2, \ldots, N$, are N upper semi-continuous functions where $N \in \mathbb{N}$.

1.4.2 Examples of Semi-continuous Functions

In this subsection, we deal with several semi-continuous functions which will be useful for the further investigation. In this subsection, X is treated as an inner space.

1. *Indictor function* $\delta(x, A)$. Let $A \subset X$ be a nonempty set, the indicator function of $A\delta(x, A)$: $X \to \mathbb{R}(\infty)$ is defined as

$$\delta(x, A) = \begin{cases} 0 & x \in A, \\ \infty & x \notin A. \end{cases}$$

The continuity of $\delta(x, A)$ is completely determined by the set A. If A is an open set, then $\delta(x, A)$ is upper semi-continuous; If A is a closed set, then $\delta(x, A)$ is lower semi-continuous. $\delta(x, A)$ is a convex function if A is a convex set.

2. *Supporting function* $S(x, A)$. The supporting function of a set A $S(x, A)$ has been defined at the last section where $S(x, A)$ is defined as $S(x, A) = \sup\limits_{y \in A} \langle y, x \rangle$. Because the inner function $f_y(x) = \langle y, x \rangle$ is continuous, by the property of upper

semi-continuous function listed at the last subsection, $S(x, A) = \sup_{y \in A} f_y$ is a lower semi-continuous. It can be proved that when A is a finite set then $S(x, A)$ is upper semi-continuous. Thus, it is continuous.

3. *Directional derivative* $Df(x)(\upsilon)$. The directional derivative $Df(x)(\upsilon)$ has been defined at the last subsection. We here give a detailed discussion.

Lemma 1.4.1 Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and positive homogeneous function. Then $f(x) = S(x, \partial f(0))$.

Proof Because f is a positive homogeneous function, $f(0) = 0$. By the definition of subdifferential, we have $\partial f(0) = \{x^*; f(x) \geq \langle x^*, x\rangle\}$. Consequently

$$f(x) \geq \sup_{x^* \in \partial f(0)} \langle x^*, x\rangle = S(x, \partial f(0)).$$

The opposite inequality is verified by contradiction. If there is a $x_0 \in \mathbb{R}^n$ such that $f(x_0) > S(x_0, \partial f(0))$, then $(x_0, S(x_0, \partial f(0))) \notin \operatorname{epi} f$. Moreover, by the definition of epigraph $(x_0, S(x_0, \partial f(0))) \notin \operatorname{cl} \operatorname{epi} f$.

Because f is a convex function, by Theorem 1.3.1, the epigraph of f is convex. Using the Remark 2 given after Lemma 1.2.2, there exists a $(y^*, \beta) \in \mathbb{R}^n \times \mathbb{R}$ and $\varepsilon > 0$ such that

$$\langle x, y^*\rangle + \alpha_x \cdot \beta < \langle x_0, y^*\rangle + \beta S(x_0, \partial f(0)) - \varepsilon \tag{1.4.3}$$

holds for every $(x, \alpha_x) \in \operatorname{epi} f$.

We now prove $\beta < 0$ and $y^* \in \partial f(0)$. Depending on the definition of supporting function, Inequality (1.4.4) is equivalent to the following one

$$\langle x, y^*\rangle + \alpha_x \cdot \beta < \langle x_0, y^*\rangle + \beta \cdot \sup_{x^* \in \partial f(0)} \langle x_0, x^*\rangle - \varepsilon. \tag{1.4.4}$$

Because α_x can trend to positive infinite, the validation of Inequality (1.4.5) requires $\beta \leq 0$. We assert that $\beta < 0$. Otherwise, if $\beta = 0$, Inequality (1.4.5) leads to $\langle x, y^*\rangle < \langle x_0, y^*\rangle - \varepsilon$. It is impossible if $x = x_0$. Without loss of generality, we can fix $\beta = -1$. Inequality (1.4.5) becomes

$$\langle x, y^*\rangle - \alpha_x < \langle x_0, y^*\rangle - \sup_{x^* \in \partial f(0)} \langle x_0, x^*\rangle - \varepsilon. \tag{1.4.5}$$

Let x be replaced by λx, and α_x be replaced by $f(\lambda x)$. Then Inequality (1.4.6) is

$$\langle \lambda x, y^*\rangle - f(\lambda x) = \lambda \big(\langle x, y^*\rangle - f(x)\big) < \langle x_0, y^*\rangle - \sup_{x^* \in \partial f(0)} \langle x_0, x^*\rangle - \varepsilon.$$

Let $\lambda \to \infty$. Then the above inequality implies $\langle x, y^*\rangle - f(x) \leq 0$ for every $x \in \operatorname{dom} f$. Hence $y^* \in \partial f(0)$.

Because $(0, 0) \in \operatorname{epi} f$, replacing (x, α_x) by $(0, 0)$, Inequality (1.4.6) leads to

$$\sup_{x^* \in \partial f(0)} \langle x_0, x^* \rangle < \langle x_0, y^* \rangle - \varepsilon. \tag{1.4.6}$$

Inequality (1.4.7) is not valid since $y^* \in \partial f(0)$. We then conclude for every $x \in \mathbb{R}^n$, $f(x) \leq S(x, \partial f(0))$.

The theorem is verified. □

Lemma 1.4.1 illustrates that a convex function is positive homogeneous then it has to be a supporting function of a convex set.

By Theorem 1.3.4, for $(x, \upsilon) \in \mathbb{R}^n \times \mathbb{R}^n$, the directional differential $Df(x)(\upsilon)$ exists. Moreover, for a fixed x, $Df(x)(\upsilon)$ is convex and positively homogeneous. Thus, we can define a function $Df(\bullet)(\circ) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by $(x, \upsilon) \mapsto Df(x)(\upsilon)$. The following theorem shows the function is upper semi-continuous.

Theorem 1.4.5 Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, then the function $(x, \upsilon) \mapsto Df(x)(\upsilon)$ is upper semi-continuous.

Proof The effective domain of f is $\mathbb{R}^n$; by the conclusion given in the Problem 6 of this section, we have

$$Df(x)(\upsilon) = S(\upsilon, \partial f(x)). \tag{1.4.7}$$

We now apply Theorem 1.4.1 to prove Theorem 1.4.5.

Suppose the sequence $\{(x_i, \upsilon_i)\}$ holds the properties that (1) $\{(x_i, \upsilon_i)\}$ is convergent to (x_0, υ_0); (2) $\{Df(x_i)(\upsilon_i)\}$ is also convergent to $\alpha \in \mathbb{R}$. If we can verify $\alpha \leq Df(x_0)(\upsilon_0)$, then Theorem 1.4.1 asserts Theorem 1.4.5 is true. The fact is proved by two steps.

(1) $\underset{i}{\cup} \partial f(x_i)$ is a bounded set, where x_i is the element in the domain of pair (x_i, υ_i).

By Eq. (1.4.8), $Df(x_i)(\upsilon_i) = S(\upsilon_i, \partial f(x_i)) = \sup\limits_{\bar{x} \in \partial f(x_i)} \langle \upsilon_i, \bar{x} \rangle$. Because $Df(x_i)(\upsilon_i)$ is finite for every pair (x_i, υ_i), $\partial f(x_i)$ is a bounded set. If $\underset{i}{\cup} \partial f(x_i)$ is boundless, then there $\bar{x}_i \in \partial f(x_i)$ for every i, such that $\bar{x}_i \to \infty$ $(i \to \infty)$. However,

$$Df(x_i)(\upsilon_i) = \sup_{\bar{x} \in \partial f(x_i)} \langle \upsilon_i, \bar{x} \rangle \geq \langle \upsilon_i, \bar{x}_i \rangle. \tag{1.4.8}$$

When $i \to \infty$, the left side of Inequality (1.4.9) is convergent by the selection of $\{(x_i, \upsilon_i)\}$, but the right side diverges. It is a contradiction. Hence, $\underset{i}{\cup} \partial f(x_i)$ is bounded.

(2) $\alpha \leq Df(x_0)(\upsilon_0)$. $\partial f(x_i)$ is a closed set (Problem 10, Sect. 1.3); hence there exists $\widehat{x}_i \in \partial f(x_i)$ such that $Df(x_i)(\upsilon_i) = \langle \upsilon_i, \widehat{x}_i \rangle$ for every i. $\underset{i}{\cup} \partial f(x_i)$ is bounded; consequently, $\{\widehat{x}_i\}$ has a convergent subsequence. Without loss of generality, we assume $\widehat{x}_i \to x^*$. Thus, $\alpha = \langle \upsilon_0, x^* \rangle$.

On the other hand, $\widehat{x_i} \in \partial f(x_i)$, by the definition of subdifferential, we have

$$f(y) - f(x_i) \geq \langle \widehat{x_i}, y - x_i \rangle \tag{1.4.9}$$

for any $y \in \mathbb{R}^n$. The theorem has assumed f is convex, so is locally bounded; hence f is continuous at x_0 by the Remark 2 of Lemma 1.3.1. Let $i \to \infty$, Inequality (1.4.10) yields

$$f(y) - f(x_0) \geq \langle x^*, y - x_0 \rangle ,$$

i.e., $x^* \in \partial f(x_0)$. By the definition of supporting function, we conclude

$$Df(x_0)(\upsilon_0) = S(\upsilon_0, \partial f(x_0)) \geq \langle \upsilon_0, x^* \rangle = \alpha.$$

The proof is completed. □

4. *Conjugate function* f^*. To end this section, we deal with the conjugate function for a semi-continuous function.

Definition 1.4.2 Let $f : \mathbb{R}^n \to \mathbb{R}(\infty)$ be a normal function.[14] $f^* : \mathbb{R}^n \to \mathbb{R}(\infty)$ is the conjugate function of f, if $f^*(x^*) = \sup_{x \in X} \{\langle x^*, x \rangle - f(x)\}$.[15] □

By the definition, f^* is always lower semi-continuous no matter what is the f. From the definition, we also have

$$f^*\left(x^*\right) + f(x) \geq \langle x^*, x \rangle . \tag{1.4.10}$$

Inequality (1.4.11) is known as Fenchel inequality.

As an example, let us consider the indicator function $\delta(x, A)$ where $A \subset \mathbb{R}^n$. By Definition 1.4.2, the conjugate function of $\delta(x, A)$ is

$$f^*\left(x^*\right) = \sup_{x \in \mathbb{R}^n} \{\langle x^*, x \rangle - \delta(x, A)\} = \sup_{x \in A} \langle x^*, x \rangle = S\left(x^*, A\right).$$

It shows that the conjugate function of the indicator function of a set A is the supporting function of the set. Now we prove that if A is a convex set, then the indicator function is also the conjugate function of supporting function for the identical set. We need a lemma to verify the conclusion.

Lemma 1.4.2 Let $\overline{B}$ be the closed unit ball of $\mathbb{R}^n$, and $A \subset \mathbb{R}^n$ is a convex set, then

$$d(x, A) = \sup_{y \in \overline{B}} \{\langle y, x \rangle - S(y, A)\} .$$

[14] Recall that a function is normal if its effective domain is nonempty.

[15] The argument of $f^*(x^*)$ is x^*, it is independent the x in $f(x)$.

Proof For a given $x \in \mathbb{R}^n$, if we denote $\rho_x = d(x, A)$, then $x \in \text{cl}\left(A + \rho_x \overline{B}\right)$. By Problem 4 of the last section, we have $\langle y, x\rangle \le S\left(y, \text{cl}\left(A + \rho_x \overline{B}\right)\right)$ for every $y \in \mathbb{R}^n$. Particularly, when $y \in \overline{B}$, we have

$$S\left(y, \text{cl}\left(A + \rho_x \overline{B}\right)\right) = \sup_{z \in \text{cl}\left(A + \rho_x \overline{B}\right)} \langle y, z\rangle \le \sup_{x_A \in A} \langle y, x_A\rangle + \sup_{b \in \overline{B}} \rho_x \langle y, b\rangle \le S(y, A) + \rho_x.$$

The inequality implies that $\rho_x \ge \langle y, x_A\rangle - S(y, A)$ for every $y \in \overline{B}$. Now we verify that $\rho_x = \sup\limits_{y \in \overline{B}} \{\langle y, x_A\rangle - S(y, A)\}$.

Denote $a = \pi(x, \text{cl}\, A)$, then $a \in$ cl A and $\rho_x = \|x - a\|$. We further denoted $y_0 = x - a / \|x - a\|$, then $y_0 \in \overline{B}$, and

$$\langle y_0, x\rangle - \langle y_0, a\rangle = \langle y_0, x - a\rangle = \left\langle \frac{x - a}{\|x - a\|}, x - a \right\rangle = \|x - a\| = \rho_x.$$

When $a \in A$, then the above equation implies $\rho_x = \langle y_0, x\rangle - \langle y_0, a\rangle$; if $a \in \text{cl}A$, then there is a sequence $\{a_k\} \subset A$, $a_k \to a \, (k \to \infty)$. The conclusion is proved. □

Consider now the conjugate function of $S(x^*, A)$. By the definition[16]

$$\begin{aligned} S^*(x, A) &= \sup_{x^* \in X} \{\langle x, x^*\rangle - S(x^*, A)\} \\ &= \sup_{x^* \in X} \cdot \|x^*\| \left\{\left\langle x, \tfrac{x^*}{\|x^*\|}\right\rangle - S\left(\tfrac{x^*}{\|x^*\|}, A\right)\right\}, \end{aligned}$$

Now we fix $\|x^*\| = \alpha$, then $S^*(x, A) = \sup\limits_{\alpha \in \mathbb{R}^+} \alpha d(x, A)$. Because α can be arbitrarily positive real number, we have

$$\sup_{\alpha \in \mathbb{R}^+} \alpha d(x, A) = \begin{cases} 0, & x \in A, \\ \infty, & x \notin A, \end{cases}$$

i.e., $S^*(x, A) = \sup\limits_{\alpha \in \mathbb{R}^+} \alpha d(x, A) = \delta(x, A)$. The equation holds due to the definition of indicator function.

The discussion on the indicator function $\delta(x, A)$ illustrates that when A is a convex set, then $\delta^{**}(x, A) = \delta(x, A)$. The relation of $f^{**}(x) = f(x)$ is not always true. The following theorem gives the condition for the validation of the equation.

By the definition of conjugate function,

$$f^{**}(x) = \sup_{x^*} \left\{\langle x, x^*\rangle - f^*\left(x^*\right)\right\}.$$

Using the Fenchel inequality, it is direct that $f^{**}(x) \le f(x)$.

[16] Here we denote x for x^{**}.

Theorem 1.4.6 If $f : \mathbb{R}^n \to \mathbb{R}(\infty)$ is a convex and lower semi-continuous function, then $f^{**}(x) = f(x)$.

Proof If $f(x) \equiv \infty$ for all $x \in X$, then by the definition of conjugate function,

$$f^*\left(x^*\right) = \sup_{x \in \mathbb{R}^n} \{\langle x, x^* \rangle - f(x)\} = -\infty.$$

It follows $f^{**}(x) \equiv \infty = f(x)$. Hence, we can assume that the effective domain $\operatorname{dom} f \neq \varnothing$.

We have mentioned the key issue is to prove $f^{**}(x) \geq f(x)$. We prove it by contradiction. If there is a $x_0 \in \operatorname{dom} f$ such that $f^{**}(x_0) < f(x_0)$. It follows that $(x_0, f^{**}(x_0)) \notin \operatorname{epi} f$, Theorem 1.4.2 asserts that $\operatorname{epi} f$ is a convex and closed set. Thus, there is $(y^*, \beta) \in \mathbb{R}^n \times \mathbb{R}$ and $\varepsilon > 0$ such that

$$\langle y^*, x \rangle + \beta \alpha_x < \langle y^*, x_0 \rangle + \beta f^{**}(x_0) - \varepsilon \tag{1.4.11}$$

for every $(x, \alpha_x) \in \operatorname{epi} f$, especially. If $\beta > 0$, by the property of epigraph, α_x can go to positive infinite, so is the left side of Inequality (1.4.12). It is a contradiction since the right side is independent of α_x. Therefore, $\beta \leq 0$. If $\beta = 0$, then Inequality (1.4.12) reduces to $\langle y^*, x \rangle < \langle y^*, x_0 \rangle - \varepsilon$. Take $x = x_0$, the inequality results in $0 < -\varepsilon$. It is impossible. Thus, we ensure $\beta < 0$. Dividing β from both sides, we have

$$\begin{aligned}\left\langle \tfrac{y^*}{|\beta|},\ x_0 \right\rangle - f^{**}(x_0) &> \sup_{(x,\alpha) \in \operatorname{epi} f} \left\{ \left\langle \tfrac{y^*}{|\beta|}, x \right\rangle - \alpha_x \right\} \\ &> \sup_{x \in \operatorname{dom} f} \left\{ \left\langle \tfrac{y^*}{|\beta|}, x \right\rangle - f(x) \right\} = f^*\left(\tfrac{y^*}{|\beta|}\right),\end{aligned}$$

Denote $\frac{y^*}{|\beta|} = x^*$, the inequality is just that $f^{**}(x_0) + f^*(x^*) < \langle x^*, x_0 \rangle$ which contradicts the Fenchel inequality. Theorem 1.4.6 is verified. □

The last conclusion is related to the subdifferential.

Theorem 1.4.7 Let $f : A \to \mathbb{R}(\infty)$ be a convex function. If $\operatorname{dom} f \neq \varnothing$ and $x \in \operatorname{dom} f$, then $f(x) + f^*(x^*) = \langle x^*, x \rangle$ if and only if $x^* \in \partial f(x)$.

Proof If $x^* \in \partial f(x)$, then by the definition of subdifferential $f(y) - f(x) \geq \langle x^*, y - x \rangle$ for all $y \in \operatorname{dom} f$. It follows that $-f(x) + \langle x^*, x \rangle \geq -f(y) + \langle x^*, y \rangle$. By again the definition of conjugate function, the yields inequality

$$f^*\left(x^*\right) = \sup_{y \in \operatorname{dom} f} \left(\langle x^*, y \rangle - f(y)\right) \leq \langle x^*, x \rangle - f(x). \tag{1.4.12}$$

Combining with Fenchel Inequality, we obtain $f(x) + f^*(x^*) = \langle x^*, x \rangle$.

If we start with Inequality (1.4.12), by the inverse procedure, we can obtain $f(y) - f(x) \geq \langle x^*, y - x \rangle$, i.e., $x^* \in \partial f(x)$. □

Theorems 1.4.6 and 1.4.7 present two conditions under which Fenchel inequality becomes equation. Theorem 1.4.6 asserts that when f is convex and lower semi-continuous then the equality holds; Theorem 1.4.7 proposes if f is a convex function, then the equation is only valid on the set of $\partial f(x)$. It is also a characteristic of $\partial f(x)$.

Problems

1. Prove that a function is continuous at x_0 and only if it is upper semi-continuous and lower semi-continuous at x_0.
2. Suppose $f : \mathbb{R} \to \mathscr{R}$ is a lower semi-continuous function, and its effective domain is nonempty. Let A be a bounded subset of $\mathbb{R}$. Then f holds lower boundary on A. If A is compact then f can reach its lower boundary.
3. Prove all equivalent statements given after Definition 1.4.1.
4. Prove all conclusions for the operations of upper semi-continuous functions given at the end of Sect. 1.4.1.
5. If $\operatorname{dom} f \neq \varnothing$ and f is convex and lower semi-continuous, then $\operatorname{dom} f^* \neq \varnothing$.
6. If f is a convex function and $x \in \text{int } (\operatorname{dom} f)$, then $Df(x)(\upsilon) = S(\upsilon, \partial f(x))$ (hint: using Lemma 1.4.1).
7. Let $f : A \to \mathbb{R}$ be a convex function. If $x_0 \in \operatorname{dom} f$ and f is continuous at x_0, then $f(x) = f^{**}(x)$.

References

Conway JB (1985) A course in functional analysis [M]. Springer, New York

de Bruim JCA, Doris A, van de Wouw et al (2009) Control of mechanical motion systems with non-collocation of actuation and friction: a Popov criterion approach for input-to-state stability and set-valued nonlinearities [J]. Automatica 45:405–415

Chapter 2
Set-Valued Mappings and Differential Inclusions

This chapter deals with two fundamental concepts for the control systems described by deferential inclusions. They are set-valued mappings and differential inclusions. The first two sections introduce the set-valued mappings and the succeeding four sections involve the differential inclusions.

2.1 Set-Valued Mappings

The set-valued mapping is an important foundation in the theory of differential inclusions. From the word of "set-valued", it is obvious that the mapping maps a variable to a set. Hence, the image of one variable may have more than one elements. In order to distinguish from the set-valued mapping, the mapping defined in Chap. 1, where the image of a variable only has one element, is often called by single-valued mapping in this chapter. This section will give the definition and elemental properties of set-valued mappings and extend the concepts of continuity and derivative to set-valued mappings.

2.1.1 Definition of the Set-Valued Mappings

X and Y are supposed to be two normed spaces. $\|\cdot\|_X$ and $\|\cdot\|_Y$ are the norms of X and Y, respectively. When there is no confusion, we often omit their subscripts.

Definition 2.1.1 $A \subset X$ is a set. A mapping $F : A \to Y$ is said to be a set-valued mapping defined on A, if for every $x \in A$, $F(x)$ is a subset of Y.

Z. Han et al., *Theory of Control Systems Described by Differential Inclusions*,
Springer Tracts in Mechanical Engineering, DOI 10.1007/978-3-662-49245-1_2

The set A is called by the domain of F. The set of $\{x;\ \ F(x) \neq \varnothing\}$ is called by the effective domain of F and is denoted by dom F. $F(x)$ is called by the image of x, the set $\underset{x\in A}{\cup} F(x) \subset Y$ is the range of F, and gra $F = \{(x, y)\,; y \in F(x)\}$ is called by the graph of F. □

From the definition, it seems to be better to define the set-valued mapping from A to the power set of Y, i.e., $F : A \to \mathscr{P}Y$. The definition can meet the requirement of single-valued mapping. The reason we have considered is that Y is a normed space, and we need to apply the norm of Y and topology induced by the norm.

From now on, capital letters, such as $F, G, \ldots$ are used to express the set-valued mappings; and correspondingly, the lowercases, such as $f,\ g,\ \ldots,$ are to express the single-valued mappings. We assume in this book that dom $F \neq \varnothing$ for every set-valued mapping F.

We now present more remarks for the set-valued mappings.

Remark 1 If the domain of set-valued mapping $F : A \to Y$ is not equal to X, i.e., $A \neq X$, then we can extend the definition of F to $\overline{F}$ by

$$\overline{F}(x) = \begin{cases} F(x) & x \in A, \\ \varnothing & x \notin A. \end{cases}$$

By the alteration, the domain of $\overline{F}$ becomes X, but its effective domain of is exactly equal to that of F. Hence we always assert that every set-valued mapping is defined on whole space X. □

Remark 2 For $y \in Y$, the inverse of y is defined as $\{x; y \in F(x)\}$ and is denoted by $F^{-1}(y)$. It is obvious that $F^{-1}(y)$ is a set of X. Let $G \subset Y$ be a set. There are two ways to define the inverse of G for a set-valued mapping F. The set $\{x; F(x) \subset G\} \subset X$ is defined as the strong inverse of G and denoted by $F^{-1}(G)\,(\mathrm{s})$ where the letter "s" means ***strong***. The set $\{x; F(x) \cap G \neq \varnothing\} \subset X$ is called by the weak inverse of G and denoted by $F^{-1}(G)\,(\mathrm{w})$ where "w" stands for ***weak***. It is obvious that $F^{-1}(G)\,(\mathrm{w}) \supset F^{-1}(G)\,(\mathrm{s})$. At the most time, we apply the definition of weak inverse, hence, $F^{-1}(G)\,(\mathrm{w})$ is simplified as $F^{-1}(G)$. But we keep the notation of $F^{-1}(G)\,(\mathrm{s})$. □

For these two kinds of inverses, we have the following equations and their proofs are left to readers as exercises.

$$F^{-1}\left(G^c\right)(\mathrm{w}) = \left[F^{-1}(G)\,(\mathrm{s})\right]^c, \tag{2.1.1a}$$

$$F^{-1}\left(G^c\right)(\mathrm{s}) = \left[F^{-1}(G)\,(\mathrm{w})\right]^c. \tag{2.1.1b}$$

In Eqs. (2.1.1a) and (2.1.1b), the complements at two sides are in different spaces. At the left side, $G^c = Y\backslash G$, but at the right side $\left[F^{-1}\left(\cdot\right)\right]^c = X\backslash F^{-1}\left(\cdot\right)$.

Remark 3 A set-valued mapping $F : X \to Y$ is with closed value if for every $x \in X$, $F(x)$ is a closed set of Y. If gra$F$ is a closed set of $X \times Y$, then $F(x)$ is called as a closed set-valued mapping or closed mapping for simplicity.

Replacing the word "closed" by "bounded", we can establish the definitions of "set-valued mapping with bounded value" and "bounded set-valued mapping". If it is replaced by "open" and "compact", we have the corresponding statements and the details are omitted. □

In fact, we have already applied the definition of set-valued mapping, for example, the epigraph epi$f$ is a graph of set-valued mapping, where the set-valued mapping is $F(x) = \{a; a \in \mathbb{R}, a \geq f(x)\}$. The another example is subdifferential $\partial f(x)$, it is also a set-valued mapping, although for some x it only has one element.

2.1.2 Continuities of Set-Valued Mappings

Giving definitions of continuities for set-valued mapping is a precise job. There are more than one kinds of continuities. Readers have to distinct and compare these statements, carefully.

Definition 2.1.2 Let $F : X \to Y$ be a set-valued mapping and $x_0 \in X$. F is upper semi-continuous at x_0 if for every open set $O_Y \subset Y$ which satisfies that $O_Y \supset F(x_0)$, there is a $\delta > 0$ such that $F(B(x_0, \delta)) \subset O_Y$. If F is upper semi-continuous at every point $x \in X$, then F is an upper semi-continuous set-valued mapping on X.

F is lower semi-continuous at x_0 if for every $y \in F(x_0)$ and every $\varepsilon > 0$, there is a $\delta > 0$ which may depend on y and ε such that for every $x \in B(x_0, \delta)$, $F(x) \cap B(y, \varepsilon) \neq \varnothing$. If F is lower semi-continuous at every point $x \in X$, then F is a lower semi-continuous set-valued mapping on X.

If F is both upper semi-continuous and lower semi-continuous at x_0, then F is continuous at x_0. If F is continuous at every point $x \in X$, then F is a continuous set-valued mapping on X. □

Note that in Definition 2.1.2, the words "the space X" can be replaced by "the set A". For example, we can say that F is an upper semi-continuous set-valued mapping on $A\ (\subset X)$.

For the upper semi-continuity and lower semi-continuity, we have the following equivalent statements.

Theorem 2.1.1 Let $F : X \to Y$ be a set-valued mapping. Then the following statements are equivalent.

(1) F is upper semi-continuous.
(2) For every open set $O_Y \subset Y$, $F^{-1}(O_Y)$ (s) is an open set of X.
(3) For every closed set $C_Y \subset Y$, $F^{-1}(C_Y)$ (w) is a closed set of X.
(4) Let $x \in X$ and $\{x_n\} \subset X$ be a sequence with $x_n \to x$. Then for every open set $O_Y \subset Y$ and $F(x) \subset O_Y$, there is an $N \in \mathbb{N}$, $F(x_n) \subset O_Y$ provided that $n > N$.

Proof (1) $\Rightarrow$ (2). Suppose that $O_Y \subset Y$ is an open set. If $x \in F^{-1}(O_Y)$ (s), then $F(x) \subset O_Y$. F is upper semi-continuous, by Definition 2.1.2, there is a neighborhood $B(x, \delta)$ such that $F(B(x, \delta)) \subset O_Y$, i.e., $B(x, \delta) \subset F^{-1}(O_Y)$ (s).

(2) $\Leftrightarrow$ (3). This is a direct result of Equations (2.1.1a) and (2.1.1b). The details are omitted.

(2) $\Rightarrow$ (4). Let $\{x_n\} \subset X$ be a sequence with $x_n \to x$. Let $O_Y \subset Y$ be an open set with $F(x) \subset O_Y$. Then by the conclusion (2), there is a $\delta > 0$, such that $B(x, \delta) \subset F^{-1}(O_Y)$ (s). Hence, there is an $N \in \mathbb{N}$, when $n > N$, $x_n \in B(x, \delta)$, i.e., $x_n \in F^{-1}(O_Y)$ (s), $F(x_n) \subset O_Y$.

(4) $\Rightarrow$ (1). If F is not upper semi-continuous at $x \in X$, then there is an open set $O_Y \subset Y$ such that $F(B(x, \delta)) \subset O_Y$ is not true. Let $\delta_n \downarrow 0$, there are an $x_n \in B(x, \delta_n)$, and a $y_n \in F(x_n)$, $y_n \notin O_Y$ for $n \in \mathbb{N}$. Obviously, $x_n \to x$. $y_n \in F(x_n)$, $y_n \notin O_Y$ for $n \in \mathbb{N}$. It contradicts to the conclusion (4). □

For the lower semi-continuity, we have the following corresponding conclusions.

Theorem 2.1.2 Let $F : X \to Y$ be a set-valued mapping. Then the following statements are equivalent.

(1) F is lower semi-continuous.
(2) For every open set $O_Y \subset Y$, $F^{-1}(O_Y)$ (w) is an open set of X.
(3) For every closed set $C_Y \subset Y$, $F^{-1}(C_Y)$ (s) is a closed set of X.
(4) Let $x \in X$ and $\{x_n\} \subset X$ be a sequence with $x_n \to x$. Then for every open set $O_Y \subset Y$, $F(x) \cap O_Y \neq \varnothing$, there is an $N \in \mathbb{N}$, $F(x_n) \cap O_Y \neq \varnothing$ provided that $n > N$. □

The proof of Theorem 2.1.2 is quite similar to that of Theorem 2.1.1 and is left to readers as an exercise.

There is an alternative way to define the continuity for the set-valued mappings. This, we call, is the $\varepsilon-$ continuity.

Definition 2.1.3 A set-valued mapping $F : X \to Y$ is $\varepsilon-$ upper semi-continuous at $x_0 \in X$, if for every $\varepsilon > 0$, there is a $\delta > 0$ which may depend on ε, such that for every $x \in B(x_0, \delta)$, $F(x) \subset F(x_0) + \varepsilon B$. If for every $x \in X$, F is $\varepsilon-$ upper semi-continuous at x, then F is an $\varepsilon-$ upper semi-continuous mapping on X.

A set-valued mapping $F : X \to Y$ is $\varepsilon-$ lower semi-continuous at $x_0 \in X$; if for every $\varepsilon > 0$, there is a $\delta > 0$ which may depend on ε, such that for every $x \in B(x_0, \delta)$, $F(x_0) \subset F(x) + \varepsilon B$. If for every $x \in X$, F is $\varepsilon-$ lower semi-continuous at x, then F is an $\varepsilon-$ lower semi-continuous mapping on X. □

It is easy to see that the relation $F(x) \subset F(x_0) + \varepsilon B$ for every $x \in B(x_0, \delta)$ is equivalent to $F(B(x_0, \delta)) \subset B(F(x_0), \varepsilon)_,$ and $F(x_0) \subset F(x) + \varepsilon B$ for every $x \in B(x_0, \delta)$ is equivalent to $F(x_0) \subset \bigcap\limits_{x \in B(x_0, \delta)} B(F(x), \varepsilon)$.

From $F(B(x_0, \delta)) \subset B(F(x_0), \varepsilon)$, it can see that if F is upper semi-continuous, then it is $\varepsilon-$ upper semi-continuous, since $B(F(x_0), \varepsilon)$ is an open set. But the inverse statement may fail to be true. A counter example will be given later.

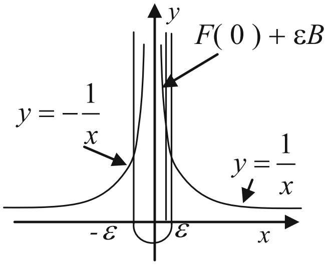


Fig. 2.1 The graph of $F(x)$ defined in Example 2.1.1

For the $\varepsilon-$ lower semi-continuous, the situation is different. The inclusion $F(x_0) \subset F(x) + \varepsilon B$ implies that for every $y_0 \in F(x_0)$, there is a $y \in F(x)$ such that $\|y - y_0\| < \varepsilon$, i.e., $F(x) \cap B(y_0, \varepsilon) \neq \varnothing$. On the other hand, if $F(x) \cap B(y_0, \varepsilon) \neq \varnothing$, then there is a $y \in F(x)$ such that $\|y - y_0\| < \varepsilon$, i.e., $F(x_0) \subset F(x) + \varepsilon B$. We conclude that the lower semi-continuity is equivalent to the $\varepsilon-$ lower semi-continuity.

Example 2.1.1 Consider $F : \mathbb{R} \to \mathbb{R}^2$ which maps every $x \in \mathbb{R}$ to the set $F(x) = \{(y_1, y_2); y_1 = x, y_2 \in [0, \infty)\}$, or, $F(x) = \{(x, y);\ x \in \mathbb{R},\ y \in [0, \infty)\}$ for simplicity.

In the plane of $\mathbb{R}^2$, for every x the image of $F(x)$ is a radical line which starts at $(x, 0)$ and goes to positive infinite and is orthogonal to the x-axis (Fig. 2.1). It follows that $\cup F(x)$ is the upper half plane of $\mathbb{R}^2$.

We assert that F is not upper semi-continuous at every $x \in \mathbb{R}$. As an example, let us consider $F(0)$ and an open set $M = \{(x, y);\ |xy| < 1\} \subset \mathbb{R}^2$. It is obvious that $M \supset F(0)$. However, for every $x,\ x \neq 0$, M cannot contain $F(x)$. It follows that for every $\delta > 0$, the relation $M \supset F(B(0, \delta))$ is not true, i.e., $F(x)$ is not upper semi-continuous at $x = 0$.

We can prove that for every $x \in \mathbb{R}$, F is lower semi-continuous at x. It is because that for every $(x, y_0) \in F(x)$ and every $\varepsilon > 0$, we can take $\delta = \varepsilon$, then $F(B(x, \delta)) \cap B((x, y_0), \varepsilon) = B((x, y_0), \varepsilon)$.

By Definition 2.1.3, the mapping $F(x)$ is $\varepsilon-$ upper semi-continuous. Because for every $\varepsilon > 0$, we can select $\delta = \frac{\varepsilon}{2}$, it follows that $F(x) = \{(x, y); y \in [0, \infty)\} \subset F(x_0) + \varepsilon B$ □

In Example 2.1.1, the open set M cannot be written as the form of $F(0) + \varepsilon B$. This is the key issue for the failure of $F(x)$ to be an upper semi-continuous mapping.

In Remark 3 after Definition 2.1.1, we have defined a set-valued mapping with closed value and closed set-valued mapping. By these definitions, the following conclusions are obvious.

(1) If $F : X \to Y$ is a set-valued mapping with open value, then it is an open set-valued mapping.
(2) If $F : X \to Y$ is a closed set-valued mapping, then it is with closed value.

Readers can find the inverse instatements of the above conclusion are not valid. By using the concepts of continuities, we have the following conclusion.

Theorem 2.1.3 If $F : X \to Y$ is an upper semi-continuous set-valued mapping, then it is closed if and only if it is with closed value.

Proof By the above statement (2), it is sufficient to show if $F : X \to Y$ is with closed value then it is a closed mapping.

Suppose $\{(x_n, y_n)\,,\ n = 1, 2, \ldots\} \subset \text{gra } F$ such that $(x_n, y_n) \to (x_0, y_0)$. We are required to verify $(x_0, y_0) \in \text{gra } F$. By the condition of Theorem 2.1.1, F is upper semi-continuous; hence, it is $\varepsilon-$ upper semi-continuous. For every $\varepsilon > 0$, there exists a $\delta > 0$, such that $F(x) \subset F(x_0) + \varepsilon B$ for every $x \in B(x_0, \delta)$. Because $x_n \to x_0$, there is an $N \in \mathbb{N}$, when $n > N$, $x_n \in B(x_0, \delta)$. It follows that $y_n \in F(x_n) \subset F(x_0) + \varepsilon B$ where $\varepsilon > 0$ can be selected arbitrary. Consequently, $y_0 \in \text{cl } F(x_0)$ since $y_n \to y_0$. F is with closed value, hence $F(x_0)$ is a closed set, $F(x_0) = \text{cl } F(x_0)$. Thus, $(x_0, y_0) \in \text{gra } F$ □

Readers are suggested to state a similar conclusion for lower semi-continuous set-valued mappings.

Before giving Theorem 2.1.4, we prove a general result which can be treated as an extension of separation axiom in Topology.

Lemma 2.1.1 Let Y be a Banach space, $A \subset Y$ is a compact set, and $M \supset A$ is an open set. Then there is an $\varepsilon > 0$ such that $M \supset A + \varepsilon B$.

Proof M is an open set, hence its complement M^c is a closed set. $M^c \cap A = \varnothing$. Now consider the distance between M^c and A, $d(M^c, A)$. We conclude that $d(M^c, A) > 0$ since A is compact and M^c is closed. Let $\varepsilon = \frac{1}{2} d(M^c, A)$. $M^c \cap (A + \varepsilon B) = \varnothing$, i.e., $M \supset A + \varepsilon B$. □

Theorem 2.1.4 If the set-valued mapping $F : X \to Y$ is with compact value, and Y is complete, then F is upper semi-continuous if and only if it is $\varepsilon-$ upper semi-continuous.

Proof It is sufficient to show that under the conditions of Theorem 2.1.4, an $\varepsilon-$ upper semi-continuous set-valued mapping is upper semi-continuous.

By the condition given by the theorem, for every $x \in X$, $F(x)$ is a compact set. Now let M be an open set which contains $F(x)$. Then by Lemma 2.1.1, there is an $\varepsilon > 0$ such that $F(x) + \varepsilon B \subset M$. $F(x)$ is $\varepsilon-$ upper semi-continuous, hence there is a $\delta > 0$, such that for every $x_1 \in B(x, \delta)$, $F(x_1) \subset F(x) + \varepsilon B$. It follows $F(B(x, \delta)) \subset F(x) + \varepsilon B \subset M$. □

In Example 2.1.1, $F(x)$ is $\varepsilon-$ upper semi-continuous, but for every $x \in \mathbb{R}$, $F(x)$ is not a compact set. Hence, it fails to be upper semi-continuous.

We now turn to the operation of two mappings.

For two set-valued mappings $F(x)$ and $G(x)$, because their images are in a normed space, we can define linear operation $\alpha F(x)+\beta G(x)$. But at the most time, we prefer to define $F(x) \cup G(x)$ and $F(x) \cap G(x)$.

Theorem 2.1.5

(1) If $F(x)$ and $G(x)$ are upper semi-continuous, then $F(x) \cup G(x)$ is upper semi-continuous;
(2) If $F(x)$ and $G(x)$ are lower semi-continuous, then $F(x) \cup G(x)$ is lower semi-continuous.

Proof The theorem is verified by using the conclusion 4 in Theorem 2.1.1 and Theorem 2.1.2, respectively. Let $O_Y \subset Y$ be an open set. Let $\{x_n\} \subset X$ be a sequence and $x_n \to x$.

(1) If $O_Y \supset F(x) \cup G(x)$, then $O_Y \supset F(x)$. It follows from the upper semi-continuity of $F(x)$, $O_Y \supset F(x_n)$ for some N_1 and $n > N_1$. Similarly, there is an N_2 such that for $n > N_2$, $O_Y \supset G(x_n)$. Thus, when $n > \max\{N_1, N_2\}$, $O_Y \supset F(x_n) \cup G(x_n)$. The first conclusion is verified.

(2) If $O_Y \cap (F(x) \cup G(x)) \neq \varnothing$, then without loss of generality, we assume that $O_Y \cap F(x) \neq \varnothing$. If follows from the lower semi-continuity, $O_Y \cap F(x_n) \neq \varnothing$ for some N_1 and $n > N_1$. Thus, $O_Y \cap (F(x_n) \cup G(x_n)) \neq \varnothing$ when $n > N_1$. The second conclusion is also proved.

The situation for the intersections is more complicated. We present them in the following two theorems.

Theorem 2.1.6 Let $F(x)$ and $G(x)$ be two set-valued mappings from X to Y. If they are all lower semi-continuous and satisfy that:

(1) F and G have all convex values, i.e., $F(x)$ and $G(x)$ are convex sets for every $x \in X$.
(2) $G(x)$ is locally bounded, i.e., for every $x_0 \in X$, there are $\delta > 0$ and $M > 0$ which may depend on x_0 such that if $x \in B(x_0, \delta)$ then $F(x) \subset MB_Y$ where B_Y is the open unit ball of Y.
(3) There exists a $\gamma > 0$, such that for every $x \in X$ and every $z \in \gamma B_Y$, $F(x) \cap (G(x) + z) \neq \varnothing$.

Then $F(x) \cap G(x)$ is lower semi-continuous.

Proof Let $x_0 \in X$ and $y_0 \in F(x_0) \cap G(x_0)$. Because $G(x)$ is locally bounded, there are $\delta_1 > 0$ and $M > 0$ such that $G(x) \subset MB_Y$ for every $x \in B(x_0, \delta_1)$. For a given $\varepsilon > 0$, define $\eta = \frac{\varepsilon}{1+4M\gamma^{-1}}$. By the lower semi-continuity of $F(x)$ and $G(x)$, there exists a $\delta \leq \delta_1$, such that $F(x) \cap B(y_0, \eta) \neq \varnothing$, $G(x) \cap B(y_0, \eta) \neq \varnothing$ simultaneously for $x \in B(x_0, \delta)$.

Take a point $g(x) \in G(x) \cap B(y_0, \eta)$ for every $x \in B(x_0, \delta)$. The $g(x) \in B(y_0, \eta)$ and $F(x) \cap B(y_0, \eta) \neq \varnothing$, hence there is a $f(x) \in F(x)$ such that $\|g(x) - f(x)\| < 2\eta$, i.e., $g(x) \in F(x) + 2\eta z$ for some $z = z(x) \in B_Y$ (Fig. 2.2).

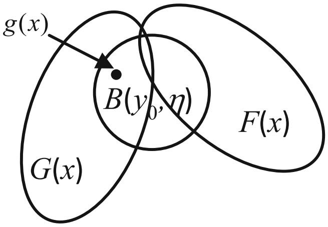


Fig. 2.2 The relation among $F(x)$, $G(x)$ and $B(y_0, \eta)$

By the third condition of Theorem 2.1.6, for this $z = z(x)$, there are $\overline{f}(x) \in F(x)$ and $\overline{g}(x) \in G(x)$, such that $\overline{g}(x) = \overline{f}(x) - \gamma z$. Now, we define $\lambda = \frac{\gamma}{\gamma+2\eta}$, it follows $1 - \lambda = \frac{2\eta}{\gamma+2\eta} = \frac{2\lambda\eta}{\gamma}$. At last

$$\begin{aligned}\lambda g(x) \in \lambda F(x) + 2\lambda\eta z &= \lambda F(x) + \tfrac{2\lambda\eta}{\gamma}\left(\overline{f}(x) - \overline{g}(x)\right)\\ &\subset \lambda F(x) + (1-\lambda)\,F(x) - (1-\lambda)\,\overline{g}(x)\ .\end{aligned}$$

Because both $F(x)$ and $G(x)$ are convex, $\lambda g(x) + (1-\lambda)\,\overline{g}(x) \in F(x) \cap G(x)$. On the other hand,

$$\begin{aligned}\|\lambda g(x) + (1-\lambda)\,\overline{g}(x) - y_0\| &\le \lambda\,\|g(x) - y_0\| + (1-\lambda)\,\|\overline{g}(x) - y_0\|\\ &\le \lambda\eta + (1-\lambda)\,2M\\ &= \lambda\eta + \frac{4M\lambda\eta}{\gamma}\\ &= \lambda\eta\left(1 + \frac{4M}{\gamma}\right)\\ &\le \varepsilon,\end{aligned}$$

$\lambda g(x) + (1-\lambda)\,\overline{g}(x) \in B\,(y_0, \varepsilon)$. Thus, $(F(x) \cap G(x)) \cap B\,(y_0, \varepsilon) \neq \varnothing$∘ □

This is a constructive proof, it gives an element in $(F(x) \cap G(x)) \cap B\,(y_0, \varepsilon)$. The next theorem is related to upper semi-continuous mappings.

Theorem 2.1.7 Let $F : X \to Y$ be an upper semi-continuous mapping. If the following conditions are satisfied:

(1) F is with compact value.
(2) $\operatorname{gra} G$ is a closed set in $X \times Y$.

Then $F(x) \cap G(x)$ is upper semi-continuous on X.

Proof Let $O_Y \subset Y$ be an open set and $O_Y \supset F\,(x_0) \cap G\,(x_0)$ for $x_0 \in X$. Then the target of verification is to find a $\delta > 0$ such that $O_Y \supset F(x) \cap G(x)$ for every $x \in B\,(x_0, \delta)$.

If $O_Y \supset F(x_0)$, then by the upper semi-continuity of $F(x)$ at x_0, there is a $\delta > 0$, such that $O_Y \supset F(x) \supset F(x) \cap G(x)$ for every $x \in B(x_0, \delta)$. The conclusion is proved. Hence it is sufficient to prove the case that $F(x_0) \not\subset O_Y$. If $F(x_0) \not\subset O_Y$, then $F(x_0) \cap O_Y^c \neq \varnothing$. Denote $K = F(x_0) \cap O_Y^c$, K is a compact set. We conclude that $K \cap G(x_0) = \varnothing$. Otherwise $G(x_0) \cap F(x_0) \cap O_Y^c \neq \varnothing$, it is contradicted to $O_Y \supset F(x_0) \cap G(x_0)$. Therefore, for every $y \in K$, $(x_0, y) \notin$ gra G. gra$G$ is a closed set, there exist $\delta_y > 0$ and $\varepsilon_y > 0$ such that $\left(B\left(x_0, \delta_y\right) \times B\left(y, \varepsilon_y\right)\right) \cap$ gra $G = \varnothing$. These $B(y, \varepsilon_y)$'s construct a open covering of K. There is a finite subcovering of K, i.e., $\bigcup\limits_{i=1}^{n} B\left(y_i, \varepsilon_{y_i}\right) \supset K$. Denote $M = \bigcup\limits_{i=1}^{n} B\left(y_i, \varepsilon_{y_i}\right)$, M is an open set. Recall the construction of $B(y, \varepsilon_y)$, every ε_{y_i} corresponds to a δ_{y_i}. Let $\delta_1 = \min\left\{\delta_{y_i}; i = 1, 2, \ldots, n\right\}$. For every $x \in B(x_0, \delta_1)$, $(\{x\} \times M) \cap$ gra $G = \varnothing$, i.e., $M \cap G(x) = \varnothing$.

Because $M \supset K = F(x_0) \cap O_Y^c$, $O_Y \cup M \supset F(x_0)$. F is upper semi-continuous, for the open set $M \cup O_Y$, there is a δ_2 such that for every $x \in B(x_0, \delta_2)$, $O_Y \cup M \supset F(x)$. Let $\delta = \min(\delta_1, \delta_2)$. When $x \in B(x_0, \delta)$, it happens $O_Y \cup M \supset F(x)$ and $M \cap G(x) = \varnothing$. Thus,

$$\begin{aligned} F(x) \cap G(x) &\subset (M \cup O_Y) \cap G(x) \\ &= (M \cap G(x)) \cup (O_Y \cap G(x)) \\ &= O_Y \cap G(x) \subset O_Y. \end{aligned}$$

We have prove that $O_Y \supset F(x) \cap G(x)$ always holds in a neighborhood of x_0. □

In Sect. 1.4, we have mentioned that f is a single-valued upper semi-continuous function if and only if $-f$ is lower semi-continuous. But for set-valued mappings, there is no such a simple relation.

We now turn to deal with the Lipshitzian property of a set-valued mappings.

Definition 2.1.4 $F : X \to Y$ is a set-valued mapping and $x_0 \in X$. If there are two positive constants L and δ such that

$$F(x) \subset F(x_0) + L\|x - x_0\|_X B_Y \tag{2.1.2}$$

for every $x \in B(x_0, \delta)$ where B_Y is the open unit ball in Y. Then F is a locally Lipschitzian set-valued mapping at x_0 and L is its Lipschitzian constant.

F is said to be a global Lipschitzian set-valued mapping at x_0, if the relation (2.1.2) is valid for all $x \in X$. Moreover, F is a (global) Lipschitzian set-valued mapping, if (2.1.2) is valid for any $x, x_0 \in X$. □

If the relation (2.1.2) is satisfied for $F : X \to Y$, then for $y \in F(x)$, there exist $y_0 \in F(x_0)$ and $b \in B_Y$ such that $y = y_0 + L\|x - x_0\| b$, it is equivalent to

$$\|y - y_0\| = L\|x - x_0\| \|b\| < L\|x - x_0\|. \tag{2.1.3}$$

Conversely, if Inequality (2.1.3) holds for any two vectors $y \in F(x)$ and $y_0 \in F(x_0)$, then the relation (2.1.2) holds. Hence Inequality (2.1.3) is an equivalent condition for Lipschitzian set-valued mappings.

It is obvious that if $F : X \to Y$ is a local Lipschitzian set-valued mapping at $x_0 \in X$, then F is $\varepsilon-$ upper semi-continuous at x_0. The following example shows the Lipschitzian property of a set-valued mapping does not imply the upper semi-continuity. This is an important difference from the single-valued mapping.

Example 2.1.2 Let $F : \mathbb{R} \to \mathbb{R}^2$ be as follows: for $x_0 \in \mathbb{R}$, its image is

$$F(x_0) = \left\{(x, y)\,;\,(x - x_0)^2 + y^2 < 1\right\}.$$

It is obvious for every $x_0 \in \mathbb{R}$, $F(x_0)$ is an open set. Hence, let $O_Y = F(x_0)$, then there is no $\delta > 0$ such that $F(B(x_0, \delta)) \subset O_Y$. However, F is a global Lipschitzian set-valued mapping, its Lipschitzian constant can be 1. □

Theorem 2.1.8 Let $F : X \to Y$ be a Lipschitzian set-valued mapping, its Lipschitzian constant is L. Let $\overline{F}(x) : X \to Y$ be an extension of F with $\overline{F}(x) = \overline{\text{co}}\ F(x)$. Then $\overline{F}$ is still a Lipschitzian set-valued mapping whose Lipschitzian constant is still L.

Proof Let $x_0 \in X$ and consider $\overline{F}(x_0) = \overline{\text{co}} F(x_0) \subset Y$. By the property of closed hull, for $y_0 \in \overline{F}(x_0)$ and arbitrarily $\varepsilon > 0$, there exist $y_{i0} \in F(x_0)$, $i = 1, 2, \ldots, s$ and λ_i, $i = 1, 2, \ldots, s$ with $\lambda_i \geq 0$ and $\sum_{i=1}^{s} \lambda_i = 1$ such that $\left\|y_0 - \sum_{i=1}^{s} \lambda_i y_{i0}\right\| < \varepsilon$.

We now consider $x \in X$, and take s vectors y_i from $F(x)$, i.e., $y_i \in F(x)$, $i = 1, 2, \ldots, s$. We have $\left\|y_0 - \sum_{i=1}^{s} \lambda_i y_i\right\| \leq \left\|y_0 - \sum_{i=1}^{s} \lambda_i y_{i0}\right\| + \left\|\sum_{i=1}^{s} \lambda_i y_{i0} - \sum_{i=1}^{s} \lambda_i y_i\right\| \leq \varepsilon + \sum_{i=1}^{s} \lambda_i \|y_{i0} - y_i\| < L\|x_0 - x\| + \varepsilon \varepsilon$ can be selected arbitrarily, consequently, $y_0 \in \overline{\text{co}}\ F(x) + L\|x - x_0\| B_Y$. □

Theorem 2.1.8 illustrates the Lipschitzian property can be extended to its convex hull, and the Lipschitzian constant is invariant.

2.1.3 Tangent Cones and Normal Cones

Before considering the derivative of a set-valued mapping, we deal with the tangent cone and normal cone for a set-valued mapping.

1. Definition of cones

Recall the concept of the cone which is mentioned when we introduce the convex sets. A set K is called by a cone if $x \in K$ then $\lambda x \in K$ for every $\lambda \in \mathbb{R}\,(\geq 0)$. K is

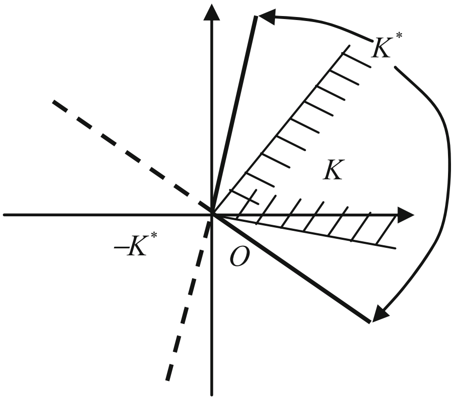


Fig. 2.3 Cone K and conjugate cone K^* and polar cone $-K^*$

said to be a convex cone if and only if for $x, y \in K$ and arbitrarily $\lambda, \mu \in \mathbb{R} (\geq 0)$ such that $\lambda x + \mu y \in K$.

By the definition, it is easy to verify if K_1 and K_2 are two cones, then $K_1 \cup K_2$, $K_1 \cap K_2$, $K_1 + K_2$ and $K_1 \backslash K_2$ are also cones. If K_1 and K_2 are two convex cones, then $K_1 \cap K_2$ and $K_1 + K_2$ are convex cones.

Let K be a cone, the conjugate cone of K is denoted by K^*

$$K^* = \{x^*;\ \langle x^*, x\rangle \geq 0, \forall x \in K\}.$$

Moreover, $-K^*$ is called by the polar cone of K. For simplicity, we often apply N to denote a polar cone. It is obvious that both K^* and $-K^*$ are convex and closed no matter what is the cone K. Figure 2.3 gives an illustration of relations of K, K^* and $-K^*$ in $\mathbb{R}^2$.

By the definition of conjugate cone, it is direct to verify that $(K_1 + K_2)^* = K_1^* \cap K_2^*$ for two cones K_1 and K_2.

In order to be convenient to application later, we deal with the support function of a cone. Let $K \subset \mathbb{R}^n$ be a cone. Then it support function $S(x, K)$ is defined by $S(x, K) = \sup\limits_{x_K \in K} (\langle x, x_K\rangle)$. When $x^* \in -K^*$, then $\langle x^*, x_K\rangle \leq 0$ for every $x_K \in K$. Since for all $\alpha \geq 0$, $\alpha x_K \in K$, $S(x^*, K) = \sup\limits_{x_K \in K} (\langle x^*, \alpha x_K\rangle) \to 0$, $(\alpha \to 0)$. When $x^* \notin -K^*$, there is a $x_K \in K$, such that $\langle x^*, x_K\rangle > 0$, then $S(x^*, K) = \sup\limits_{x_K \in K} (\langle x^*, \alpha x_K\rangle) \to \infty$, $(\alpha \to \infty)$. Therefore, we can conclude that

$$S\left(x^*, K\right) = \delta\left(x^*, -K^*\right). \tag{2.1.4}$$

The fact can be explained form Fig. 2.3

The following is a fundamental result for a cone and its polar cone.

Theorem 2.1.9 Let $K \subset X$ be a closed and convex cone, N be the polar cone of K. $x \in X$ is a vector and $\pi(x, K)$ is the projection of x on K. Denote $\pi(x, N) = x - \pi(x, K)$. Then we have the following conclusions

(1) $\pi(x, N) \in N$.
(2) If $x \notin N$, then $\pi(x, N)$ is the projection of x on N.
(3) $\|x\|^2 = \|\pi(x, K)\|^2 + \|\pi(x, N)\|^2$.

Proof (1) By Inequality (1.2.13), for every $y \in K$, we have

$$\langle x - \pi(x, K), y - \pi(x, K)\rangle \le 0. \tag{2.1.5}$$

K is a cone, hence $0 \in K$, and by the definition of projection $\pi(x, K) \in K$, replacing y by 0 and $2\pi(x, K) \in K$, respectively, Inequality (2.1.5) yields, $\langle x - \pi(x, K), \pi(x, K)\rangle \ge 0$ and $\langle x - \pi(x, K), \pi(x, K)\rangle \le 0$. Consequently,

$$\langle x - \pi(x, K), \pi(x, K)\rangle = 0. \tag{2.1.6}$$

Also by Eq. (2.1.6), Inequality (2.1.5) results in $\langle x - \pi(x, K), y\rangle \le 0$ for every $y \in K$. It is just that $\pi(x, N) = x - \pi(x, K) \in N$ by the definition of polar cone. The first conclusion is verified.

(2) Let $\upsilon^* \in N$. We conclude that

$$\langle \upsilon^* - \pi(x, N), x - \pi(x, N)\rangle = \langle \upsilon^*, x - \pi(x, N)\rangle - \langle \pi(x, N), x - \pi(x, N)\rangle \le 0.$$

It is because $\pi(x, K) = x - \pi(x, N)$, the first inner product is less than or equal to zero, and the second inner product is zero by Eq. (2.1.6). By Remark 2 of Lemma 1.2.2, $\pi(x, N)$ is the projection of x on N.

(3) We have

$$\begin{aligned}\|x\|^2 &= \|x - \pi(x, K) + \pi(x, K)\|^2 \\ &= \|x - \pi(x, K)\|^2 + \|\pi(x, K)\|^2 + 2\langle x - \pi(x, K), \pi(x, K)\rangle \\ &= \|x - \pi(x, K)\|^2 + \|\pi(x, K)\|^2 .\end{aligned}$$

The conclusion 3 is now verified. □

2. Tangent cone and normal cone

In order to define the derivative for set-valued mappings, we consider tangent cone and normal cone of a set.

Definition 2.1.5 Let $A \subset X$ be a set and $x \in A$. The set

$$T(x, A) = \mathrm{cl} \bigcup_{\lambda > 0} \frac{A - x}{\lambda} \tag{2.1.7}$$

is the tangent cone of A at x; $-T^*\,(x, A)$, the polar cone of $T(x, A)$, is defined as the normal cone of A at x and denoted by $N(x, A)$. □

Remark $T(x, A)$ is a cone. At first $x \in A$, hence $0 \in T\,(x, A)$. Moreover, if $y \in T\,(x, A)$, and there is a $\lambda > 0$ such that $y \in \frac{A-x}{\lambda}$, it follows that $\mu y \in \frac{A-x}{\lambda\mu^{-1}}$, for every $\mu > 0$. For vectors in derived set of $\underset{\lambda>0}{\cup}\dfrac{A-x}{\lambda}$, the proof is similar and omitted. □

If $X = \mathbb{R}^n$ and $x \in$ int A, then $T(x, A)= \mathbb{R}^n$. It is a trivial case. Hence it is only meaningful to deal with the tangent cones at the boundary of A.

Lemma 2.1.1 Let A be a convex set, and two real numbers λ and μ satisfy $0 < \lambda < \mu$. Then for $x \in A$, $\frac{A-x}{\mu} \subset \frac{A-x}{\lambda}$.

Proof Let $y \in A$ be an arbitrarily element of A. Define $z = \left(1 - \frac{\lambda}{\mu}\right) x + \frac{\lambda}{\mu} y$, then $z \in A$ since A is convex. It is equivalent to that $y = \frac{\mu}{\lambda} z + \left(1 - \frac{\mu}{\lambda}\right) x$. Thus,

$$\frac{y-x}{\mu} = \mu^{-1}\left(\frac{\mu}{\lambda} z - \frac{\mu}{\lambda} x + x - x\right) = \frac{z-x}{\lambda} \in \frac{A-x}{\lambda}.$$

□

Theorem 2.1.10 Let A be a convex set, and $x \in A$. Then

$$T\,(x, A) = \left\{ \upsilon; \lim_{\lambda \downarrow 0} \lambda^{-1} d\,(x + \lambda\upsilon, A) = 0 \right\}.$$

Before proving the theorem, we give an explanation. $d\,(x + \lambda\upsilon, A)$ is the distance of point $x + \lambda\upsilon$ to set A. The theorem asserts that, firstly, the tangent cone is a set-valued mapping and its effect domain is A; secondly, if $d\,(x + \lambda\upsilon, A)$ is a higher infinitesimal when λ intends to zero, then the υ belongs to the tangent cone. It meets well the definition of tangent line in calculus.

The proof of Theorem 2.1.10: If $\upsilon \in T\,(x, A)$, then by the Definition of 2.1.5, for every $\varepsilon > 0$, there is a positive real λ_0 such that $\upsilon \in \frac{A-x}{\lambda_0} + \varepsilon B_X$ where B_X is the unit ball of X. By Lemma 2.1.1, it follows that $\upsilon \in \frac{A-x}{\lambda} + \varepsilon B_X$ for every λ with $0 < \lambda \leq \lambda_0$. For the λ, we have

$$\lambda^{-1} d\,(x + \lambda\upsilon, A) = \lambda^{-1} \inf_{y\in A} \|x + \lambda\upsilon - y\| = \inf_{y\in A} \left\|\upsilon - \lambda^{-1}\,(y - x)\right\| \leq \varepsilon. \quad (2.1.8)$$

Because $\varepsilon > 0$ can be selected arbitrarily, $\lambda^{-1} d\,(x + \lambda\upsilon, A) \to 0$ as $\lambda \downarrow 0$.

Conversely, if $\lim\limits_{\lambda\downarrow 0} \lambda^{-1} d\,(x + \lambda\upsilon, A) = 0$, then, for every $\varepsilon > 0$, there is a $\delta > 0$, when $\lambda < \delta$, we have Inequality (2.1.8). The last inequality means $\upsilon \in \mathrm{cl}\frac{A-x}{\lambda} + \varepsilon B_X$, i.e., $\upsilon \in T\,(x, A)$. □

Theorem 2.1.11 Let A be a convex set and $x \in A$. Then

$$N\,(x, A) = \{\upsilon^*; \langle \upsilon^*, y - x\rangle \leq 0, \ y \in A\}.$$

Proof By Definition 2.1.5, $A - x \subset T(x, A)$, and by the definition of normal cone $N(x, A) = -T^*(x, A)$, we have $\langle \upsilon^*, y - x \rangle \leq 0$ for every $y \in A$ and every $\upsilon^* \in N(x, A)$. Therefore, $N(x, A) \subset \{\upsilon^*; \langle \upsilon^*, y - x \rangle \leq 0, \ y \in A\}$.

Conversely, let $\upsilon \in T(x, A)$, then for $\varepsilon > 0$ there is a $\lambda > 0$ such that $\upsilon \in \frac{A-x}{\lambda} + \varepsilon B$. It can be rewritten by $\left\| \upsilon - \frac{y-x}{\lambda} \right\| \leq \varepsilon$ for some $y \in A$. Now, if $\langle \upsilon^*, y - x \rangle \leq 0$ holds for every $y \in A$, then

$$\begin{aligned}\langle \upsilon^*, \upsilon \rangle &= \left\langle \upsilon^*, \upsilon - \frac{y-x}{\lambda} + \frac{y-x}{\lambda} \right\rangle \\ &= \left\langle \upsilon^*, \upsilon - \frac{y-x}{\lambda} \right\rangle + \left\langle \upsilon^*, \frac{y-x}{\lambda} \right\rangle \\ &\leq \left\langle \upsilon^*, \upsilon - \frac{y-x}{\lambda} \right\rangle \\ &\leq \varepsilon \left\| \upsilon^* \right\|.\end{aligned}$$

We have applied Schwarz Inequality at the last step. Since the ε can be selected arbitrarily, $\langle \upsilon^*, \upsilon \rangle \leq 0$ for every $\upsilon \in T(x, A)$. It follows $\upsilon^* \in N(x, A)$. The conclusion is verified. □

The last theorem in this subsection reveals the relation of $T(x, A)$ and $N(x, A)$, when they are treated as set-valued mappings.

Theorem 2.1.12 If $T(x, A)$ is compact and convex for every $x \in A \subset X$, then gra $N(x, A)$ is closed if and only if $T(x, A)$ is lower semi-continuous.

Proof The proof of the theorem applies the conclusion given in Problem 2 of this section. Let $(x_n, \upsilon_n^*) \in$ gra $N(x, A)$ and $(x_n, \upsilon_n^*) \to (x, \upsilon^*)$. If we can prove $\upsilon^* \in N(x, A)$, then gra $N(x, A)(x \in A)$ is closed. Let $\upsilon \in T(x, A)$. $T(x, A)$ is lower semi-continuous, hence there is a $\upsilon_n \in T(x_n, A)$, $\upsilon_n \to \upsilon$. It follows that $\langle \upsilon_n, \upsilon_n^* \rangle \leq 0$. $0 \geq \lim_{n\to\infty} \langle \upsilon_n^*, \upsilon_n \rangle = \langle \upsilon^*, \upsilon \rangle$. Thus, $\upsilon^* \in N(x, A)$.

For a vector $\upsilon \in T(x, A)$, and a sequence $\{x_n\}$ with $x_n \to x$, we should construct a sequence $\{\upsilon_n\}$ such that $\upsilon_n \in T(x_n, A)$ and $\upsilon_n \to \upsilon$. Because $\upsilon \in T(x, A)$, and $T(x, A)$ is convex and compact, by using Theorem 2.1.9, we conclude there are $\upsilon_n = \pi(\upsilon, T(x_n, A)) \in T(x_n, A)$ and $\upsilon_n^* = \pi(\upsilon, N(x_n, A)) \in N(x_n, A)$ such that $\upsilon = \upsilon_n + \upsilon_n^*$. Thus, $\left\| \upsilon_n^* \right\| \leq \|\upsilon\|$, $\{\upsilon_n^*\}$ has a convergent subsequence, without loss of generality, we can assume $\{\upsilon_n^*\}$ is convergent, $\upsilon_n^* \to \upsilon^*$. gra $N(x, A)$ is closed, hence $\upsilon^* \in N(x, A)$. Furthermore, $\upsilon_n = \upsilon - \upsilon_n^* \to \upsilon - \upsilon^*$. If we can prove $\upsilon^* = 0$, then $\upsilon_n \to \upsilon$. The conclusion is verified.

By Eq. (2.1.6), $\langle \upsilon^*, \upsilon - \upsilon^* \rangle = \lim_{n\to\infty} \langle \upsilon_n^*, \upsilon - \upsilon_n^* \rangle = 0$, i.e.,

$$0 = \langle \upsilon^*, \upsilon - \upsilon^* \rangle = \langle \upsilon^*, \upsilon \rangle - \langle \upsilon^*, \upsilon^* \rangle .$$

Hence, $\|\upsilon^*\|^2 = \langle \upsilon^*, \upsilon^* \rangle = \langle \upsilon^*, \upsilon \rangle \leq 0, \|\upsilon^*\| = 0.$ □

2.1.4 Derivative of Set-Valued Mappings

This subsection starts with the discussion of tangent cone. At the most time, we restricted ourselves in $\mathbb{R}^n$.

Definition 2.1.6 Let $A \subset X$, and $x \in A$. The set defined as

$$T_+ (x, A) = \left\{ \upsilon; \lim_{\lambda \downarrow 0} \lambda^{-1} d (x + \lambda \upsilon, A) = 0 \right\}$$

is called by the tangent cone of A at the x.

The set

$$T_- (x, A) = \left\{ \upsilon; \lim_{\lambda \downarrow 0} \inf \lambda^{-1} d (x + \lambda \upsilon, A) = 0 \right\}$$

is called by the contingent cone of A at the x. □

Recall the definition of limitation inferior of a function, if $\lim_{x \to x_0} \inf f(x) = c$ then (1) there exists a convergent $\{x_n\}$ sequence $x_n \to x_0$, such that $f(x_n) \to c$; (2) let $\{x_n\}$ be a convergent sequence, $x_n \to x_0$ and $f(x_n) \to \overline{c}$, then $\overline{c} \geq c$.

Remark 1 By Definition 2.1.6, it is obvious that both $T_+ (x, A)$ and $T_- (x, A)$ are closed and $T_+ (x, A) \subset T_- (x, A)$. □

Remark 2 If A is a convex set, then $\lambda^{-1} d (x + \lambda \upsilon, A)$ is decreasing with the decrease of λ. Hence, Theorem 2.1.10 concludes that the limitation $\lim_{\lambda \downarrow 0} \lambda^{-1} d (x + \lambda \upsilon, A)$ exists for every $\upsilon \in A$, and $T_+ (x, A) = T (x, A)$. Furthermore, $T_+ (x, A) = T_- (x, A)$. □

Example 2.1.3 Consider a single-valued mapping

$$f(x) = \begin{cases} |x| \sin \frac{1}{|x|}, & x \neq 0, \\ 0, & x = 0, \end{cases}$$

and $A = \cup \text{epi } f(x)$. Then the tangent cone $T_+ (0, A)$ and the contingent cone $T_- (0, A)$ are shown in Fig. 2.4. □

Let $A \subset \mathbb{R}^n$ and $\alpha > 0$. Define a set $D_\alpha \subset \mathbb{R}^n \times \mathbb{R}$ as follows

$$D_\alpha = \left\{ \left(x, x^0\right) \in \mathbb{R}^n \times \mathbb{R},\ x^0 \geq \alpha d (x, A) \right\}.$$

In $\mathbb{R}^n \times \mathbb{R}$, the set D_α looks as a solid basin (Fig. 2.5). Hence D_α is called by the basin of A with the slope α.

Theorem 2.1.13 Let $A \subset \mathbb{R}^n$, $\alpha > 0$ and $\hat{x} \in A$. Then

$$T_+ ((\hat{x}, 0), D_\alpha) \supset \left\{ \left(v, v^0\right) \in \mathbb{R}^n \times \mathbb{R},\ v^0 \geq \alpha d (v, T_+ (\hat{x}, A)) \right\}.$$

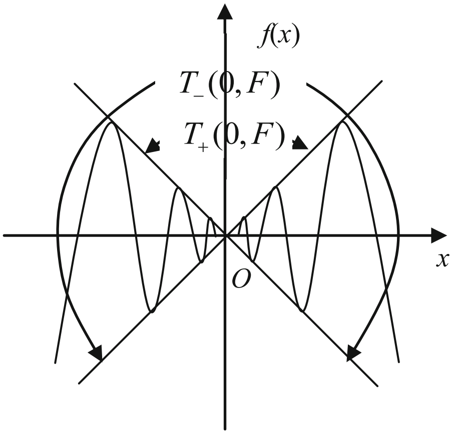


Fig. 2.4 The tangent cone and contingent cone of Example 2.1.3

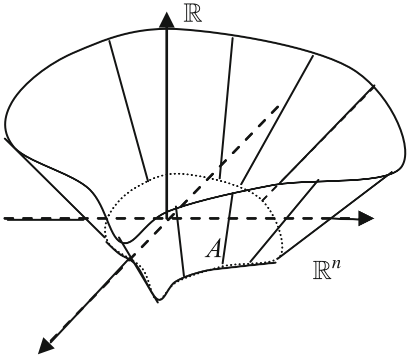


Fig. 2.5 The figure of D_α

Proof Let $v^0 \in \mathbb{R}$ and $v^0 \geq \alpha d\left(v, T_+\left(\hat{x}, A\right)\right)$. Because $T_+\left(\hat{x}, A\right)$ is closed, there exists a $\omega \in T_+\left(\hat{x}, A\right)$ such that $d\left(v, T_+\left(\hat{x}, A\right)\right) = \|v - \omega\|$. Then $v^0 \geq \alpha \|v - \omega\|$. Moreover,

$$\alpha d\left(\hat{x} + \lambda v, A\right) \leq \alpha d\left(\hat{x} + \lambda \omega, A\right) + \alpha \lambda \|v - \omega\| \leq \alpha d\left(\hat{x} + \lambda \omega, A\right) + \lambda v^0.$$

Dividing both sides by λ, since $\omega \in T_+\left(\hat{x}, A\right)$, we have $\lim\limits_{\lambda \downarrow 0} \lambda^{-1} \alpha d\left(\hat{x} + \lambda v, A\right) \leq v^0$. It means $\lambda^{-1} \alpha d\left(\hat{x} + \lambda v, A\right) \leq v^0 + \varepsilon$ for small λ, i.e., $\left(\hat{x} + \lambda v, \lambda v^0 + \lambda \varepsilon\right) \in D_\alpha$. Equivalently,

$$\lim_{\lambda \downarrow 0} \lambda^{-1} d\left(\left(\begin{bmatrix} \hat{x} \\ 0 \end{bmatrix} + \begin{bmatrix} \lambda v \\ \lambda v^0 \end{bmatrix}\right), D_\alpha\right) \leq \varepsilon.$$

$\varepsilon > 0$ can be selected arbitrarily. Hence, $\left(v, v^0\right) \in T_+\left(\left(\hat{x}, 0\right), D_\alpha\right)$. □

Lemma 2.1.2 Let $F : X \rightarrow Y$ be a Lipschitzian mapping with Lipschitzian constant l. Then

$$d\left((x, y), \text{gra } F\right) \leq d\left(y, F(x)\right) \leq (1 + l)\, d\left((x, y), \text{gra } F\right).$$

Proof We have mentioned that in the Cartesian product space $X \times Y$, the norm can be defined as $(x, y) \in X \times Y,\ \ \|(x, y)\| = \|x\|_X + \|y\|_Y$. By the definition, there exists a $z \in F(x)$ such that $d\left(y, F(x)\right) \geq \|y - z\| - \varepsilon$ for every $\varepsilon > 0$. Hence, we have

$$d\left((x, y), \text{gra } F\right) \leq \|(x, y) - (x, z)\|_{X \times Y} = \|y - z\|_Y \leq d\left(y, F(x)\right) + \varepsilon.$$

Because $\varepsilon > 0$ can be selected arbitrarily, it leads to

$$d\left((x, y), \text{gra } F\right) \leq d\left(y, F(x)\right). \tag{2.1.9}$$

The left inequality is obtained.

Similarly, for $\varepsilon > 0$, there is $(x_0, y_0) \in \text{gra } F$ such that

$$d\left((x, y), \text{gra } F\right) \geq \|(x, y) - (x_0, y_0)\|_{X \times Y} - \varepsilon = \|x - x_0\|_X + \|y - y_0\|_Y - \varepsilon.$$

Because F is a Lipschitzian mapping, for this x, there is a $z \in F(x)$ such that $\|z - y_0\|_Y \leq l\|x - x_0\|_X$. Then,

$$\begin{aligned} d\left(y, F(x)\right) &\leq \|y - z\|_Y \\ &\leq \|y - y_0\|_Y + \|y_0 - z\|_Y \\ &\leq \|y - z\|_Y + l\|x - x_0\|_X \\ &\leq (1 + l)\left(\|y - z\|_Y + \|x - x_0\|_X\right) \\ &\leq (1 + l)\left(d\left((x, y)\, \text{gra } F\right) + \varepsilon\right)\ . \end{aligned}$$

The right side inequality can be obtained since $\varepsilon > 0$ can be selected arbitrarily. □

To end this section, we define the derivative for set-valued mappings. We start with a recall of single-valued function. Let $y = f(x)$ where $x, y \in \mathbb{R}$, be a function. In the plane of $\mathbb{R}^2$, the graph of $y = f(x)$ is a set of $\{(x, f(x))\,;\ x \in \mathbb{R}\}$, i.e., gra $f = \{(x, f(x))\}$. Let $f'(x_0)$ be the derivative of $f(x)$ at x_0. Then, the tangent vector at $(x_0, f(x_0))$ is $(x, f'(x_0)x)$. It is clear that the set $\{(x, f'(x_0)\, x)\,; x \in \mathbb{R}\ (\geq 0)\}$ forms a cone. The idea is extended to the set-valued mapping.

Definition 2.1.7 Let $F : X \rightarrow Y$ be a set-valued mapping. Suppose $(x_0, y_0) \in \text{gra } F$. Then a set-valued mapping $D_+F\,(x_0, y_0) : X \rightarrow Y$ is defined by

$$\text{gra } D_+F\,(x_0, y_0) = T_+\left((x_0, y_0),\ \text{gra } F\right).$$

$D_+F(x_0, y_0)$ is called by the derivative of F at (x_0, y_0). Similarly, the set-valued mapping $D_-F(x_0, y_0) : X \to Y$ defined by

$$\text{gra}\, D_-F(x_0, y_0) = T_-((x_0, y_0),\ \text{gra}\, F)$$

is called by contingent derivative of F at (x_0, y_0). □

The statement of $(x, y) \in$ gra $D_+F(x_0, y_0)$ has an alternative explanation that $y \in D_+F(x_0, y_0)(x)$. By the notion, $D_+F(x_0, y_0)$ is qualified as the derivative of set-valued mapping F at (x_0, y_0). We have to denote it by two variables x_0 and y_0 since for every x_0 there exist more than one $y \in F(x_0)$. Correspondingly, $D_+F(x_0, y_0)(x)$ is the differential. Similarly, $D_-F(x_0, y_0)(x)$ is the contingent differential of F at (x_0, y_0). The following theorem gives an explanation of differential.

Theorem 2.1.14 Suppose $F : X \to Y$ is a Lipschitzian set-valued mapping, then

$$D_+F(x_0, y_0)(x) = \left\{y;\ \lim_{\lambda \downarrow 0} \lambda^{-1} d(y_0 + \lambda y, F(x_0 + \lambda x)) = 0\right\};$$

$$D_-F(x_0, y_0)(x) = \left\{y;\ \lim_{\lambda \downarrow 0} \text{inf} \lambda^{-1} d(y_0 + \lambda y, F(x_0 + \lambda x)) = 0\right\}.$$

Proof Let l be the Lipschitzian constant of F. Then By Lemma 2.1.2, we have

$$\begin{aligned} d((x_0 + \lambda x, y_0 + \lambda y), \text{gra}\, F) &\le d(y_0 + \lambda y, F(x_0 + \lambda x)) \\ &\le (1 + l)\, d((x_0 + \lambda x, y_0 + \lambda y), \text{gra}\, F)\ . \end{aligned} \tag{2.1.10}$$

Multiplying both side by λ^{-1},and let $\lambda \downarrow 0$. If $(x, y) \in T_+((x_0, y_0),\ \text{gra}\, F)$, then

$$d((x_0 + \lambda x, y_0 + \lambda y), \text{gra}\, F) = 0.$$

Hence $\lim\limits_{\lambda \downarrow 0} \lambda^{-1} d\Big(y_0 + \lambda y, F(x_0 + \lambda x)\Big) = 0$.

Conversely, if $\lim\limits_{\lambda \downarrow 0} \lambda^{-1} d\Big(y_0 + \lambda y, F(x_0 + \lambda x)\Big) = 0$, then by using the first inequality of (2.1.10), we have $\lim\limits_{\lambda \downarrow 0} d((x_0 + \lambda x, y_0 + \lambda y), \text{gra}\, F) = 0$, i.e., $(x, y) \in T_+((x_0, y_0),\ \text{gra}\, F)$.

The proof for the contingent derivative is similar to the case of tangent derivative. Hence it is omitted. □

Theorem 2.1.14 shows that the computation of tangent derivative for the set-valued mapping is quite similar to that of single-valued function.

Problems

1. $F : X \to Y$ is a set-valued mapping, $G \subset Y$ is a set. Prove the following equations.

$$F^{-1}\left(G^{c}\right)(\mathrm{w}) = \left[F^{-1}(G)\,(\mathrm{s})\right]^{c},$$

$$F^{-1}\left(G^{c}\right)(\mathrm{s}) = \left[F^{-1}(G)\,(\mathrm{w})\right]^{c}.$$

2. Prove Theorem 2.1.2.
3. Let $F : X \to Y$ be set-valued mapping. Prove the following statements are equivalent:

   (1) F is lower semi-continuous at x_0.
   (2) For every $y_0 \in F\left(x_0\right)$ and $\varepsilon > 0$, there is a $\delta > 0$ such that for every $x \in B\left(x_0, \delta\right)$ there exists a $y \in F(x) \cap B\left(y_0, \varepsilon\right)$.
   (3) For every $y_0 \in F\left(x_0\right)$, there exists a sequence $\{x_n; x_n \in X\}$ such that $x_n \to x_0$, moreover, there is a sequence $\{y_n; y_n \in F\left(x_n\right)\}$ such that $y_n \to y_0$.

4. A set-valued mapping $F(x)$ is lower semi-continuous. Is it $\varepsilon-$ lower semi-continuous? If you answer "no", then give a counter example.
5. Prove the following statements:

   (1) If the set-valued mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is upper semi-continuous, and $g : \mathbb{R}^p \to \mathbb{R}^m$ is continuous single-valued function, then $F \circ g : \mathbb{R}^p \to \mathbb{R}^n$ is upper semi-continuous.
   (2) If the set-valued mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is lower semi-continuous, and $g : \mathbb{R}^p \to \mathbb{R}^m$ is continuous single-valued function, then $F \circ g : \mathbb{R}^p \to \mathbb{R}^n$ is lower semi-continuous.

6. Let $F : X \to Y$ be set-valued mapping. Prove the following statements are equivalent:

   (1) F is $\varepsilon-$ upper semi-continuous.
   (2) $\lim\limits_{x \to x_0} \sup F(x) = F\left(x_0\right)$ (By the definition of upper limitation, it is equivalent to $\{u_0; \exists x_n \to x_0, u_n \in F\left(x_n\right), u_n \to u_0\} \subset F\left(x_0\right)$);
   (3) If $u \notin F\left(x_0\right)$, there exist open sets W and V such that $u \in W \subset Y$, $x_0 \in V \subset X$ and $V \cap F^{-1}(W) = \varnothing$.

7. Give examples to show that

   (1) $F : X \to Y$ is an open set-valued mapping, but there is an $x_0 \in X$ such that $F(x_0)$ is not an open set.
   (2) $F : X \to Y$ is a set-valued mapping with closed value, but F is not a closed mapping.

8. Let $F : X \to Y$ be set-valued mapping. Suppose gra F is a closed set, and for every $x_0 \in X$ and $\delta > 0$, the set
$$M_\delta = \text{cl } \{F(x);\ x \in B\,(x_0, \delta)\}$$
is compact. Then F is upper semi-continuous.
9. Let $K_1, K_2 \subset \mathbb{R}^n$ be two closed and convex cones. If re int $K_1 \cap$ re int $K_2 \neq \varnothing$, then cl $(K_1 \cap K_2) =$ cl$K_1 \cap$ clK_2. Moreover, $(K_1 \cap K_2)^* = (\text{cl}K_1 \cap \text{cl}K_2)^* = \text{cl}\left(K_1^* + K_2^*\right)$.
10. Let $K_1, K_2 \subset \mathbb{R}^n$ be two closed and convex cones. Prove the following conclusions and further to discuss which statement is still valid if the condition of "closed" or "convex" is removed.

    (1) $(K_1 + K_2)^* = K_1^* \cap K_2^*$;
    (2) $K_1^{**} = K_1$;
    (3) If $\Lambda : \mathbb{R}^m \to \mathbb{R}^n$ is a linear mapping, and $\Lambda\mathbb{R}^m - K_1 = \mathbb{R}^n$, then $\left(\Lambda^{-1}K_1\right)^* = \Lambda^* K_1^*$ (If Λ is treated as a matrix, then Λ^* is its conjugate matrix);
    (4) $\partial\delta\,(x, K_1) = -K_1^*$.

11. Let $x_i \in \mathbb{R}^n$, $i = 1, 2, \ldots, m$ and denote $K = \{y; \langle y, x_i\rangle \geq 0, i = 1, 2, \ldots m\}$. Then K is a closed and convex cone. Moreover, $= \overline{\text{conx}}\,(x_i, i = 1, 2, \ldots, m)$.
12. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a set-valued mapping with convex and compact value (i.e., for every x, $F(x)$ is a convex and compact set). Then F is continuous if and only if the supporting mapping $x \to S\,(y^*, F(x))$ is continuous for every y^*.
13. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a set-valued mapping. If gra F is closed and the set
$$M_\delta = \text{cl}\{F(x);\ ||x - x_0|| < \delta\}$$
is compact for some $\delta > 0$, then F is upper semi-continuous at x_0.
14. Let $F : \mathbb{R}^n \to A$ be a set-valued mapping, and $A \subset \mathbb{R}^m$ be a compact set. If gra F is closed, then F is with closed value and is upper semi-continuous.

2.2 Selection of Set-Valued Mappings

In the first section of this chapter, we have introduced the set-valued mapping and pointed out that the image of a set-valued mapping $F(x)$ is a set for every x in the domain. Consequently, it is possible to select a $y \in F(x)$ for every $x \in$ dom F, then a single-valued mapping $y = f(x)$ yields. It looks a very simple task. The task is called selection of a set-valued mapping. Alternatively, a selection is to give a way to construct a single-valued mapping $y = f(x)$ such that $f(x) \in F(x)$ for every $x \in A$. If there is an $x \in A$, $F(x)$ holds more than one elements, then the selection is not unique. The requirement is to choose a mapping $f(x)$ with some satisfied properties, for example, it is a continuous mapping, or a Lipschitzian mapping, etc. Readers will find that obtaining a nice selection is a hard work. It is quite extravagant to obtain a Lipschitzian selection.

Usually, there are three kinds selection: continuous selection, measurable selection and approximate selection. This book only considers continuous selection and approximate selection. The later is to look for a sequence $\{f_n(x)\}$ such that the limitation of $f_n(x)$ belongs to $\mathrm{cl}\, F(x)$. After the theory of selection, we spend some time to deal with the problem of fixed points which is a very important title in topology.

2.2.1 The Minimal Selection

Definition 2.2.1 $A \subset X$ is a set. If there is a vector $x_0 \in A$ such that $\|x_0\| = \min\{\|x\|, \ \ x \in A\}$, then the x_0 is called by a minimal norm element of A. The set of minimal norm elements is denoted by $m(A)$. □

In general, we cannot guarantee the existence of minimal norm element for a set. But if the space is complete and the set is closed, then we can assure $m(A) \neq \varnothing$. Moreover, when the set A is convex and closed, then $m(A)$ holds only one element. If $m(A)$ has only one element, we call it the minimal norm element and also denote it by $m(A)$ for simplicity.

Suppose A is a convex and closed set. Because $m(A) = \underset{x\in A}{\arg\min}\, \|x\|$[1] We have $m(A) = \pi\,(0, A)$. By Remark 2 of Lemma 1.2.2, $\langle m(A) - x_A, m(A)\rangle \leq 0$ for every $x_A \in A$ and if there is an element x such that $\langle x - x_A, x\rangle \leq 0$ for every $x_A \in A$, then $x = m(A)$.

Theorem 2.2.1 Suppose that the set-valued mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is continuous and with convex and closed value. Then $f(x) = m\,(F(x))$ is a continuous selection.

Proof By the condition of theorem, $F(x)$ is a convex and closed set for every $x \in \mathbb{R}^n$. Hence, $m(F(x))$ has only one element. We now prove that the single-valued mapping $x \mapsto m\,(F(x))$ is continuous.

Denote $y_0 = m\,(F\,(x_0))$. We first consider the case that $y_0 = 0 \in \mathbb{R}^m$. By the condition that $F(x)$ is continuous and is lower semi-continuous at x_0. For every $\varepsilon > 0$, there is $\delta > 0$ such that if $x \in B\,(x_0, \delta)$, then $F(x) \cap B\,(0, \varepsilon) \neq \varnothing$. It means there is a $y(x) \in F(x) \cap B\,(0, \varepsilon)$. From the definition of the minimal norm element, we have $\|m\,(F(x))\| \leq \|y(x)\| \leq \varepsilon$, i.e., $f(x)$ is continuous at x_0.

We now turn to the case where $y_0 \neq 0$. $F(x)$ is upper semi-continuous at x_0. Hence, for every $\varepsilon > 0$, there is $\delta > 0$ such that if $x \in B\,(x_0, \delta)$, then

[1] The notation $\underset{x\in A}{\arg\min} f(x)$ expresses a set. If $x_0 \in \underset{x\in A}{\arg\min} f(x)$ then $f\,(x_0) = \underset{x\in A}{\min} f(x)$. Similarly, we can define $\underset{x\in A}{\arg\max} f(x)$. The set $\underset{x\in A}{\arg}\,\{f(x) = 0\}$ is the set of roots of $f(x) = 0$ and $x \in A$.

$$F(x) \subset F(x_0) + \frac{\varepsilon}{\|y_0\|} B_m. \tag{2.2.1}$$

There exists a $y(x_0) \in F(x_0)$ such that $m(F(x)) \in \Big(y(x_0) + \frac{\varepsilon}{\|y_0\|} B_m\Big)$, or equivalently, $\|m(F(x)) - y(x_0)\| < \frac{\varepsilon}{\|y_0\|}$. It follows $|\langle y_0, m(F(x)) - y(x_0)\rangle| \leq \|y_0\| \, \|m(F(x)) - y(x_0)\| < \varepsilon$.

By the remark given behind Definition 2.2.1, $\langle y_0 - y(x_0), y_0\rangle \leq 0$, i.e., $0 < \langle y_0, y_0\rangle \leq \langle y_0, y(x_0)\rangle$. It leads to

$$\begin{aligned}
\langle y_0, m(F(x))\rangle &= \Big\langle y_0, y(x_0)\Big)\Big\rangle + \langle y_0, m(F(x)) - y(x_0)\rangle \\
&\geq \Big\langle y_0, y(x_0)\Big)\Big\rangle - \varepsilon \\
&\geq \langle y_0, y_0\rangle - \varepsilon,
\end{aligned}$$

or,

$$\langle y_0, m(F(x)) - y_0\rangle \geq -\varepsilon.$$

From the Problem 1 of this section, the relation (2.2.1) implies $\|m(F(x))\| \leq \|y_0\| + \frac{\varepsilon}{\|y_0\|}$. Thus,

$$\begin{aligned}
\langle m(F(x)) - y_0, m(F(x)) - y_0\rangle &= \langle m(F(x)), m(F(x)) - y_0\rangle - \langle y_0, m(F(x)) - y_0\rangle \\
&= \Big\langle m\Big(F(x), m(F(x))\Big\rangle - \langle m(F(x)), y_0\rangle \\
&\quad - \Big\langle y_0, m(F(x)) - y_0\Big\rangle \\
&= \|m(F(x))\|^2 - \Big\langle m\Big(F(x) - y_0 + y_0, y_0\Big\rangle \\
&\quad - \Big\langle y_0, m(F(x)) - y_0\Big\rangle \\
&= \|m(F(x))\|^2 - \|y_0\|^2 - 2\langle y_0, m(F(x)) - y_0\rangle.
\end{aligned}$$

On the other hand, $\|m(F(x))\|^2 \leq \Big(\|y_0\| + \frac{\varepsilon}{\|y_0\|}\Big)^2 = \|y_0\|^2 + 2\varepsilon + \Big(\frac{\varepsilon}{\|y_0\|}\Big)^2$. The above inequality leads to

$$\|m(F(x)) - y_0\|^2 \leq 4\varepsilon + \left(\frac{\varepsilon}{\|y_0\|}\right)^2.$$

We then conclude that $f(x) = m(F(x))$ is continuous at x_0. □

Corollary 2.2.1 If $F : \mathbb{R}^n \to \mathbb{R}^m$ is continuous and with convex and closed value, then for every $a \in \mathbb{R}^n, f(x) = \pi(a, F(x))$ is a continuous selection.

Proof Let $\overline{x} = x - a$ Then $\pi(a, F(x)) = \pi(0, F(\overline{x} + a)) = \pi\left(0, \overline{F}(\overline{x})\right) = m\left(\overline{F}(\overline{x})\right)$. It is clear that $\overline{F}$ is continuous and with convex and closed value. $\overline{f}(\overline{x}) = m\left(\overline{F}(\overline{x})\right)$ is a continuous selection by Theorem 2.2.1. Hence, $f(x) = \overline{f}(x - a)$ is continuous. □

Remark 1 Theorem 2.2.1 and Corollary 2.2.1 are still valid, if F is a set-valued mapping with convex and compact value form a normed space X to a Hilbert space Y. □

Remark 2 Recall the proof of Theorem 2.2.1, we have applied the lower semi-continuity only for the case that $y_0 = 0 \in F(x_0)$. Hence, if $0 \notin F(x)$ then it is sufficient for Theorem 2.2.1 that F is upper semi-continuous and with convex and closed value. Accordingly, Corollary 2.2.1 can be rewritten as "there is a $a \in \mathbb{R}^m$, $a \notin F(x)$ for all $x \in \mathbb{R}^n$, then $F(x)$ holds a continuous selection provided that $F(x)$ is upper semi-continuous and with convex and closed value. □

To extend the conclusion to a more general result, we give a lemma now.

Lemma 2.2.1 Let A be a convex and closed set of $\mathbb{R}^n$. Then the mapping of projection $x \mapsto \pi(x, A)$ is a Lipschitzian mapping and its Lipschitzian constant can be 1.

Proof Suppose $x_1, x_2 \in \mathbb{R}^n$. By Remark 2 of Lemma 1.2.2, we have $\langle \pi(x, A) - x_A, \pi(x, A) - x\rangle \leq 0$. Replacing x by x_1, x_2 and replacing x_A by $\pi(x_2, A)$ and $\pi(x_1, A)$, respectively, we obtain

$\langle \pi(x_1, A) - \pi(x_2, A), \pi(x_1, A) - x_1\rangle \leq 0$ and $\langle \pi(x_2, A) - \pi(x_1, A), \pi(x_2, A) - x_2\rangle \leq 0$.

The second inequality can be rewritten as $\langle \pi(x_1, A) - \pi(x_2, A), x_2 - \pi(x_2, A)\rangle \leq 0$ and is added to the first inequality, we obtain

$$\langle \pi(x_1, A) - \pi(x_2, A), \pi(x_1, A) - x_1 + x_2 - \pi(x_2, A)\rangle \leq 0.$$

It leads to

$$\begin{aligned} \|\pi(x_1, A) - \pi(x_2, A)\|^2 &\leq \langle x_1 - x_2, \pi(x_1, A) - \pi(x_2, A)\rangle \\ &\leq \|x_1 - x_2\| \, \|\pi(x_1, A) - \pi(x_2, A)\| . \end{aligned}$$

When $\pi(x_1, A) \neq \pi(x_2, A)$, it leads to

$$\|\pi(x_1, A) - \pi(x_2, A)\| \leq \|x_1 - x_2\| . \tag{2.2.2}$$

Inequality (2.2.2) is also valid if $\pi(x_1, A) = \pi(x_2, A)$. The lemma is verified. □

Theorem 2.2.2 Suppose that the set-valued mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is continuous and with convex and closed value, and $g : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous single-valued function, then $f(x) = \pi(g(x), F(x))$ is a continuous selection of F.

Proof $F(x)$ is convex and closed, it follows $f(x) = \pi\,(g(x), F(x))$ is a single valued mapping. Consider now the continuity of $f(x)$.

$$\begin{aligned} f\left(x_1\right) - f\left(x_2\right) &= \pi\left(g\left(x_1\right), F\left(x_1\right)\right) - \pi\left(g\left(x_2\right), F\left(x_2\right)\right) \\ &= \pi\left(g\left(x_1\right), F\left(x_1\right)\right) - \pi\left(g\left(x_1\right), F\left(x_2\right)\right) + \pi\left(g\left(x_1\right), F\left(x_2\right)\right) \\ &\quad - \pi\left(g\left(x_2\right), F\left(x_2\right)\right) \ . \end{aligned}$$

If we take $a = g\left(x_1\right)$, by Corollary 2.2.1

$$\pi\left(g\left(x_1\right), F\left(x_1\right)\right) - \pi\left(g\left(x_1\right), F\left(x_2\right)\right) \to 0, \quad \left(x_1 - x_2 \to 0\right);$$

On the other hand, if we treat $F(x_2)$ as set A, by Lemma 2.2.1, we have

$$\pi\left(g\left(x_1\right), F\left(x_2\right)\right) - \pi\left(g\left(x_2\right), F\left(x_2\right)\right) \to 0, \quad \left(x_1 - x_2 \to 0\right).$$

Consequently, $f\left(x_1\right) - f\left(x_2\right) \to 0$ as $x_1 - x_2 \to 0, f(x)$ is continuous. □

The selection obtained by Theorem 2.2.1 is called by the minimal selection since the selected element is the vector with minimal norm. Because Theorem 2.2.2 deduces from Theorem 2.2.1, it is called as patulous minimal selection theorem or minimal selection theorem for simplicity. The conclusion is still valid if the $\mathbb{R}^n$ and $\mathbb{R}^m$ are replaced by Hilbert Spaces X and Y. There is an example to show the minimal selection is continuous but may fail to be Lipschitzian mapping. The readers who are interested in the example are referred to Aubin and Cellina (1984).

2.2.2 Michael Selection Theorem

The most famous result in selection theory of set-valued mapping is the Michael selection theorem. It is notable that in the proof of Theorem 2.2.2, we have applied the operation of inner production. Hence, if we try to apply Theorem 2.2.2, the image space should be Hilbert space. The Michael selection theorem improves the condition of Theorem 2.2.1, the image space is only a Banach space. In this subsection, we do not restrict ourselves in $\mathbb{R}^n$. Let us start with a conclusion of unit decomposition.

1. Unit decomposition

Decomposing the integer 1 into a summation of several nonnegative functions defined on a normed space is known as the unit decomposition, or the partition of 1. In the theory of functional analysis, there are many kinds unit decomposition. This book only introduces the Lipschitzian unit decomposition.

Let X be a normed space, $A \subset X$ be a set. $\{\Omega_i; i \in I\}$ is an open covering of A.[2] The open covering $\{\Omega_i; i \in I\}$ is said to be a locally finite open covering if for every $\Omega_{i_0} \in \{\Omega_i; i \in I\}$, there are only finite open sets $\Omega_j \in \{\Omega_i; i \in I\}, j = 1, 2, \ldots, J$ such that $\Omega_{i_0} \cap \Omega_j \neq \varnothing$.

Let $\{\Omega_i; i \in I\}$ be an open covering of $A \subset X$ and $\{p_i(x);\ i \in I\}$, where $p_i(x) : A \to \mathbb{R}\,(\geq 0)$, be a set of functions. The set $\{p_i(x); i \in I\}$ is a unit decomposition attached to $\{\Omega_i; i \in I\}$, if it satisfies the following conditions:

(1) $p_i(x) \geq 0,\ x \in \Omega_i;\ p_i(x) = 0,\ x \notin \Omega_i;$
(2) For every $x \in A$, $\sum\limits_{i \in I} p_i(x) = 1$.

If $p_i(x)$ is continuous for every $i \in I$, then $\{p_i(x); i \in I\}$ is said to be a continuous decomposition. If $p_i(x)$ is a Lipschitzian function for every $i \in I$, then $\{p_i(x); i \in I\}$ is said to be a Lipschitzian decomposition. If the open covering $\{\Omega_i; i \in I\}$ is determined, we often omit the words "attached to $\{\Omega_i; i \in I\}$" for simplicity.

Lemma 2.2.2 Let $A \subset \mathbb{R}^n$, and $\{\Omega_i; i \in I\}$ be an open covering which is nontrivial and locally finite. Then there exists a local Lipschitzian unit composition.

Proof For every $i \in I$, define a function $q_i(x) : A \to \mathbb{R}\,(\geq 0)$, $q_i(x) = d\,(x, A \backslash \Omega_i)$. By Problem 2 in this section, $q_i(x)$ is a Lipschitzian function and its Lipschitzian constant is 1. Moreover, $q_i(x) = 0, x \notin \Omega_i$ and $q_i(x) \geq 0, x \in \Omega_i$ since Ω_i is an open set. The $p_i(x)$ is defined as follows:

$$p_i(x) = \frac{q_i(x)}{\sum\limits_{j \in I} q_j(x)}.$$

By the definition of $q_i(x)$, it is obvious that the set $\{p_i(x); i \in I\}$ satisfies the properties of unit decomposition. The remaining problem is that it is a locally Lipschitzian function. We now prove this fact.

Let $x_0 \in A$. Then there is an $i_0 \in I$, such that $x_0 \in \Omega_{i_0}$. Ω_{i_0} is an open set, there is a neighborhood $B(x_0, \varepsilon)$ such that $B\,(x_0, \varepsilon) \subset \Omega_{i_0}$. The open covering is locally finite; hence, there are only finite $\Omega_{j_k} \in \{\Omega_i, i \in I\}$, $k = 1, 2, \ldots, N$ such that $B\,(x_0, \varepsilon) \cap \Omega_{j_k} \neq \varnothing$. It follows that when $x \in B\,(x_0, \varepsilon)$, there are at most N functions $q_{j_k}(x)$ are nonzero in $\{q_i(x); i \in I\}$. On the other hand, there exist $0 < m < M$, such that $m \leq \sum\limits_{j \in I} q_j(x) \leq M$ on the compact set cl$B(x_0, \varepsilon)$. Thus, when $x_1, x_2 \in B\,(x_0, \varepsilon)$, we have

[2] If there is an Ω_{i_0}, $i_0 \in I$, such that cl $\Omega_{i_0} \supset A$, then the covering is trivial. We do not deal with such a special case.

$$
\begin{aligned}
|p_i(x_1) - p_i(x_2)| &= \left| \frac{q_i(x_1)}{\sum\limits_{j\in I} q_j(x_1)} - \frac{q_i(x_2)}{\sum\limits_{j\in I} q_j(x_2)} \right| \\
&= \left| \frac{q_i(x_1)\sum\limits_{j\in I} q_j(x_2) - q_i(x_2)\sum\limits_{j\in I} q_j(x_1)}{\sum\limits_{j\in I} q_j(x_1)\sum\limits_{j\in I} q_j(x_2)} \right| \\
&\leq \tfrac{1}{m^2}\sum_{j\in I} \left| q_i(x_1)\, q_j(x_2) - q_i(x_2)\, q_j(x_1) \right| \\
&\leq \tfrac{1}{m^2}\sum_{j\in I} \Big(\left| q_i(x_1)\, q_j(x_2) - q_i(x_1)\, q_j(x_1) \right| \\
&\quad + \left| q_i(x_1)\, q_j(x_1) - q_i(x_2)\, q_j(x_1) \right| \Big) \\
&\leq \tfrac{1}{m^2}\sum_{j\in I} \Big(|q_i(x_1)| \left| q_j(x_2) - q_j(x_1) \right| + \left| q_j(x_1) \right| |q_i(x_1) - q_i(x_2)| \Big) \\
&\leq \tfrac{1}{m^2}(\sum_{j\in I} \left| q_j(x_2) - q_j(x_1) \right|) + (|q_i(x_1) - q_i(x_2)| \sum_{j\in I} \left| q_j(x_1) \right|) \\
&\leq \tfrac{1}{m^2}(N + MN)\,|x_2 - x_1| \ .
\end{aligned}
$$

The conclusion is verified. □

2. Michael selection Theorem

This subsection proves the main result of this section: Michael selection theory. We introduce the following two facts.

Lemma 2.2.3 Suppose that X and Y are two normed spaces, $\varphi : X \to Y$ is a continuous single-valued mapping, and $\varepsilon : X \to \mathbb{R}\,(> 0)$ is a continuous functional. If $G : X \to Y$ is a lower semi-continuous set-valued mapping, then the mapping defined by

$$F(x) = B\left(\varphi(x), \varepsilon(x)\right) \bigcap G(x)$$

is lower semi-continuous. □

The proof of this lemma is a classical exercise of application of $\varepsilon - \delta$ language. Hence we leave it to readers.

Another fact is about paracompactness. A set is said to be paracompact if for every open covering of the set there exists a fined open covering which is locally finite. Using mathematical language, it can be defined as follows. Let A be set, and $\{U_\alpha\}$ is an open covering of A. Then there exists another open covering $\{V_\beta\}$. $\{V_\beta\}$ is locally finite, and for every $V_\beta \in \left\{V_\beta\right\}$, there is a $U_\alpha \in \{U_\alpha\}$ such that $V_\beta \subset U_\alpha$. The $\{V_\beta\}$ is called a fineness of $\{U_\alpha\}$.

In 1948, A.H. Stone proved that a metric space is paracompact. The proof of this conclusion is quite complicated, we do not try to provide it in this book. The readers are referred to textbook for topology for the proof.

Theorem 2.2.3 (Michael selection theorem) Let X and Y be two normed spaces, and Y be complete. If F is a lower semi-continuous and with convex and closed value, then F exists a continuous selection.

Proof The proof of Theorem 2.2.3 consists of three steps. The critical step is construct a continuously single-valued mapping ϕ_ε: $X \rightarrow Y$ which satisfies $d\left(\varphi_\varepsilon(x), F(x)\right) < \varepsilon$. We now deal with the first step.

(1) For a given $\varepsilon > 0$, there is a continuously single-valued mapping ϕ_ε: $X \rightarrow Y$ such that $d\left(\varphi_\varepsilon(x), F(x)\right) < \varepsilon$.

$F : X \rightarrow Y$ is lower semi-continuous, hence, for a given $\varepsilon > 0$ and every $x \in X$, there is a $y_x \in F(x)$, such that $F\left(\overline{x}\right) \cap B\left(y_x, \varepsilon\right) \neq \varnothing$ provided that $\overline{x} \in B\left(x, \delta\right)$ for some $\delta > 0$. We now fix the vector y_x and the neighborhood $B(x, \delta)$ for every $x \in X$.

Thus the set $\{B\left(x, \delta\right), \ x \in X\}$ is an open covering of X. X is a normed space; consequently, by Stone theorem, $\{B\left(x, \delta\right), \ x \in X\}$ holds a fined open covering $\{U_i\}$ which is locally finite, i.e., for every U_i, there is a $B(x_i, \delta_i)$ such that $U_i \subset B\left(x_i, \delta_i\right)$. Lemma 2.2.2 shows that the $\{U_i\}$ provides a Lipschitzian unit partition $\{p_i(x)\}$.

For every $x \in X$, there are only finite $U_1, \ldots, U_{N_x} \in \{U_i\}$, such that $x \in U_j, \ j = 1, 2, \ldots, N_x$. Then, there is a $B(x_j, \delta_i)$ such that $U_j \subset B\left(x_j, \delta_i\right)$. From this x_j there is a corresponding y_{x_j} such that $y_{x_j} \in B\left(x_j, \delta_j\right)$ and $F\left(\overline{x}\right) \cap B\left(y_{x_j}, \varepsilon\right) \neq \varnothing$ provided that $\overline{x} \in B\left(x_j, \delta_j\right)$. We now define the following mapping.

$$\varphi_\varepsilon(x) = \sum_{j \in I} p_j(x) y_{x_j},$$

where I is the index set of i. $\phi_\varepsilon(\xi)$ is then a local Lipschitzian mapping, so it is continuous. Because $x \in U_j \subset B\left(x_j, \delta_j\right)$, $F(x) \cap B\left(y_{x_j}, \varepsilon\right) \neq \varnothing$, i.e., $y_{x_j} \in F(x) + \varepsilon B_m$. $F(x)$ is a convex set, hence $\varphi_\varepsilon(x) = \sum\limits_{j \in I} p_j(x) y_{x_j} \in F(x) + \varepsilon B_m$, i.e., $d\left(\varphi_\varepsilon(x), F(x)\right) < \varepsilon$ for every $x \in X$.

It is worth to note that at the above verification, we did not apply the condition that $F(x)$ is with closed value.

(2) A Cauchy sequence $\{f_n(x)\}$ is constructed where every $f_n(x)$ is continuous at X.

Let $\varepsilon = \frac{1}{2}$. By the construction method proposed in Step (1) we can obtain a Lipschitzian functional $\varphi_{\frac{1}{2}}(x)$ such that $d\left(\varphi_{\frac{1}{2}}(x), F(x)\right) \leq \frac{1}{2}$ for every $x \in X$. Now we denote $f_1(x) = \varphi_{\frac{1}{2}}(x)$. Let us consider the set-valued mapping $F_1(x) = B\left(f_1(x), \left(\frac{1}{2}\right)^2\right) \cap F(x)$, $F_1(x)$ is lower semi-continuous by Lemma 2.2.2. Taking $\varepsilon = \frac{1}{4}$, by the method proposed in Step (1) again, a Lipschitzian functional $\varphi_{\frac{1}{4}}(x)$ can be constructed such that $d\left(\varphi_{\frac{1}{4}}(x), F_1(x)\right) < \frac{1}{4}$. Let $f_2(x) = \varphi_{\frac{1}{4}}(x)$. By definition,

$F_1(x) \subset F(x) \cap B\left(f_1(x), \frac{1}{4}\right)$, therefore $d\left(f_2(x), F(x)\right) \leq d\left(f_2(x), F_1(x)\right) < \frac{1}{4}$ and

$$\|f_1(x) - f_2(x)\| \leq d\left(f_1(x), F_1(x)\right) + d\left(f_2(x), F_1(x)\right) \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Inductively, when we construct $f_n(x)$ which satisfies that $d\left(f_n(x), F(x)\right) < \left(\frac{1}{2}\right)^n$ and $\|f_n(x) - f_{n-1}(x)\| \leq \left(\frac{1}{2}\right)^{n-1}$, we let $F_n(x) = B\left(f_n(x), \left(\frac{1}{2}\right)^{n+1}\right) \cap F(x)$, then $F_n(x)$ is lower semi-continuous by Lemma 2.2.2. There is a $f_{n+1}(x) = \varphi_{\frac{1}{2^{n+1}}}(x)$ which is Lipschitzian and satisfies $d\left(f_{n+1}(x), F_n(x)\right) < \left(\frac{1}{2}\right)^{n+1}$. By the definition of $F_n(x)$, $d\left(f_{n+1}(x), F(x)\right) \leq d\left(f_{n+1}(x), F_n(x)\right) < \left(\frac{1}{2}\right)^{n+1}$ and $\|f_{n+1}(x) - f_n(x)\| \leq \left(\frac{1}{2}\right)^n$. Thus,

$$\begin{aligned}\|f_{n+k}(x) - f_n(x)\| &\leq \|f_{n+k}(x) - f_{n+k-1}(x)\| + \cdots + \|f_{n+1}(x) - f_n(x)\| \\ &< \left(\tfrac{1}{2}\right)^{n+k-1} + \cdots + \left(\tfrac{1}{2}\right)^n \\ &< \left(\tfrac{1}{2}\right)^{n-1}.\end{aligned}$$

$\{f_n(x)\}$ is a Cauchy sequence.

(3) The selection $f(x)$ is obtained.

Let $f(x) = \lim_{n\to\infty} f_n(x)$. Then $f(x)$ is continuous. Moreover, the space Y is complete and $F(x)$ is a closed set; consequently, $f(x) \in F(x)$. □

2.2.3 Lipschitzian Approximation

If f is a continuous single-valued mapping, then f may not be a Lipschitzian mapping. But its opposite statement is true. However, a Lipschitzian set-valued mapping may not be continuous. This subsection will prove that an upper semi-continuous set-valued mapping can be approximated by a sequence of Lipschitzian mappings.

Theorem 2.2.4 Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be an upper semi-continuous set-valued mapping. It is bounded and is with closed and convex value. Then there exist set-valued mappings $F_k : \mathbb{R}^n \to \mathbb{R}^m$, $k = 1, 2, \ldots$ which satisfy the following requirements.

(1) $F_k : \mathbb{R}^n \to \mathbb{R}^m$, $k = 1, 2, \ldots$ are all Lipschitzian mappings and are bounded and with closed and convex values.

(2) $F(x) \subset \cdots \subset F_{k+1}(x) \subset F_k(x) \subset \cdots \subset F_1(x) \subset F_0(x)$.

(3) For every $\varepsilon > 0$, there is a $K = K(x, \varepsilon)$ such $F_k(x) \subset F(x) + \varepsilon B_m$ provided that $k > K$.

Proof Suppose $F(x) \subset bB_m$ for every $x \in X$ where $b \in \mathbb{R}\,(> 0)$ and B_m is the open unit ball of $\mathbb{R}^n$. Denote $e_i = [0 \ldots \underset{i}{1} \ldots 0]$, i.e., its ith component is 1 and others are zeros. $\{e_i;\ i = 1, 2, \ldots, n\}$ is a set of normal orthogonal basis of $\mathbb{R}^n$. Let $\mathbb{Z}^n = \underbrace{\mathbb{Z} \times \mathbb{Z} \times \cdots \times \mathbb{Z}}_{n} \subset \mathbb{R}^n$ be the set in which components of every vector are all integers. The following proof consists of three steps.

(1) Constructing a sequence of Lipschitzian set-valued mappings $\{F_k(x)\}$.

Fix $\rho_0 = 1$ and define $\mathbb{Z}_0^n = \rho_0 \mathbb{Z}^n = \mathbb{Z}^n$. In addition, we define a set

$$O_0 = \{x; x = a_1 e_1 + a_2 e_2 + \cdots + a_n e_n,\ a_i \in (-\rho_0, \rho_0)\},$$

O_0 is an open cube whose center is the origin and lengths of edges are all equal to $2\rho_0$. O_0 is obviously convex. $\mathbb{Z}_0^n + O_0 = \left\{z_0 + O_0,\ z_0 \in \mathbb{Z}_0^n\right\}$ is qualified as a locally finite open covering of $\mathbb{R}^n$. Then the open covering yields a Lipschitzian unit decomposition $p_{z_0}(x)$. For every $z_0 \in \mathbb{Z}_0^n$, we can have a set $C_{z_0} = \text{cl co } F(z_0 + 2O_0)$, then set-valued mapping $F_0(x)$ is defined as follows:

$$F_0(x) = \sum_{z_0 \in \mathbb{Z}_0^n} p_{z_0}(x) C_{z_0}.$$

Because $p_{z_0}(x)$ is a local Lipschitzian function, $F_0(x)$ is then a Lipschitzian set-valued mapping and with closed and convex value. $F(x) \subset bB_m$ leads to that $C_{z_0} \subset bB_m$, moreover, $F_0(x) \subset bB_m$ since $p_{z_0}(x)$ is a unit decomposition.

We now construct $F_1(x)$. Define $\rho_1 = \frac{1}{3}\rho_0$ and $\mathbb{Z}_1^n = \rho_1 \mathbb{Z}_0^n$. The element of $\mathbb{Z}_1^n$ can be denoted by $z_1 = \rho_1 z_0$ with $z_0 \in \mathbb{Z}_0^n$. A set O_1 is defined below

$$O_1 = \left\{x; x = a_1 e_1 + a_2 e_2 + \cdots + a_n e_n,\ a_i \in \left(-\frac{1}{3}, \frac{1}{3}\right)\right\}.$$

Then $\mathbb{Z}_1^n + O_1$ forms an open covering of $\mathbb{R}^n$ and it is locally finite. The open covering holds a unit partition $p_{z_1}(x),\ z_1 \in \mathbb{Z}_1^n$. For every $z_1 \in \mathbb{Z}_1^n$, define a set $C_{z_1} = \text{cl co } F(z_1 + 2O_1)$ and a set-valued mapping $F_1(x)$ can be obtained as follows:

$$F_1(x) = \sum_{z_1 \in \mathbb{Z}_1^n} p_{z_1}(x) C_{z_1}.$$

$F_1(x) \subset bB_m$ and is also Lipschitzian and with closed and convex value.

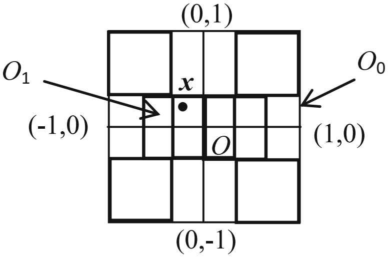


Fig. 2.6 The relation of O_0 and O_1

Generally, in the kth step, we obtain a set-valued mapping $F_k(x)$, which is Lipschitzian and with closed and convex value, and is bounded by bB_m. We then turn to the $(k+1)$th step. By defining $\rho_{k+1} = \frac{1}{3}\rho_k$, the set-valued mapping $F_{k+1}(x)$ can be defined with a similar way, i.e., we define $\mathbb{Z}^n_{k+1} = \rho_{k+1}\mathbb{Z}^n_0$ and O_{k+1}, then we obtain a unit partition $p_{z_{k+1}}(x)$, $z_{k+1} \in \mathbb{Z}^n_{k+1}$ and closed sets $C_{z_{k+1}} = \text{cl co } F(z_{k+1} + 2O_{k+1})$ for each $z_{k+1} \in \mathbb{Z}^n_{k+1}$, and at last we define $F_{k+1}(x) = \sum\limits_{z_{k+1} \in \mathbb{Z}^n_{k+1}} p_{z_{k+1}}(x) C_{z_{K+1}}$.

(2) Proof of $F(x) \subset F_{k+1}(x) \subset F_k(x)$.

If we can prove $F(x) \subset F_1(x) \subset F_0(x)$, then by a similar way, we can prove $F(x) \subset F_{k+1}(x) \subset F_k(x)$.

For a fixed $x \in \mathbb{R}^n$, there are only finite $z_{0_i} \in \mathbb{Z}^n_0$, $i = 1, 2, \ldots, N_{0x}$ such that $x \in z_{0_i} + O_0$. Denote $Z_0(x) = \{z_{0_i}; i = 1, 2, \ldots, N_{0x}\}$ for the x. Similarly, there are only finite $z_{1_j} \in \mathbb{Z}^n_1$, $j = 1, 2, \ldots, N_{1x}$ such that $x \in z_{1_j} + O_1$, and denote $Z_1(x) = \left\{z_{1_j}; j = 1, 2, \ldots, N_{1x}\right\}$. A scheme diagram for the relation of $Z_0(x)$ and $Z_1(x)$ is given in Fig. 2.6.

We now prove that for every $z_{0_i} \in Z_0(x) \subset \mathbb{Z}^n_0$ and every $z_{1_j} \in Z_1(x) \subset \mathbb{Z}^n_1$, $z_{1_j} + 2O_1 \subset z_{0_i} + 2O_0$.

Let $y \in z_{1_j} + 2O_1$. Then $\left\|y - z_{1_j}\right\| \leq 2\rho_1 = \frac{2}{3}\rho_0$. In addition, we have

$$\left\|z_{0_i} - z_{1_j}\right\| \leq \left\|z_{0_i} - x\right\| + \left\|x - z_{1_j}\right\| \leq \rho_0 + \rho_1 = \frac{4}{3}\rho_0.$$

Therefore, $\left\|y - z_{0_i}\right\| \leq \left\|y - z_{1j}\right\| + \left\|z_{1_j} - z_{0_i}\right\| \leq \frac{2}{3}\rho_0 + \frac{4}{3}\rho_0 = 2\rho_0$, i.e., $y \in z_{0_i} + 2O_1$.

The relation $z_{1_j} + 2O_1 \subset z_{0_i} + 2O_0$ implies $C_{z_1} \subset C_{z_0}$. C_{z_1} is convex; hence,

$$C_{z_1} = \sum_{z_0 \in Z_0(x)} p_{z_0}(x) C_{z_1} \subset \sum_{z_0 \in Z_0(x)} p_{z_0}(x) C_{z_0}.$$

It follows that,

$$\begin{aligned}
F_1(x) = \sum_{z_1 \in \mathbb{Z}_1^n} p_{z_1}(x) C_{z_1} &= \sum_{z_1 \in Z_1(x)} p_{z_1}(x) C_{z_1} \\
&\subset \sum_{z_1 \in Z_1(x)} p_{z_1}(x) \sum_{z_0 \in Z_0(x)} p_{z_0}(x) C_{z_0} \\
&= \sum_{z_1 \in Z_1(x)} p_{z_1}(x) \sum_{z_0 \in \mathbb{Z}_0^n} p_{z_0}(x) C_{z_0} \\
&= \sum_{z_1 \in Z_1(x)} p_{z_1}(x) F_0(x) \\
&= F_0(x) \ .
\end{aligned}$$

Simultaneously, when $x \in \mathbb{R}^n$ and $z_{1_j} \in Z_1(x), x \in z_{1j} + O_1$ and $F(x) \subset C_{z_1}$.

$$F(x) \subset \sum_{z_1 \in Z_1(x)} p_{z_1}(x) C_{z_1} = \sum_{z_1 \in Z_1^n} p_{z_1}(x) C_{z_1} = F_1(x).$$

(3) Convergence of $\{F_k(x)\}$.

At last, we verify the convergence of $F_k(x)$. Because $F(x)$ is upper semi-continuous for $x \in \mathbb{R}^n$ and $\varepsilon > 0$, there is a $\delta = \delta(x, \varepsilon)$ such that if $y \in B(x, \delta)$, $F(y) \subset F(x) + \frac{\varepsilon}{2} B_m$. For the x and ε, there is a $K = K(x, \varepsilon)$, $x + O_{K(x,\varepsilon)} \subset B(x, \delta)$. For every $k > K$, denote $Z_k(x) = \left\{ z_{k_j}; j = 1, 2, \ldots, N_{kx} \right\}$, where $x \in z_{k_j} + O_k$. Thus, we have

$$z_{k_j} + 2O_k \subset x + 3O_k \subset x + O_K.$$

It illustrates that $y \in z_{k_j} + 2O_k$ for every $z_{k_j} \in Z_k(x)$, $F(y) \subset F(x) + \frac{\varepsilon}{2} B_m$. Furthermore,

$F\left(z_{k_j} + 2O_k\right) \subset F(x) + \frac{\varepsilon}{2} B_m$. $C_{z_{k_j}} = \text{clco} F_k \left(z_{k_j} + 2O_k\right) \subset F(x) + \frac{\varepsilon}{2} \overline{B}_m \subset F(x) + \varepsilon B_m$. $F(x) + \varepsilon B_m$ is convex, hence, when $k > K$,

$$F_k(x) = \sum_{z_k \in \mathbb{Z}_k^n} p_{z_k}(x) C_{z_k} = \sum_{z_{k_i} \in Z_k(x)} p_{z_k}(x) C_{z_{k_j}} \subset F(x) + \varepsilon B_m.$$

The theorem is now verified. □

From Problem 7 of this section the readers will conclude if the conditions given in Theorem 2.2.4 are satisfied, then for every $\varepsilon > 0$, there is a single-valued mapping $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}^m$ which is continuous and $f_\varepsilon(x) \in F(x) + \varepsilon B$. The fact leads to the following definition.

Definition 2.2.2 Suppose that X and Y are two metric spaces, and $F : X \to Y$ is a set-valued mapping. If for every $\varepsilon > 0$, there is a single-valued mapping $f_\varepsilon : X \to Y$ such that gra $f_\varepsilon \subset$ gra $F + \varepsilon B$, then F is approximately selectable, and f_ε is an $\varepsilon-$ approximate selection. □

From Definition 2.2.2, $f_\varepsilon(x) \in F(x) + \varepsilon B$ is only a sufficient condition for the approximate selection. Let us consider the following example.

Example 2.2.1 Consider a set-valued mapping $F : \mathbb{R} \to \mathbb{R}$

$$F(x) = \begin{cases} -1, & x < 0, \\ [-1, \ \ 1], & x = 0, \\ 1, & x > 0. \end{cases}$$

$F(x)$ is a monotonous mapping which will be studied at the end of this chapter. It is obvious that $F(x)$ has no a continuous selection. But the sigmoid function

$$f_\beta(x) = \frac{e^{\beta x} - e^{-\beta x}}{e^{\beta x} + e^{-\beta x}}$$

is continuous. It is direct to show that for $1 > \varepsilon > 0$, when $\beta > \frac{1}{2\varepsilon} \ln \frac{2-\varepsilon}{\varepsilon}, f_\varepsilon(x)$ is an $\varepsilon-$ approximate selection (Fig. 2.7). □

Example 2.2.1 really gives a new way for the selection since for every $1 > \varepsilon > 0$, except $x = 0, f_\varepsilon(x) \cap F(x) = \varnothing$. Furthermore, it can be seen that no matter how small ε is determined, there still is a $\delta > 0$, when $x \in (0, \delta)$, $d\,(f_\varepsilon(x), F(x)) > \frac{1}{2}$, or, in other words, $f_\varepsilon(x) \notin F(x) + \varepsilon B$.

Theorem 2.2.5 Let X be a metric space, Y be a Banach space, and $F : X \to Y$ be a set-valued mapping, which is upper semi-continuous and with closed and convex value. Then F holds an $\varepsilon-$ approximate selection; moreover, the selection is a Lipschitzian mapping.

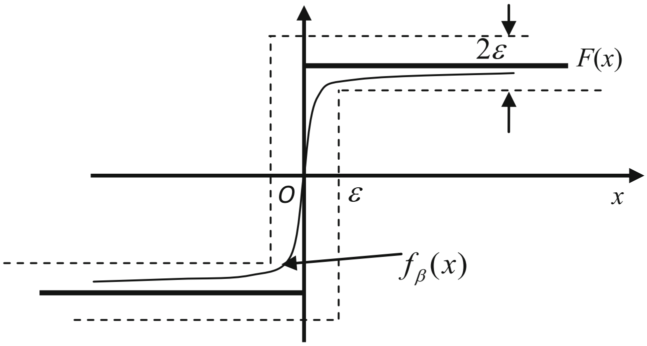


Fig. 2.7 $\varepsilon-$ approximate selection of Example 2.2.1

Proof $F(x)$ is upper semi-continuous, then for a given $\varepsilon > 0$ and every $x_0 \in X$, there is a $\delta_1 = \delta_1(x_0, \varepsilon)$ such that when $x \in B(x_0, \delta_1)$, $F(x) \subset F(x_0) + \frac{\varepsilon}{2}B_Y$. Let $\delta(x_0) = \min\left(\frac{\varepsilon}{2},\ \delta_1\right)$. Then these neighborhoods $B\left(x_0, \frac{\delta(x_0)}{4}\right)$ form an open covering of X if x_0 goes over the space $X.X$ is a metric space; hence, there exists a fined locally finite sub-covering $\{U_\alpha\}$. It follows that there is a Lipschitzian unit decomposition $p_i(x)$ attached to $\{U_\alpha\}$. Because $\{U_\alpha\}$ is a fined covering of $\{B\left(x_0, \frac{\delta(x_0)}{4}\right), x_0 \in X\}$, for every $U_i \in \{U_\alpha\}$, we can find a x_i such that $U_i \subset B\left(x_i, \frac{\delta(x_i)}{4}\right)$. We now fix the x_i for U_i and construct a mapping f_ε as follows

$$f_\varepsilon(x) = \sum_i p_i(x) m\left(F(x_i)\right), \tag{2.2.3}$$

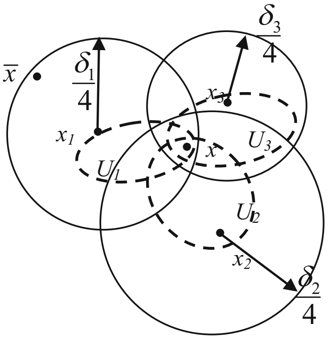
where $m(F(x_i))$ is the minimal norm element in $F(x_i)$. We analyze the properties of the f_ε. At first, because $p_i(x)$ is a Lipschitzian mapping, so is the f_ε. Secondly, because co gra$F$ is convex, $(x_i, m(F(x_i))) \in$ co gra F implies then $(x, f_\varepsilon(x)) \in$ co gra F. At last, we prove that $f_\varepsilon(x) \in F\left(x_j\right) + \frac{\varepsilon}{2}B_Y$ for some $x_j \in X$.

$\{U_i\}$ is a locally finite open covering; consequently, for every $x \in X$, there are only finite nonzero terms in $\sum_i p_i(x) m(F(x_i))$. For a fixed x, let $I(x) = \{i;\ p_i(x) \neq 0\} = \{i;\ x \in U_i\}$, then $I(x)$ is a finite set. By the definition of U_i, an x_i can be found such that $U_i \subset B\left(x_i, \frac{\delta_i}{4}\right)$. Define $\delta_{i_0} = \max\{\delta_i = \delta(x_i), i \in I(x)\}$, the δ_{i_0} corresponds to $x_{i_0} \in X$. Then for every $x_i, i \in I(x)$, $d(x, x_i) < \frac{\delta(x_i)}{4} \leq \frac{\delta\left(x_{i_0}\right)}{4}$, and $d(x_{i_0}, x_i) \leq d(x_{i_0}, x) + d(x, x_i) < \frac{\delta\left(x_{i_0}\right)}{2}$, furthermore, $U_i \subset B(x_{i_0}, \delta_{i_0})$, $i \in I(x)$. Hence, for $i \in I(x)$

$$m\left(F(x_i)\right) \in F(x_i) \subset F(x_{i_0}) + \frac{\varepsilon}{2}B_Y. \tag{2.2.4}$$

Figure 2.8 is used to illustrate the relations of x and U_i, x_i, δ_i. $x \in U_i \subset B\left(x_i, \frac{\delta_i}{4}\right)$, then $d(x_i, x_2) \leq d(x_i, x) + d(x, x_2) < \frac{\delta_2}{2}$ (in Fig. $\delta_2 = \max\{\delta_i, i \in I(x)\}$), i.e., $x_i \in$

Fig. 2.8 The relation of x and U_i



$B\left(x_2, \frac{\delta_2}{2}\right)$, and a point $\overline{x} \in B(x_i, \delta_i)$, $d(\overline{x}, x_2) \leq d(\overline{x}, x_i) + d(x_i, x_2) < \frac{\delta_2}{4} + \frac{\delta_2}{2} < \delta_2$. Thus, $B(x_2, \delta_2) \supset B(x_i, \delta_i) \supset U_i$.

We have assumed that $F(x_{i_0})$ is convex, hence, from Relation (2.2.4)

$$f_\varepsilon(x) = \sum_i p_i(x) m(F(x_i)) \in F(x_{i_0}) + \frac{\varepsilon}{2} B_Y,$$

i.e., there exists a $y_{i_0} \in F(x_{i_0})$ such that $\|f_\varepsilon(x) - y_{i_0}\| < \frac{\varepsilon}{2}$. It follows that

$$d((x, f_\varepsilon(x)), \text{gra}\, F) \leq d\Big((x, f_\varepsilon(x)), (x_{i_0}, y_{i_0}) \leq d(x, x_{i_0}) + d(f_\varepsilon(x), y_{i_0}) < \varepsilon.$$

□

We sum up the procedure of proof of Theorem 2.2.5. For a giving $x \in X$ and $\varepsilon > 0$, the target is to construct $f_\varepsilon(x)$ the image of x. A neighbourhood $B(x, \delta)$ is constructed firstly for the upper semi-continuous, then an open covering $\left\{B\left(x, \frac{\delta}{4}\right)\right\}$ is applied to yield a locally finite fined open covering $\{U_\alpha\}$. $\{U_\alpha\}$ yields a Lipschitzian unit decomposition and a set $I(x)$ which is finite. By $i \in I(x)$, x_i is found. At last $m(F(x_i))$ is used to construct the $f_\varepsilon(x)$. To prove the conclusion, let $\delta_{i_0} = \max\{\delta_i, i \in I(x)\}$, then $B(x_{i_0}, \delta_{i_0}) \supset B\left(x_i, \frac{\delta_i}{4}\right)$. It leads to $m(F(x_i)) \in F(x_{i_0}) + \frac{\varepsilon}{2} B_Y$ and $f_\varepsilon(x) \in F(x_{i_0}) + \frac{\varepsilon}{2} B_Y$.

It is meaningful to compare Theorem 2.2.5 with Theorem 2.2.4. Theorems 2.2.4 and 2.2.5 do not require the spaces X and Y to be finite dimensions. But, Theorem 2.2.5 does not require $F(x)$ is bounded, because the vector selected in Theorem 2.2.5 is $m(F(x_i))$ which holds determined meaning. However, Theorem 2.2.5 only provides an approximate selection.

A mapping $\varphi : X \to Y$ is said to be locally compact if for every $x \in X$ there is a compact set $K_x \subset Y$.[3] and a $\delta > 0$ such that $\varphi(B(x, \delta)) \subset K_x$. Accordingly, we can define locally bounded. Let $\varphi_\varepsilon : X \to Y$ be mappings, $\varphi_\varepsilon(x)$ is locally equicompact if there is a $\varepsilon_0 > 0$, when $\varepsilon_0 > \varepsilon > 0$, there exists a δ which may depend on x but be independent of ε such that $\varphi_\varepsilon(B(x, \delta)) \subset K_x$.

Corollary 2.2.2 If all of the conditions of Theorem 2.2.5 are satisfied and the mapping $x \mapsto m(F(x))$ is locally compact, then the selection $f_\varepsilon(x)$ is locally equicompact.

Proof Because $x \mapsto m(F(x))$ is locally compact, by the definition of local compactness, there is an $\eta_x > 0$, such that $m\Big(F(B(x, \eta_x)) \subset C_x$ where C_x is a compact set. Take $\varepsilon_0 = \frac{\eta_x}{2}$ and $\delta_0 = \frac{\eta_x}{4}$. Now let $\varepsilon \in (0, \varepsilon_0)$, we prove

[3] The subscript x is used to express that the compact set K depends on variable x. η_x has the same meaning.

$f_\varepsilon(B(x,\delta_0)) \subset K_x$. For every $\overline{x} \in B(x,\delta_0)$, the above proof asserts $f_\varepsilon(\overline{x}) = \sum p_i(\overline{x}) m(F(\overline{x}_i))$, where $\overline{x}_i \in B\left(\overline{x}_i, \frac{\delta_i}{4}\right)\Big) \subset B\left(\overline{x}, \frac{\delta_{i_0}}{2}\right) \subset B\left(\overline{x}, \frac{\delta_0}{2}\right)$. The last containing relation comes from the fact that all $\delta(x) < \frac{\varepsilon}{2} < \frac{\varepsilon_0}{2} = \delta_0$. Furthermore, $\overline{x} \in B(x,\delta_0)$, hence $\overline{x}_i \in B(x,2\delta_0) \subset B(x,\eta_x)$ and $m(F(\overline{x}_i)) \in C_x$. Let $K_x = \mathrm{co}C_x$. Then K_x is compact and $f_\varepsilon(\overline{x}) \in K_x$. □

2.2.4 Theorems for Fixed Points

To end this section, we deal with the fixed points. The existence of fixed points is an important issue in both topology and functional analysis. We will extend these results to set-valued mappings. We start with a new definition for selection.

Definition 2.2.3 Let $F : X \to Y$ be a set-valued mapping, and $X, \ Y$ be two normed spaces. If there exist set-valued mappings $F_k : X \to Y, \ k = 1, 2, \ldots$ such that

(1) For every k, gra F_k is a closed set, and F_k holds a continuous selection;

(2) For every $x \in X, \ \ F_{k+1}(x) \subset F_k(x)$ and $\bigcap_{k=1}^{\infty} F_k(x) = F(x)$;

then the F is said to be $\sigma-$ selectable and $\{F_k, k = 1, 2, \ldots\}$ is a $\sigma-$ selectable approximate sequence. □

By Theorem 2.2.4, if $F : \mathbb{R}^n \to \mathbb{R}^m$ is an upper semi-continuous mapping, and it is bounded and with closed and convex value, then F is $\sigma-$ selectable.

Definition 2.2.4 Let $f : X \to X$ be a single-valued mapping. $x \in X$ is said to be a fixed point of f, if $f(x) = x$.

Let $F : X \to X$ be a set-valued mapping. $x \in X$ is said to be a fixed point of F, if $x \in F(x)$. □

It is obvious, from Definition 2.2.4, that the concept of fixed point for set-valued mapping is exactly an extension of that for single-valued mapping.

Theorem 2.2.6 is known as Brouwer fixed point theorem and is a famous result.

Theorem 2.2.6 Let $A \subset \mathbb{R}^n$ be convex and compact set, $f : A \to A$ is a continuous single-valued mapping, then there is a $x \in A$ such that $f(x) = x$. □

The conclusion was extended to general normed spaces by Schauder. Their proofs are quite complicated and omitted.

Theorem 2.2.7 Let $A \subset \mathbb{R}^n$ be a convex and compact set, $F : A \to A$ be a $\sigma-$ selectable set-valued mapping, then there is a $x \in A$ such that $x \in F(x)$.

Proof Let $\{F_k(x)\}$ be the $\sigma-$ selectable approximate sequence and $f_k(x)$ is a continuous selection of $F_k(x)$. Consider a continuous mapping from A to A defined as $x \mapsto \pi(f_k(x), A)$. By Theorem 2.2.6, there is a fixed point $x_k \in A$ such that $x_k = \pi(f_k(x_k), A)$, i.e., $||f_k(x_k) - x_k|| = d(f_k(x_k), A)$.

Because $F_k \downarrow F$, when $m \leq k$, we have $f_k(x_k) \in F_k(x_k) \subset F_m(x_k)$ and

$$d(x_k, F_m(x_k)) \leq d(x_k, F_k(x_k)) \leq ||f_k(x_k) - x_k|| = d(f_k(x_k), A)$$

A is a compact set and $\{x_k\} \subset A$, hence, $\{x_k\}$ has an accumulation $\overline{x} \in A$. Without loss of generality, we assume $x_k \to \overline{x}$. By Problem 5 of the last section, F_m is an upper semi-continuous set-valued mapping; consequently, for a given $\varepsilon > 0$, we have $F_m(x_k) \subset F_m(\overline{x}) + \varepsilon B_n$ for some large $k\,(> m)$. Thus,

$$d(\overline{x}, F_m(\overline{x})) \leq ||\overline{x} - x_k|| + d(x_k, F_m(x_k)) + \varepsilon \leq ||\overline{x} - x_k|| + d(f_k(x_k), A) + \varepsilon \tag{2.2.5}$$

$F_m \downarrow F$, it leads to $F_m(\overline{x}) \subset F(\overline{x}) + \varepsilon B_n$ for some large m. Therefore,

$$f_k(x_k) \in F_k(x_k) \subset F_m(x_k) \subset F_m(\overline{x}) + \varepsilon B_n \subset F(\overline{x}) + 2\varepsilon B_n \subset A + 2\varepsilon B_n$$

i.e., $d(f_k(x_k), A) \leq 2\varepsilon$. Substituting it into Inequality (2.2.5) results in $d(\overline{x}, F_m(\overline{x})) \leq 4\varepsilon$. By $F_m(\overline{x}) \subset F(\overline{x}) + \varepsilon B_n$ again, $d(\overline{x}, F(\overline{x})) \leq 5\varepsilon$. Because ε can be selected arbitrarily, we obtain $\overline{x} \in F(\overline{x})$. □

By Theorem 2.2.4, a direct corollary can be obtained.

Corollary 2.2.3 Let $A \subset \mathbb{R}^n$ be a convex and compact set, $F : A \to A$ be an upper semi-continuous set-valued mapping and with convex and closed value. Then F holds a fixed point on A. □

It $\mathbb{R}^n$ is replaced by a Banach space Y, Corollary 2.2.3 is still valid and known as Kakutani fixed point theorem. We restate it below.

Theorem 2.2.8 Let Y be a Banach space and $A \subset Y$ be convex and compact set. If $F : A \to A$ is an upper semi-continuous set-valued mapping and with closed and convex value, then F holds a fixed point on A. □

Problems

1. If $A_1, A_2 \subset \mathbb{R}^n$ and $A_1 \subset A_2 + \varepsilon B$, then for every $x_0 \in \mathbb{R}^n$, $d(x_0, A_1) \leq d(x_0, A_2) + \varepsilon$.
2. Let X be a normed space, $A \subset X$ be a set, then $d(x, A)$ is a Lipschitzian functional of $X \to \mathbb{R}$ and its Lipschitzian constant can be 1.
3. If $F : \mathbb{R}^n \to \mathbb{R}^m$ is a Lipschitzian set-valued mapping and with closed and convex value. Furthermore, for every $x \in \mathbb{R}^n$, $F(x)$ is bounded then $F(x)$ is $\varepsilon-$continuous.
4. Prove Lemma 2.2.3.
5. Let $A \subset \mathbb{R}^n$ be a convex and compact set, x_0 be an inner point of A. $f : A \to \mathbb{R}^n$ is a continuous single-valued mapping. If for every $x \in \text{bd } A$, $\|f(x) - x\| \leq \|x_0 - x\|$ then $x_0 \in f(A)$. (Hint: applying Brouwer fixed point theorem and Minkowski function.)

6. Give an example to show that $f_\varepsilon(x) \in F(x) + \varepsilon B$ is not a necessary condition for the approximate selection.
7. Suppose that all conditions given in Theorem 2.2.4 are satisfied, then for every $\varepsilon > 0$, there exists a single-valued mapping $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}^m$ such that $f_\varepsilon(x) \in F(x) + \varepsilon B$. (Hint: showing the set-valued mapping constructed in Theorem 2.2.4 is also lower semi-continuous)
8. Try to verify Theorem 2.2.4 is still valid if all conditions given in Theorem 2.2.4 hold but F is replaced that $F : X \to Y$ where X and Y are all Banach spaces.

2.3 Differential Inclusions and Existence Theorems

This section deals with differential inclusion. It starts with the definition which is presented by a comparison with the differential equation. Then motivation of investigation of differential inclusions is presented. The main content of this section is existence theorems of solutions of differential inclusions. At last we extend the conclusion to time-delayed differential inclusions. The extension can show the advantages of differential inclusion theory.

2.3.1 Differential Equations and Differential Inclusions

The differential equation considered takes an explicit form that

$$\dot{x} = f(t, x) \tag{2.3.1}$$

where $f : [t_0, t_1] \times \mathbb{R}^n \to \mathbb{R}^n$ is a single-valued mapping, $t \in [t_0, t_1] \subset \mathbb{R}\,(\geq 0)$ is the interval of time, $x : \left[t_0, t_f\right] \to \mathbb{R}^n$ is a derivable function to be solved and $\dot{x}$ is its derivate related to time t. $[t_0, t_f]$ is the interval where the solution exists. t_0 is the initial time and $t_f \leq t_1$ is the final time. From geometrical viewpoint, $f(t, x)$ gives a vector field in $\mathbb{R}^n$. Solving Eq. (2.3.1) is equivalent to looking for a curve which, in time t, passes through $x \in \mathbb{R}^n$, with its tangent $f(t, x)$. Thus every vector in $\mathbb{R}^n$ yields a curve, i.e., a solution. If Eq. (2.3.1) satisfies the uniqueness condition, then these curves do not intersect. The fact illustrates the solution of Eq. (2.3.1) contains a vector of $\mathbb{R}^n$ as its parameter. Such a solution is called general solution. Usually for the initial time t_0 the starting vector $x(t_0)$ is assigned. When the initial condition is given, Eq. (2.3.1) becomes

$$\dot{x} = f(t, x), x(t_0) = x_0.$$

Solving this differential equation with initial condition is called Cauchy problem. The solution of Cauchy problem is called the special solution.

If f in Eq. (2.3.1) does not explicitly contain the variable t, i.e.,

$$\dot{x} = f(x) \tag{2.3.2}$$

the equation is then time-invariant. For the time-invariant differential equation, the initial time is usually fixed at $t_0 = 0$.

The primitive target for the investigation of differential equation is to get the general solution or special solution $x(t)$ in an analytic form. As well-known for almost all differential equations, we have no effective method to obtain their analytic solutions. Then researchers cleverly transferred their energy to study the qualitative properties of the solutions such as existence, uniqueness, stability, continuities on the parameters, and so on.

If the function $f(t, x)$ in the right side of Eq. (2.3.1) is replaced by a set-valued mapping $F : [t_0, t_1] \times \mathbb{R}^n \to \mathbb{R}^n$, then Eq. (2.3.1) becomes

$$\dot{x} \in F(t, x) \tag{2.3.3}$$

Relation (2.3.3) is called differential inclusion and denoted by Inc. (2.3.3) from now on. Accordingly, if F is only a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$, then the differential inclusion is time-invariant. The book mostly considers the time-invariant case, and the inclusion extended from Eq. (2.3.2), i.e.,

$$\dot{x} \in F(x), \quad x(0) = x_0 \tag{2.3.4}$$

To solve Inc. (2.3.4) is called the Cauchy problem of differential inclusion. For the case, the initial time is fixed at $t_0 = 0$. We apply $x(t; [0, t_f],\ x_0)$ to denote a solution of Inc. (2.3.4). In the notation $x(t; [0, t_f],\ x_0)$, time t is the argument, x_0 is the initial condition, t_f is the final time and $[0, t_f]$ is the time interval where the solution exists. When there is no any confusion, we apply $x(t, x_0)$ or $x(t)$ to substitute the troublesome notation $x(t; [0, t_f],\ x_0)$ for simplicity. It is evident Inc. (2.3.4) holds many solution if $F(x)$ is a nontrivial set-valued mapping.

From the definition of the differential inclusion, one may suggest a way to solve the Cauchy problem of Inc. (2.3.4). Firstly, we try to find a selection $f(x) \in F(x)$, then solve the Cauchy problem of the differential equation $\dot{x} = f(x)$, $x(0) = x_0$. It is really a way that we used to prove many conclusions such as the existence of solutions; however, the scheme is almost impracticable to serve to solve the differential inclusions. At first, finding a satisfactory selection with an explicit description borders on a fantasy. Secondly, even if a simple selection has been obtained, we still lack a method to solve the Cauchy problem except some very special cases. Thirdly, even if a solution of the Cauchy problem is obtained, it still cannot analyze the properties of all solutions. Therefore such a way has been abandoned to solve Inc. (2.3.4) by almost all researchers.

There are two main differences from the investigation of the differential equation. At first, the solution of Inc. (2.3.4) is not unique; hence, we usually do not investigate its uniqueness of solution except some special form of Inc. (2.3.4).[4] Secondly, the

[4] The monotonous differential inclusions are a special case which will be studied at the last section of this chapter.

solution $x(t)$ of a differential equation is required to be differentiable, and Eq. (2.3.1) should be held for every $t \in \left[t_0, t_f\right]$. But for the differential inclusion, the solution $x(t)$ is only required to be absolutely continuous; hence, it allows at some time $x(t)$ is not be differentiable. When $x(t)$ is not differentiable, the left side of Inc. (2.3.4) becomes meaningless. Hence, we only require Inc. (2.3.4) is valid for almost all t in the interval $[0, t_f]$, i.e., it is allowed Inc. (2.3.4) to fail on a set whose measure is zero. The relaxed requirement brings lots advantages for the research of differential inclusions.

$S_{[0,T]}(F, x_0)$ is used to denote the set of solutions of Inc. (2.3.4) where $T = \min\left\{t_f\right\}$ and $[0, T]$ is the common interval where the solutions all exist, i.e., $S_{[0,T]}\left(F, x_0\right) = \underset{\dot{x}\in F(x)}{\cup} x\left(t; [0, T], x_0\right)$. Moreover, the set

$$S_{[0,T]}\left(F, C\right) = \underset{\dot{x}\in F(x)}{\cup} \left\{x\left(t, [0, T], x_0\right), x_0 \in C\right\}$$

is denoted for the set of solutions whose initial values are in set $C \subset \mathbb{R}^n$. If the time interval is not a key issue in the discussing problem, we often omit the low subscript $[0, T]$.

The following example is given to illustrate the solutions of differential inclusion.

Example 2.3.1 Consider the following differential inclusion

$$\dot{x} \in [0.1,\ 0.3]\,x, \quad x(0) = 1\ . \tag{2.3.5}$$

The solutions of Inclusion (2.3.5) are full of the shadow area in Fig. 2.9. But it is not true that every curve in the shadow area is its solution. The trajectories in $S(F, 1)$ hold the following features: (1) the curve is derivable at almost all points on the curve; (2) all trajectories start at (0,1); (3) if one trajectory arrives at $A(t_1, x_1)$ (Fig. 2.9), then the trajectory will be restricted at the darker shadow area where the starting angle is restricted between $0.1x$ and $0.3x$. Hence the trajectory is always monotonically increasing; (4) two curves on the boundary of the shadow area, i.e., $x = e^{0.1t}$ and $x = e^{0.3t}$ are solutions, moreover, once a solution $x\left(t_1\right) = B$ where B (Fig. 2.9) is on the boundary, then the $x(t)$ is always at the boundary before t_1. □

We note that, throughout of the book, the set-valued mapping $F(x)$ in Inc. (2.3.4) is with convex and closed value. The integration adopted is in the meaning of Lebesgue, and $x(t) \in AC\left([0, T], \mathbb{R}^n\right)$ or $x(t) \in AC$ i.e., $x(t)$ is an absolutely continuous single-valued mapping.

2.3.2 *Why Do We Propose Differential Inclusions?*

Maybe there is a question for some readers: Why do we need to consider the differential inclusions? The differential equation is quite powerful so that lots of natural phenomena can be described by differential equations precisely. For

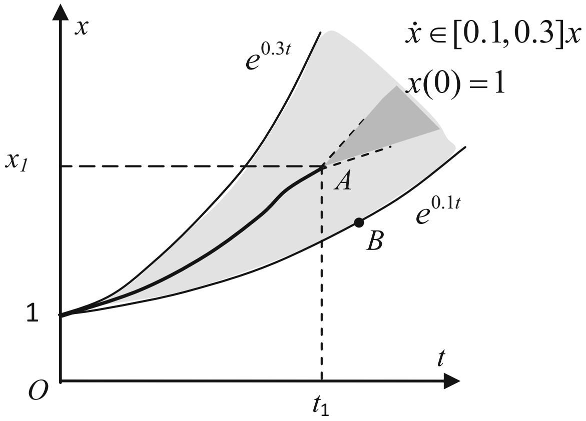


Fig. 2.9 The solutions area of Example 2.3.1

example, the motion of objects can be described by using Newton's laws which are described by differential equations. Another example is about the circuit. Because capacitance, inductance and operational amplifier all can be depicted by differential (or integral) equations, almost all circuits can be described by using differential equations. Kirchhoff's law provides a method to obtain the differential equation. Hence, do we really need differential inclusions?

The following examples are given to explain the reason why we really need to introduce differential inclusions.

1. Need from real systems

It was known for us every mechanical system suffers from friction and every modern electronic system has to apply diodes. These two basic elements should be depicted by differential inclusions. We now give a detailed discussion with friction in mechanical system.

In 1902, Stribeck studied the dry friction (friction for the objects without using lubricant) of a friction pair. He found that the change of friction coefficient has four phases (Fig. 2.10), i.e., static friction which exists when no motion happens in the friction pair; Stribeck friction called by later researchers which happens in a very lower velocity, with the speed increasing the friction coefficient decreases; Coulomb friction, where the friction coefficient reaches its minimal value; viscous friction, where the friction efficient increases with the speed increasing. He then concluded that the friction is a very complicatedly nonlinear phenomenon. The most undetermined case is the first phase where the friction coefficient is zero but the function of friction exists.

In 2001, Glocker suggested the friction force λ should be depicted by the following set-valued mapping

$$\lambda \in F_f\left(\dot{q}\right) = -\mu\lambda_N \operatorname{sgn}\left(\dot{q}\right) + F_S\left(\dot{q}\right)$$

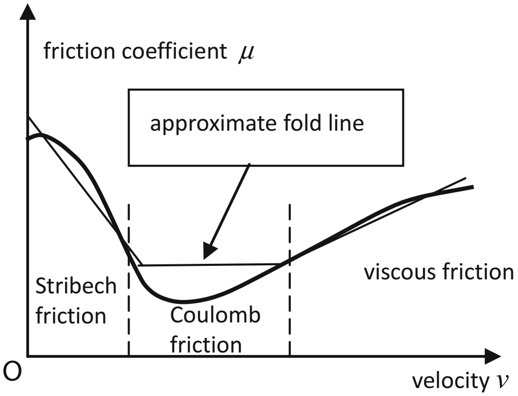


Fig. 2.10 Stribeck plot

where q is the displacement of object, $\dot{q}$ is then the velocity, the sliding friction $F_S\left(\dot{q}\right)$ is a single-valued function of $\dot{q}$, μ is its friction coefficient, λ_N is the pressure and sgn $\left(\dot{q}\right)$ is the sign function defined as follows

$$\operatorname{sgn}\left(\dot{q}\right) = \begin{cases} 1 & \dot{q} > 0, \\ [-1, 1] & \dot{q} = 0, \\ -1 & \dot{q} < 0. \end{cases} \tag{2.3.6}$$

Then the moving of the mechanical object which suffers from only dry friction can be described by the following equation

$$M\ddot{q} + D\dot{q} + Kq = Su + T\lambda \tag{2.3.7}$$

where M,D,K are the mass matrix, damping matrix and stiff matrix, respectively, q is genelized displacement which contains both translational motion and rotational motion, S is the input matrix and u is the control whose components are forces or moments, T is gain matrix for the friction and λ is the vector of friction where the ith component takes the form of

$$\lambda_i \in -\mu_i\lambda_{Ni}\operatorname{sgn}\left(T_i^T\dot{q}\right) + F_{Si}\left(T_i^T\dot{q}\right)$$

The relation is the friction that happens in the ith touch point. In the above relation, T_i is the ith column of T and $T_i^T\dot{q}$ is the relative sliding moving happens in the ith touch point.

Equation (2.3.7) can be rewitten by its state description

$$\begin{aligned} \dot{x} &= Ax + Gw + Bu, \\ z &= Hx, \\ y &= Cx, \\ w &\in -\varphi(z), \end{aligned} \tag{2.3.8}$$

where $x = \left[q^T \ \dot{q}^T \right]^T$ is the state, z and w are the input and output of set-valued mapping $\varphi(\cdot)$, respectively. The meanings of other matrices can be known from Eq. (2.3.7) and are omitted. Equation (2.3.8) is called Luré differential inclusion which will be dealt with in Chap. 5.

In classical nonlinear control theory, the nonlinear phenomenon depicted by Eq. (2.3.6) is called by relay nonlinearity, which is very popular in engineering area. Hence, lots of engineering systems can be described by using Luré differential inclusions.

2. Control system theory

Since 1980s, the development of differential inclusion is motivated by needs of control systems theory. Let us consider a state model of a control system. Generally its state equation is described by a differential equation

$$\dot{x} = f(x, u) \tag{2.3.9}$$

In the almost all textbooks of control theory, there is a paragraph of explanation attached to the equation: *in Eq.* (2.3.9), *x is the state which is usually an n-dimensional real vector, and u* $\in$ *U is an m-dimensional control input where U is the admissible control set*. The statement really suggests that if we apply differential inclusion, the state space model should be described as

$$\dot{x} \in F(x, U) = \underset{u \in U}{\cup} f(x, u) \tag{2.3.10}$$

Inc. (2.3.10) is usually called as state inclusion, by the state inclusion, the model of whole control system becomes

$$\begin{bmatrix} \dot{x} \\ y \end{bmatrix} \in \begin{bmatrix} F(x, U) \\ G(x, U) \end{bmatrix}$$

The second is the output mapping which is a set-valued mapping altered from output equation.

There are some effective investigation for the Inc. (2.3.10). Commonly, the control u is a mapping form time interval $\left[0, t_f\right] \subset \mathbb{R}(\geq 0)$ to $\mathbb{R}^r$; hence, U is a set of r-dimensional functions. Without loss of generality, it is commonly assumed that the U is a compact set of piecewise continuous functions. When U is a compact set, the continuity of $f(x, u)$ implies the continuity of set-valued mapping $F(x, U)$. Furthermore, if $f(x, u)$ is a uniform Lipschitzian mapping for every fixed u, then $F(x, U)$ is also a Lipschitzian set-valued mapping. These facts illustrate that $F(x, U)$ inherits lots of properties of $f(x, u)$, and $f(x, u)$ is only a continuous selection of $F(x, U)$. At problems of this section, we will give more properties of Inc. (2.3.10).

In the design of control system, the design target can be often transformed into performance of state $x(t)$. For example, the stability is that every trajectory of Inc. (2.3.10) converges to the origin (it is strongly asymptotically stable, to see Sect. 2.5). Let $S(2.3.10)$ denote the set of solutions of Inc. (2.3.10), and $x(t)$ be the ideal trajectory satisfying required performance. For the open-loop control, the design of control system consists of two steps. The first step is to check whether or not $x(t) \in S(2.3.10)$; if the first step succeeds, then the second is to find a $u(t) \in U$, such that $\dot{x}(t) = f(x(t), u(t))$. In case $x(t) \notin S(2.3.10)$, we have two ways to complete the design. One is to extend $x(t)$ to a set $\{x_\lambda(t), \lambda \in \Lambda\}$ where x_λ's are called admissible trajectories, and then to check whether $\{x_\lambda(t)\} \cap S(2.3.10) \neq \varnothing$. Then we try to find $x(t) \in \{x_\lambda(t)\}$ and $u(t) \in U$. The another is to enlarge the control set U such that $\{x_\lambda(t)\} \cap S(2.3.10) \neq \varnothing$. The introduction of slide mode control is a meaningful practice of extension of control ability. For the case of closed-loop control, the admissible control set U is replaced by $U(x)$ where $U(x)$ is the admissible feedback. Repeating the above statements, we have the design procedure of closed-loop system design. Of course, these only give an outline of system design, the check of $\{x_\lambda(t)\} \cap S(2.3.10) \neq \varnothing$ is very troublesome since we cannot obtain an explicit expression of $S(2.3.10)$. However, the analysis of $S(2.3.10)$ can give a guide for the design of controller.

Let us continue to consider Example 2.3.1, the controlled differential inclusion is

$$\begin{aligned}&\dot{x} \in [0.1,\ 0.3]x + u, \quad u \in U,\\ &x(0) = 1\,.\end{aligned}$$

The design target is to stabilize the system by feedback, i.e., the trajectory will trend to the t-axis as the time goes. Hence, it is necessary to make the shadow area in Fig. 2.9 can contain an area which is approximate t-axis as the t increases. Suppose, at time t, the trajectory arrives at $A(t_1, x_1)$. If we require the trajectory tends to zero, then the derivative of trajectory has to be less than zero, i.e., $u + 0.1x_1 < 0$, the x_1 can be selected arbitrarily; hence $U \supset \{u; u < -0.1x,\ t > 0, x > 0\}$ is a necessary condition (Fig. 2.11).

The above discussion shows that using differential inclusion description can extend the design thought for control systems.

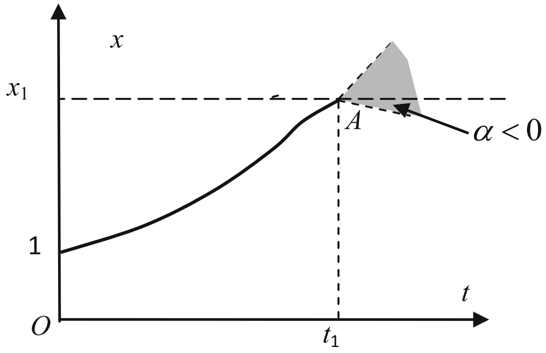


Fig. 2.11 The design of control set for Example 2.3.1

3. Discontinuous differential equations

We start with an example.

Example 2.3.2 Consider the Cauchy problem for the following discontinuous differential equation

$$\dot{x} = \begin{cases} 1 & x < 0, \\ -1 & x \geq 0, \end{cases} \quad x(0) = 0 \tag{2.3.11}$$

We consider the trajectory starting at the origin of the (t, x) plane. Because $\dot{x}(0) = -1$, the trajectory should enter the IV quadrant. However, as soon as the trajectory enters the IV quadrant, then $x<0$ and $\dot{x} > 0$. The trajectory will leave the IV quadrant and enter the I quadrant. In the I quadrant, $x>0$, and $\dot{x} < 0$, hence, the trajectory should return to the IV quadrant. It is really a trouble. The trajectory cannot also go ahead along with the t-axis since its derivative is nonzero. We conclude that the Cauchy problem has no solution. □

In Example 2.3.2, as long as $x(0) \neq 0$, the Cauchy problem has a solution. If the trajectory arrives at the t-axis at a finite time, the trajectory then finalizes.

To make these equation solvable at interval $[0,\ \infty)$, Filippov presented an improved scheme which is now called Filippov theory.

For a differential equation $\dot{x} = f(x)$, we define a set-valued mapping

$$F(x) = \bigcap_{\delta>0} \text{cl co} f\,(x + \delta B)$$

where B is the open unit ball of $\mathbb{R}^n$. $F(x)$ is called as the set-valued extension of $f(x)$, or the Filippov extension of $f(x)$. The Filippov extension holds the following properties: (1) $F(x)$ is with closed and convex value. Furthermore, if $f(x_0)$ is bounded, then F is upper semi-continuous at x_0 (Prob. 1–8 of this chapter). (2) If $f(x)$ is continuous at x, then $F(x) = \{f(x)\}$. In the next subsection, we will prove the above properties and prove that the Cauchy problem of differential inclusion

$$\dot{x} \in F(x), \quad x(0) = x_0$$

is always solvable. The solution of the above inclusion is called as Filippov solution of discontinuous differential equation $\dot{x} = f(x)$. The fact shows that Filippov extension is a powerful tool for discontinuous differential equations.

Let us return to Example 2.3.2. At $x = 0$, for every $\delta > 0, f\,(0 + \delta B) = \{-1, 1\}$. Hence, co$f\,(0 + \delta B) = [-1, 1]$, or $F(0) = [-1,\ 1]$. Then Eq. (2.3.11) becomes

$$\dot{x} \in F(x) = \begin{cases} -1 & x > 0, \\ [-1,\ 1] & x = 0, \\ 1 & x < 0. \end{cases}$$

$F(x)$ is continuous and its Cauchy problem $\dot{x} \in F(x),\ \ x(0) = x_0$ is solvable. The Filippov solution is

$$x(t) = \begin{cases} x_0 - \operatorname{sgn}(x_0)\, t & t \in [0, |x_0|]\,, \\ 0 & t \geq |x_0|\,, \end{cases}$$

where $\operatorname{sgn}(x)$ is the single-valued mapping and $\operatorname{sgn}(0) = 0$.

2.3.3 Existence Theorems of Solution of Differential Inclusions

This subsection deals with the existence of solution of differential inclusions. It is sufficient for us to show that $S\,(F, x_0) \neq \varnothing$ for Inc. (2.3.4). Because there are more than one definitions of continuity for set-valued mappings, we have to deal with the problem, separately.

In the theory of differential equation, the following conclusion is fundamental for the existence of the solution.

Consider the Cauchy problem of $\dot{x} = f(x),\ \ x(0) = x_0$, if there is neighborhood of x_0, $f(x)$ satisfies the Lipschitzian condition in the neighborhood, then there is a $\tau > 0$, the Cauchy problem is solvable at $[0, \tau]$. Furthermore, if there is neighborhood of $x(\tau)$, $f(x)$ satisfies Lipschitzian condition at that neighborhood, too. Then the solution can be extended.

By the method of using in the proof of Lemma 1.3.1, we can prove that if $f(x)$ is continuous at x_0 and x_0 is a inner point of domain of f, then $f(x)$ is a local Lipschitzian function at x_0. Therefore, some textbooks revise the theorem as *if $f(x)$ is continuous at x_0 and x_0 is an inner point of domain, then the Cauchy problem $\dot{x} = f(x),\ \ x(0) = x_0$ is locally solvable.*

We now turn to differential inclusion.

Theorem 2.3.1 Let X be a Hilbert space, and $F(x) : X \to X$ be an upper semi-continuous set-valued mapping and with closed and convex value. If the mapping $x \mapsto m\,(F(x))$ is locally compact, then the Cauchy problem of Inc. (2.3.4) has a solution.

Proof By Theorem 2.2.5, when F is an upper semi-continuous set-valued mapping and with closed and convex value, then F holds an approximate selection which is Lipschitzian. Hence for every $\varepsilon > 0$, there is a $f_\varepsilon(x)$ such that

$$f_\varepsilon(x) \in F(x) + \varepsilon B \tag{2.3.12}$$

Moreover, the selection has the form of

$$f_\varepsilon(x) = \sum_i p_i(x) m\,(F\,(x_i))$$

where $p_i(x)$ is a Lipschitzian unit decomposition, $m(F(x_i))$ is the minimal norm element of $F(x_i)$.

Let $\{\varepsilon_n\}$ be a monotonously decreasing series with that $\varepsilon_n \to 0 \ (n \to \infty)$. For every ε_n there is an approximate selection $f_{\varepsilon_n}(x)$ (for simplicity, denoted by $f_n(x)$). $f_n(x)$ is a locally Lipschitzian mapping; hence, the Cauchy problem $\dot{x} = f_n(x), \ \ x(0) = x_0$ has a solution $x_n(t)$.

By Corollary 2.2.2, $\{f_n(x)\}$ is locally equicompact, i.e., there is a $\delta > 0$ and $N \in \mathbb{N}$, such that for $n > N, f_n\left(B\left(x_0, \delta\right)\right) \subset K$, for a compact set K. It is equivalent to that $\dot{x}_n(t) \in K, t \in [0, T)$ for some $T > 0$. K is bounded hence the sequence $\{x_n(t)\}$ is equicontinuous at least for $n > N$. By Arzela-Ascoli theorem (to see Sect. 1.1 of this book), there are an absolutely continuous function $x(t)$ and a subsequence $\{x_{n_k}(t)\}$ of $\{x_n(t)\}$ such that $x_{n_k}(t)$ converges to $x(t)$ uniformly. By Theorem 1.1.7 (Alaoglu theorem), $\left\{\dot{x}_{n_k}(t)\right\}$ holds a weakly $*-$ convergent subsequence, for simplicity, without loss of generality; $\left\{\dot{x}_{n_k}(t)\right\}$ is the weakly $*-$ convergent sequence. Thus, there is a $\delta > 0$ and $v(t) \in L^1\,[0, \ \delta)$ such that $\int\limits_0^t \dot{x}_{n_k}(\tau)\,d\tau \to \int\limits_0^t v(\tau)\,d\tau$.

On the other hand, $x_{n_k}(t) = x_0 + \int\limits_0^t \dot{x}_{n_k}(\tau)\,d\tau$. Let k trend to infinite. We have $x(t) = x_0 + \int\limits_0^t v(\tau)\,d\tau$. Hence, $x(t) \in AC$ and $\dot{x}(t) = v(t)$ almost everywhere.

The approximate selection $f_{n_k}(x)$ satisfies that gra $f_{n_k} \subset$ gra $F + \varepsilon_{n_k} B$, i.e.,

$$d\left[\left(x_{n_k}(t), f_{n_k}(x(t))\right), \mathrm{gra} F(x)\right] \le \varepsilon_{n_k}$$

Let k trend to infinite again. We have $d\left[\left(x(t), \dot{x}(t)\right), \mathrm{gra} F(x)\right] = 0$, i.e., $\dot{x}(t) \in F\left(x(t)\right)$. And $x_{n_k}(0) = x_0$, consequently, $x(0) = x_0$. Thus, we prove the theorem. □

We now apply Theorem 2.3.1 to verify the existence of Filippov solution for discoutinuous differential equations.

Theorem 2.3.2 Consider the Cauchy problem of differential equation

$$\dot{x} = f(x), x_0 \in \mathbb{R}^n,$$

where $f(x)$ is bounded, but may be discoutinuous. Its Filippov extension is

$$F(x) = \bigcap_{\delta > 0} \mathrm{clco}\, f\,(x + \delta B)$$

Then the Cauchy problem differential inclusion

$$\dot{x} \in F(x), x_0 \in \mathbb{R}^n,$$

exists a solution. Moreover, let $x(t, x_0)$ be any solution of the Fillippov extension, then $\dot{x}\,(t, x_0) = f\,(x\,(t, x_0))$ if $f(x)$ is continuous at $x(t, x_0)$.

Proof Because $f(x)$ is bounded by the definition $F(x) = \bigcap\limits_{\delta>0} \text{cl co} f\,(x + \delta B)$, $F(x)$ is bounded and with closed and convex value. The local compactness of $x \mapsto m\,(F(x))$ is guaranteed. We now verify the following two facts.

(1) If $f(x)$ is continuous at x_1, then $F\,(x_1) = \{f\,(x_1)\}$.

Let $y_1 \neq f\,(x_1)$. Then we denote $\varepsilon = \|y_1 - f\,(x_1)\|$, there is a $\delta > 0$ such that $f\,(x_1 + \delta B) \subset B\left(f\,(x_1)\,,\ \frac{\varepsilon}{2}\right)$ since $f(x)$ is continuous at x_1. Hence $y_1 \notin f\,(x_1 + \delta B)$. Let z_i, $i = 1, 2, \ldots, n$ be arbitrarily n elements in $f\,(x_1 + \delta B)$. Then for $\lambda_i \geq 0$, $\sum\limits_{i=1}^{n} \lambda_i = 1$, we have

$$\left\| \left(\sum_{i=1}^{n} \lambda_i z_i \right) - f\,(x_1) \right\| \leq \sum_{i=1}^{n} \lambda_i \, \|z_i - f\,(x_1)\| \leq \frac{\varepsilon}{2}$$

i.e., $y_1 \notin \text{co} f\,(x_1 + \delta B)$. Furthermore, for $z \in \text{co} f\,(x_1 + \delta B)$,

$$\|y_1 - z\| \geq \|y_1 - f\,(x_1)\| - \|f\,(x_1) - z\| \geq \frac{\varepsilon}{2}$$

i.e., $y_1 \notin \ \text{clco} f\,(x_1 + \delta B)$

However, for every $\delta > 0$, $f\,(x_1) \in f\,(x_1 + \delta B)$, moreover, $f\,(x_1) \in \text{clco} f\,(x_1 + \delta B)$. It follows $f\,(x_1) \in F\,(x_1)$.

(2) $F(x)$ is upper semi-continuous.

This is the result of Problem 8 in the first section of this chapter.

Applying Theorem 2.3.1, the Cauchy problem of differential inclusion $\dot{x} \in F(x)$, $x(0) = x_0$ has a solution. Moreover, by the first fact proved in this theorem, we conclude $\dot{x}\,(t, x_0) = f\,(x\,(t, x_0))$ if $f(x)$ is continuous at $x(t, x_0)$. □

Theorem 2.3.1 needs a quite strong condition, the local compactness of mapping $x \mapsto m\,(F(x))$, that is difficult to be checked. The following conclusion is about the Lipschitzian differential inclusion. From the conclusion, we can feel that the Lipschitzian condition is critical for both differential equation and differential inclusion.

We will give a theorem which is very useful in the theory of differential inclusion. Before stating the theorem, we present so-called Granwall inequality.

Lemma 2.3.1 Suppose that $x(t) \in AC\,([0, T]\,, \mathbb{R})$. If

$$\left|\dot{x}(t)\right| \leq l(t)\,|x(t)| + \rho(t),$$

for two single-valued functions $l(t),\ \rho(t) \in L_1\,([0, T]\,, \mathbb{R}\,(\geq 0))$. Then for $t \in [0, T]$, we have

$$|x(t)| \leq e^{\int\limits_0^t l(s)ds} |x(0)| + \int\limits_0^t \rho(s) e^{\int\limits_0^t l(\tau)d\tau - \int\limits_0^s l(\tau)d\tau} ds. \tag{2.3.13}$$

□

Lemma 2.3.1 is known as Gronwall inequality. It is a fundamental integral inequality. Reference (Cai et al. 2009) gives a proof for the inequality. We also leave it to readers as an example since the skill of proof is fundamental in math.

Theorem 2.3.3 Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a closed Lipschitzian set-valued mapping with Lipschitzian constant l, $M \subset AC\left([0, T], \mathbb{R}^n\right)$ be a set, and $r_0 : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous single-valued mapping. If there is a function $\rho(t) \in L_1\left([0, T], \mathbb{R}\left(\geq 0\right)\right)$ such that for every $x(t) \in M$,

$$d\left(\dot{x}(t), F\left(x(t)\right)\right) \leq \rho(t),$$

and a real number $\delta \in \mathbb{R}\left(\geq 0\right)$ such that $||r_0(x) - x|| \leq \delta$ for every $x \in \mathbb{R}^n$. Then, there is a mapping $r : M \to S_{[0,T]}\left(F, C\right)$[5] where $C = \{r_0\left(x(0)\right), x(t) \in M\}$ is a set of $\mathbb{R}^n$. The mapping r holds the following properties:

1. r is continuous;
2. $r\left(x(t)\right)(0) = r_0\left(x(0)\right)$ for every $x(t) \in M$;
3. If $x(t) \in M \cap S\left(F, C\right)$ and $r_0\left(x(0)\right) = x(0)$, then $r\left(x(t)\right) = x(t)$;
4. Denote $\xi(t) = \delta e^{l|t|} + \left|\int\limits_0^t e^{l(|t|-|s|)}\rho(s)ds\right|$,[6] then

$$||r\left(x(t)\right) - x(t)|| \leq \xi(t), \tag{2.3.14}$$

and

$$\left|\left|\frac{d}{dt} r\left(x(t)\right) - \dot{x}(t)\right|\right|_{\mathbb{R}^n} \leq l\xi(t) + \rho(t) \tag{2.3.15}$$

for every $t \in [0, T]$.

Proof This is a constructive proof, it consists of four steps.

(1) Construction of function sequence $\{x_k(t)\}$

Let $x(t) \in AC\left([0, T], \mathbb{R}^n\right)$. Then a mapping $\omega\left(x(t)\right) = \pi\left(\dot{x}(t), F\left(x(t)\right)\right)$ is defined. Because

[5] Because $S_{[0,T]}\left(F, C\right) \subset AC\left([0, T], \mathbb{R}^n\right)$, r is a mapping from $AC\left([0, T], \mathbb{R}^n\right)$ to $AC\left([0, T], \mathbb{R}^n\right)$.

[6] In the current situation, the absolute value is useless. However, if the interval $[0,\ T]$ is replaced by $[-T,\ 0]$, all conclusions of the theorem are still valid. For convenience of later use, we here apply absolute $|t|$ to substitute t.

$$\begin{aligned}\|\omega(x(t))\| &\le \left\|\dot{x}(t)\right\| + \left\|\omega(x(t)) - \dot{x}(t)\right\| \\ &= \left\|\dot{x}(t)\right\| + \left\|\pi\Big(\dot{x}(t), F(x(t))\Big) - \dot{x}(t)\right\| \\ &= \left\|\dot{x}(t)\right\| + d\Big(\dot{x}(t), F(x(t))\Big) \\ &\le 2\left\|\dot{x}(t)\right\| + d\Big(0, F(x(t))\Big) \\ &\le 2\left\|\dot{x}(t)\right\| + d\Big(0, F(x(0))\Big) + l\,\|x(t) - x(0)\|,\end{aligned}$$

where we have applied the fact that $F(x)$ is with closed value in the third step, the triangle inequality $d\left(\dot{x}(t), F(x(t))\right) \le d\left(0, \dot{x}(t)\right) + d\big(0, F(x(t))\right)$ in the fourth step, and the Lipschitzian condition $F(x(t)) \subset F(x(0)) + l\,\|x(t) - x(0)\|\,B$ in the fifth step. $x(t) \in AC$, hence $\dot{x}(t) \in L_1[0, T]$, moreover, $\omega(x(t)) \in L_1([0, T], \mathbb{R}^n)$. It is obvious from the definition, if $x(t) \in S_{[0,T]}(F, C)$, then $\omega(x(t)) = \dot{x}(t)$.

Using the function $\omega(\cdot)$, a function sequence $\{x_k(t); k = 0, 1, 2, \ldots\}$ is constructed as follows:

$$\begin{aligned} & x_0(t) \in M, \\ & x_1(t) = r_0(x_0(0)) + \int_0^t \omega(x_0(\tau))\,d\tau, \\ & x_2(t) = x_1(0) + \int_0^t \omega(x_1(\tau))\,d\tau, \\ & \cdots \\ & x_{k+1}(t) = x_k(0) + \int_0^t \omega(x_k(\tau))\,d\tau, \\ & \cdots \end{aligned}$$

Note that by the definition of $x_k(t)$, we have $x_k(0) = r_0(x_0(0))$ and $x_k(t) \in AC$ for every $k \in \mathbb{N}$. Moreover, if $x(t) \in S_{[0,T]}(F, C)$ and $r_0(x(0)) = x(0)$, then $x_k(t) \equiv x(t)$ for every k.

(2) Convergence of sequences $\{x_k(t)\}$ and $\left\{\dot{x}_k(t)\right\}$

At first,

$$\|x_1(t) - x_0(t)\| \le \|r_0(x_0(0)) - x_0(0)\| + \left\| \int_0^t \left(\omega(x_0(\tau)) - \dot{x}_0(\tau)\right) d\tau \right\| \le \delta + \int_0^t \rho(\tau)\,d\tau \tag{2.3.16}$$

where the constant δ and the function $\rho(t)$ are those given by the theorem.

Secondly,

$$\left\|\dot{x}_1(t) - \dot{x}_0(t)\right\| = \left\|\omega(x_0(\tau)) - \dot{x}_0(t)\right\| = d\Big(\dot{x}_0(t), F(x_0(t))\Big) \le \rho(t) \tag{2.3.17}$$

Now we extend the Inequalities (2.3.13) and (2.3.14) into the case of $k > 1$.

$$\begin{aligned}\left\|\dot{x}_{k+1}(t) - \dot{x}_k(t)\right\| &= \left\|\omega\left(x_k(t)\right) - \dot{x}_k(t)\right\| \\ &= d\left(\dot{x}_k(t), F\left(x_k(t)\right)\right) \\ &\le d\Big(\dot{x}_k(t), F\left(x_{k-1}(t)\right)\Big) + l\left\|x_k(t) - x_{k-1}(t)\right\| \\ &= l\left\|x_k(t) - x_{k-1}(t)\right\| \ .\end{aligned} \tag{2.3.18}$$

In the third step, we have applied triangle inequality and Lipschitzian condition. In the fourth step, we have applied the fact that $\omega\left(x_{k-1}(t)\right) = \pi(\dot{x}_{k-1}(t), F\left(x_{k-1}(t)\right) \in F\left(x_{k-1}(t)\right)$.

We now prove that

$$\left\|x_{k+1}(t) - x_k(t)\right\| \le \delta\frac{|lt|^k}{k!} + \int\limits_0^t \frac{\left(l\,|t| - |\tau|\right)^k}{k!}\rho\left(\tau\right)d\tau \tag{2.3.19}$$

and

$$\left\|\dot{x}_{k+1}(t) - \dot{x}_k(t)\right\| \le l\delta\frac{|lt|^{k-1}}{(k-1)!} + l\int\limits_0^t \frac{\left(l\,|t| - |\tau|\right)^{k-1}}{(k-1)!}\rho\left(\tau\right)d\tau \tag{2.3.20}$$

By Inequality (2.3.18), Inequality (2.3.19) implies Inequality (2.3.20). Hence, it is sufficient to verify Inequality (2.3.19).

For the case of $k = 0$, Inequality (2.3.19) degenerates to Inequality (2.3.16). Hence the inequality holds for the case of $k = 0$. We assume that Inequality (2.3.19) holds for the case of k. Now we prove that it is also true for the case of $k + 1$. Since $k \ge 1$,

$$\begin{aligned}\left\|x_{k+1}(t) - x_k(t)\right\| &= \left\|\int\limits_0^t \left(\dot{x}_{k+1}\left(\tau\right) - \dot{x}_k\left(\tau\right)\right)d\tau\right\| \\ &\le l\int\limits_0^t \left\|x_k\left(\tau\right) - x_{k-1}\left(\tau\right)\right\| d\tau \\ &\le l\int\limits_0^t \left(\delta\frac{|l\tau|^{k-1}}{(k-1)!} + \int\limits_0^\tau \frac{\left(l\,|\tau| - |q|\right)^{k-1}}{(k-1)!}\rho(q)dq\right)d\tau \\ &= \delta\frac{|lt|^k}{k!} + l\int\limits_0^t\int\limits_0^\tau \frac{\left(l\,|\tau| - |q|\right)^{k-1}}{(k-1)!}\rho(q)dqd\tau \ .\end{aligned}$$

Because $\frac{d}{d\tau}\int\limits_0^\tau \frac{(l\,|\tau|-|q|)^k}{k!}\rho(q)dq = \int\limits_0^\tau l\frac{(l\,|\tau|-|q|)^{k-1}}{(k-1)!}\rho(q)dq,$

$$\delta\frac{|lt|^k}{k!} + l\int\limits_0^t\int\limits_0^\tau \frac{(l\,|\tau|-|q|)^{k-1}}{(k-1)!}\rho(q)dqd\tau = \delta\frac{|lt|^k}{k!} + \int\limits_0^t \frac{(l\,|t|-|q|)^k}{k!}\rho(q)dq.$$

Replacing q by τ, Inequality (2.3.19) is then verified.

Because $e^z = \sum\limits_{k=0}^{\infty}\frac{z^k}{k!}$, for every real number z, $\sum\limits_{k=m}^{\infty}\frac{z^k}{k!} \to 0$, $(m\to\infty)$. Therefore on the time interval $[0, T]$, both $\{x_k(t)\}$ and $\left\{\dot{x}_k(t)\right\}$ are Cauchy sequences.

(3) Definition of mapping r

From Step (2), for every $x_0(t)\in M$, we have constructed a sequence $\{x_k(t)\}$ such that $\dot{x}_k(t)\to v(t)$, $(k\to\infty)$ for a $v(t)\in L_1$. Furthermore if we treat two functions are equivalent when there are identical except at a set whose measure is zero, then the limitation is unique. Now we define

$$r\left(x_0(t)\right) = r_0\left(x_0(0)\right) + \int\limits_0^t v\left(\tau\right)d\tau \tag{2.3.21}$$

then $r\left(x_0(t)\right)\in AC$ and $r\left(x_0(t)\right)(0) = r_0\left(x_0(0)\right)$.

(4) Proof of the conclusions

From Inequality (2.3.18), we know $\lim\limits_{k\to\infty} d\Big(\dot{x}_k(t), F\left(x_k(t)\right)\Big) = 0$. Hence,

$$\left\|r\left(x_0(t)\right) - x_k(t)\right\| = \left\|\int\limits_0^t \left(\dot{x}_k\left(\tau\right) - v(t)\right)d\tau\right\| \le \int\limits_0^t \left\|\dot{x}_k\left(\tau\right) - v(t)\right\| d\tau \to 0 \tag{2.3.22}$$

Moreover,

$$0 = \lim_{k\to\infty} d\left(\dot{x}_k(t), F\left(x_k(t)\right)\right) = d\Big(v(t), F\left(r\left(x_0(t)\right)\right) = d\Big(\frac{d}{dt}r\left(x_0(t)\right), F\left(r\left(x_0(t)\right)\right)$$

F is with closed value, consequently, $\frac{dr(x_0(t))}{dt}\in F\left(r\left(x_0(t)\right)\right)$ and $r\left(x_0(t)\right)\in S\left(F, r_0\left(x_0(0)\right)\right)$, where we apply $r\left(x_0(t)\right)(0) = r_0\left(x_0(0)\right)$. The second conclusion of Theorem 2.3.3 is obtained.

By Inequality (2.3.19),

$$\|x_{k+1}(t) - x_0(t)\| \le \delta e^{|lt|} + \int\limits_0^t e^{l|t|-|\tau|}\rho\left(\tau\right)d\tau = \xi(t)$$

Similarly, from Inequality (2.3.20), we obtained

$$\left\|\dot{x}_{k+1}(t)-\dot{x}_0(t)\right\| \leq l\delta e^{|lt|}+l\int_0^t e^{l|t|-|\tau|}\rho(\tau)\,d\tau+\rho(t)=l\xi(t)+\rho(t)$$

Thus,

$$\|r(x_0(t))-x_0(t)\| \leq \|x_k(t)-x_0(t)\|+\|r(x_0(t))-x_k(t)\| \leq \xi(t)+\|r(x_0(t))-x_k(t)\|$$

Applying the last limitation in Inequality (2.3.22), then it implies $\|r(x_0(t))-x_0(t)\| \leq \xi(t)$, i.e., Inequality (2.3.14) holds. By the same method, we can prove Inequality (2.3.15) holds. The Conclusion 4 of the theorem is then proved.

If $x_0(t) \in S(F, C)$, then $\omega\left(\dot{x}_0(t)\right)=\dot{x}_0(t)$, and $r_0(x_0(0))=x_0(0)$. We conclude $x_1(t)=x_0(t)$. By such a way, we obtain $x_k(t)=x_0(t),\ \ k=1,2,\ldots$. Consequently, $v(t)=\dot{x}_0(t)$, and $r(x_0(t))=x_0(t)$. The Conclusion 3 of the theorem is also verified.

To end the proof of the theorem, we consider the continuity of r. r is a mapping from M to the space of AC. If $x_0(t) \in M$, then

$$\|x_0(t)\|_{AC}=\|x_0(0)\|_{R^n}+\int_0^T\left\|\dot{x}_0(\tau)\right\|_{R^n}d\tau.$$

We can think that $x(t) \in B(x_0(t), \varepsilon)$, then $x(0) \in B(x_0(0), \varepsilon)$, $\left\|\dot{x}(t)-\dot{x}_0(t)\right\|<\varepsilon$.

At first, $\omega(x(t))=\pi\left(\dot{x}(t), F(x(t))\right)$ is a continuous mapping from $AC([0,T])$ to $AC([0,T])$ (to see Prob. 3 of this section). It follows that $y(t)=x(0)+\int_0^t\omega(x(\tau))\,d\tau$ is a continuous mapping from $x(t)$ to $y(t)$. By the reason, for every k, the mapping from $x_0(t)$ to $x_k(t)$ is continuous. $\dot{x}_k(t)$ is uniformly convergent to $v(t)$. Hence, the mapping from $x_0(t)$ to $v(t)$ is continuous. At last the mapping r defined by Eq. (2.3.21) is continuous.

Thus, we complete the proof of the theorem. □

Theorem 2.3.3 is a very important result in the theory of differential inclusion. It is a constructive proof for the solutions of differential inclusions. One can start with a set (or a function) in the space of absolutely continuous functions to obtain a solution of a differential inclusion. It is also a critical key to very problems. The following theorem is about the existence of the solution of Cauchy problem of differential inclusion, and the theorem is a natural result of Theorem 2.3.3.

Theorem 2.3.4 Let $F:\mathbb{R}^n \to \mathbb{R}^m$ be a Lipschitzian set-valued mapping. It is with closed value. Then the following Cauchy problem

$$\dot{x}(t) \in F(x(t)),\ \ x(0)=x_0$$

has a solution.

Proof Let $M = \{0\}$, and $r_0 = x_0 + x$. Take $\rho(t) \equiv d(0, F(0))$ and $\delta = \|x_0\|$. By Theorem $r(0)(0) = r_0(0) = x_{0,}$[7] and there exists a $x(t) = r(0) \in S_{[0,\infty]}(F, x_0)$. The $x(t)$ satisfies

$$\|x(t)\| \leq \|x_0\| e^{l|t|} + \int_0^t e^{l(|t|-|\tau|)} d(0, F(0))\, d\tau.$$

□

Theorem 2.3.4 only verifies that the solution set $S_{[0,\infty]}(F, x_0)$ is not empty. The theorem does not provide all elements of $S_{[0,\infty]}(F, x_0)$. From the proofs of Theorems 2.3.3 and 2.3.4, we can find that the time interval where the solution exists depends on the interval where the Lipschitzian condition holds. When the set-valued mapping is only a local Lipschitzian mapping, then the above theorem gives a local result that there is a $\delta > 0$ such that $S_{[-\delta,\delta]}(F, x_0) \neq \varnothing$. The conclusion is very similar to that for differential equations.

We give several remarks to Theorem 2.3.3.

Remark 1 In Theorem 2.3.3, we did not require that $F(x)$ is with convex value. We only require $F(x)$ is with closed value. □

Remark 2 Theorem 2.3.3 gives an algorithm to calculate solutions from a set of absolutely continuous functions. It is an approximate method. The key is to calculate the projection of $\dot{x}(t)$ on the set $F(x(t))$. From the theoretic viewpoint, the calculation is operable. □

Remark 3 Inequalities (2.3.14) and (2.3.15) give the errors between $x_0(t) \in M$, $\dot{x}_0(t)$, and its corresponding solution in $S_{[0,\infty]}(F, x_0)$ and its derivate. These errors depend on δ and $\rho(t)$. If $M = \{x_0(t)\}$, i.e., M contains only one function, and $x_0(0) = x_0$, then δ can be zero. Therefore, the less of the distance between $\dot{x}_0(t)$ and $F(x_0(t))$ is, the faster convergence of $x_0(t)$ to $r(x_0(t))$. □

Remark 4 As mentioned before, Theorem 2.3.3 allows that $T < 0$. This is the reason why we apply the notations of $|t|$ and $|s|$. □

As an application of Theorem 2.3.4, we can estimate the boundary of a solution. We have

$$\|x(t, x_0)\| = \|r(0)\| \leq \|x_0\| e^{l|t|} + \int_0^t e^{l(|t|-|s|)} d(0, F(0))\, ds \tag{2.3.23}$$

[7] In the notation $r(0)(0)$, the zero in $r(0)$ is the zero function in M, i.e., a constant function whose value is always zero. The anther 0 is the initial time.

We have $d(0, F(0)) = \|m(F(0))\|$. Suppose $t > 0$, then Inequality (2.3.23) leads to $\|x(t, x_0)\| \le \|x_0\| e^{lt} + \frac{1}{l} e^{lt-1} m(F(0))$. The highest increasing speed is e^{lt} which is similar to the corresponding conclusion for differential equation.

2.3.4 The Existence of Solutions of Time-delayed Differential Inclusions

Before ending this section, we spend some time to deal with the extension of the existence theorem for the solution of differential inclusions. In the theory of differential inclusions, it is only required that Inc. (2.3.4) holds only almost everywhere, i.e., it is allowed the inclusion does not hold on a set whose measure is zero. The property makes great convenience for the extension of the existence theorem. As an example, we deal with the time-delayed differential inclusion since time-delayed is very popular for almost all dynamic systems.

Suppose $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a closed Lipschitzian set-valued mapping. We consider the Cauchy problem for the following differential inclusion with time-delay.

$$\dot{x}(t) \in F(x(t), x(t-\tau)); \quad x(t) = \phi(t), \ t \in [-\tau, 0], \tag{2.3.24}$$

where $\tau \in \mathbb{R} (\ge 0)$ is a given constant and used to express the time delay, and $\phi(t) \in AC([-\tau, 0])$ is the initial condition. Because $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitzian set-valued mapping, there exist two constants l_1 and l_2 such that

$$d(F(x_1, y_1), F(x_2, y_2)) \le l_1 \|x_1 - x_2\| + l_2 \|y_1 - y_2\|$$

It implies $d(F(x, y), F(0, 0)) \le l_1 \|x\| + l_2 \|y\|$, or

$$d(0, F(x, y)) \le l_1 \|x\| + l_2 \|y\| + l_3, \tag{2.3.25}$$

where $l_3 = d(0, F(0, 0))$.

Theorem 2.3.5 Let $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ be a closed Lipschitzian set-valued mapping. Then Inc. (2.3.24) has a solution.

Proof Consider the following Cauchy problem,

$$\dot{x}(t) \in F(x(t), \phi(t-\tau)); \quad x(0) = \phi(0).$$

On the time interval $[0, \tau]$, $\phi(t-\tau)$ is a absolutely continuous function, hence, $F(x(t), \ \phi(t-\tau))$ can be treated as a time-varying differential inclusion. By an appropriate extension of Theorem 2.3.3, the set of solutions $S_{[0,\tau]}(F(x(t), \phi(t-\tau)), \phi(0))$ is not a empty set.

For a solution $x(t) \in S_{[0,\tau]}\left(F\left(x(t), \phi\left(t-\tau\right)\right), \phi(0)\right)$, we define

$$x_\tau(t) = \begin{cases} \phi(t), & t \in [-\tau, 0], \\ x(t), & t \in [0, \tau], \end{cases}$$

then $x_\tau(t) \in S_{[0,\tau]}$ (2.3.23) where, for simplicity, we apply the notation $S_{[0,\tau]}$(2.3.24) for the set of solutions of Inc. (2.3.24). By Inequality (2.3.23), $S_{[0,\tau]}$(2.3.24) is a bounded set (the fact can be also verified by Inequality (2.3.25), we leave it as an problem).

Now for every $x_\tau(t) \in S_{[0,\tau]}$ (2.3.24), we consider the Cauchy problem

$$\dot{x}(t) \in F\left(x(t), x_\tau\left(t-\tau\right)\right); \quad x\left(\tau\right) = x_\tau\left(\tau\right), \tag{2.3.26}$$

with initial time is τ. On the time interval $[\tau, 2\tau]$, $x_\tau\left(t-\tau\right)$ is a determined absolutely continuous function, hence by a similar discussion to the first step, $S_{[\tau,2\tau]}\left(F\left(x(t), x_\tau\left(t-\tau\right)\right), x_\tau\left(\tau\right)\right)$ is nonempty.

For every $x(t) \in S_{[\tau,2\tau]}\left(F\left(x(t), x_\tau\left(t-\tau\right)\right), x_\tau\left(\tau\right)\right)$, we define

$$x_{2\tau}(t) = \begin{cases} x_\tau(t), & t \in [-\tau, \tau], \\ x(t), & t \in [\tau, 2\tau], \end{cases}$$

then $x_{2\tau}(t) \in S_{[0,2\tau]}$ (2.3.23). $\left\{x_{2\tau}\left(\tau\right); \ x_{2\tau}(t) \in S_{[0,2\tau]} \text{ (2.3.23)}\right\}$ is still a bounded set, thus, step by step, we can extend the solution to $[0, \infty)$. □

Note that $x_\tau(t)$ is defined on $[-\tau, \ \tau]$. At the point $t = 0$, $x_\tau(t)$ may be not differentiable. The fact meets with the requirement of solutions of differential inclusion.

Problems

1. (1) Suppose $f(x, \ u)$ is continuous for every fixed $u \in U$, then $F(x, U)$ is lower semi-continuous;
   (2) Suppose $f(x, \ u)$ is continuous at every pair (x, u) and U is compact, then $F(x, U)$ is continuous.
2. Suppose $\{f_i\left(x, \ u\right), \ i = 1, 2, \ldots\}$ is series of continuous mappings from normed space X to Banach space Y, $\{U_i, \ i = 1, 2, \ldots\}$ is a series of compact and connect subsets of Y. If they satisfy that

   (1) $F(x) \subset \cdots \subset f_{n+1}\left(x, U_{n+1}\right) \subset f_n\left(x, U_n\right) \cdots \subset f_1\left(x, U_1\right)$ for every $x \in X$;
   (2) For every $\varepsilon > 0$, there is an $N \in \mathbb{N}$, when $n > N$, it holds $f_n\left(x, U_n\right) \subset F(x) + \varepsilon B$ where B is the unit ball in Y.

   Then $F(x)$ is upper semi-continuous and with compact and connect value. And

$$F(x) = \bigcap_{i=1}^{\infty} f_i\left(x, U_i\right)$$

3. Suppose $F(x)$ is the Filippov extension of $f(x)$. If $f(x)$ is only upper semi-continuous at x_0, then what kind continuity can hold the $F(x_0)$? If $f(x)$ is only lower semi-continuous at x_0, what is the $F(x_0)$?
4. Prove the Gronwell inequality (Hint: using the comparison principle of differential equation).
5. Prove that under the conditions of Theorem 2.3.3, Inequality (2.3.15) holds.
6. Suppose $F : \mathbb{R}^n \to \mathbb{R}^m$ is closed. Prove that $\omega\left(x(t)\right) = \pi\left(\dot{x}(t), F\left(x(t)\right)\right)$ which is defined in the proof of Theorem 2.3.3 is a continuous mapping from $L_1[0, T]$ to $L_1[0, T]$.
7. Prove the continuity of mapping r defined by Eq. (2.3.21) by using $\varepsilon - \delta$ language.
8. Suppose $F : \mathbb{R}^n \to \mathbb{R}^m$ is a closed and Lipschitzian set-valued mapping and $\upsilon_0 \in F\left(x_0\right)$. Then there is a solution $x(t) \in S_{[0,T]}\left(F, x_0\right)$ such that $\dot{x}(0) = \upsilon_0$. The fact means that we can appoint the initial value of the solution as well as the value of its derivate.
9. Prove that $\left\{x_\tau\left(\tau\right),\ x_\tau(t) \in S_{[0,\tau]}\left(2.3.23\right)\right\}$ which is defined in the proof of Theorem 2.3.5 is bounded.
10. Repeat the proof of Theorem 2.3.3 if $M = \{x_1(t)\}$ with $x_1(t) \in S_{[0T]}\left(F, x_0\right)$ but the image of M is $S_{[0T]}\left(F, x_1\right),\ \ x_1 \neq x_0$.

2.4 Qualitative Analysis for Differential Inclusions

We carry through the qualitative analysis of differential inclusions in this section. The study can be treated as an inheritance and development of the similar study for differential equations since they consider almost the same questions, but methods used and conclusions are quite different. This section only considers the local properties of differential inclusions. We start with those differential inclusions which hold Lipschitzian property, then turn to those which are upper semi-continuous, at last we deal with the problem of convexification of differential inclusions.

2.4.1 Qualitative Analysis for Lipschitzian Differential Inclusions

Lipschitzian differential inclusions are studied qualitatively by using Theorem 2.3.3 which is powerful tool in the theory of differential inclusions as mentioned before.

Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitzian set-valued mapping, it is with closed and convex value. By Theorem 2.3.4, the Cauchy problem of differential inclusion

$$\dot{x}(t) \in F\left(x(t)\right),\ \ x(0) = x_0 \tag{2.4.1}$$

holds solutions on the time interval $[0, T]$ (and/or $[-T, 0]$[8]). $S_{[-T,0]}\,(F, x_0)$ is the solution set of Inc. (2.4.1). In this subsection, three qualitative conclusions are presented for $S_{[0,T]}(F, x_0)$.

Theorem 2.4.1 Let $x_1(t), x_2(t) \in S_{[0,T]}\,(F, x_0)$ be two solutions. Then there is a continuous mapping $\phi : [0,1] \to S_{[0,T]}(F, x_0)$ such that $\phi(0) = x_1(t)$ and $\phi(1) = x_2(t)$.

Proof Define $M = \{(1-\lambda)\,x_1(t) + \lambda x_2(t); \lambda \in [0,1]\}$, then $M \subset AC\,([0,T]\,, \mathbb{R}^n)$. To use Theorem 2.3.3, we check the distance $d\Big((1-\lambda)\,\dot{x}_1(t) + \lambda\dot{x}_2(t), F\,((1-\lambda)\,x_1(t) + \lambda x_2(t))$.

$$\begin{aligned}
& d\left((1-\lambda)\,\dot{x}_1(t) + \lambda\dot{x}_2(t),\;\; F\,((1-\lambda)\,x_1(t) + \lambda x_2(t))\right) \\
& \leq d\left((1-\lambda)\,\dot{x}_1(t) + \lambda\dot{x}_2(t),\;\; \dot{x}_1(t)\right) + d\Big(\dot{x}_1(t), F\,((1-\lambda)\,x_1(t) + \lambda x_2(t))\,\Big) \\
& \leq \lambda\,\left\|\dot{x}_2(t) - \dot{x}_1(t)\right\| + d\left(\dot{x}_1(t), F\,((1-\lambda)\,x_1(t) + \lambda x_2(t))\right) \\
& \leq \lambda\,\left\|\dot{x}_2(t) - \dot{x}_1(t)\right\| + d\left(\dot{x}_1(t), F\,(x_1)\right) + d\,(F\,(x_1)\,, F\,((1-\lambda)\,x_1(t) + \lambda x_2(t))) \\
& \leq \lambda\,\left\|\dot{x}_2(t) - \dot{x}_1(t)\right\| + l\lambda\,\|x_2(t) - x_1(t)\| \\
& \leq \left\|\dot{x}_2(t) - \dot{x}_1(t)\right\| + l\,\|x_2(t) - x_1(t)\|\,,
\end{aligned}$$

where we have use the triangle inequality in the second and the fourth steps, respectively, and the facts that $d\left(\dot{x}_1, F\,(x_1)\right) = 0$, Lipschitzian condition and $\lambda \leq 1$.

Denote $\rho(t) = \left\|\dot{x}_2(t) - \dot{x}_1(t)\right\| + l\,\|x_2(t) - x_1(t)\|$. Then $\rho(t) \in L_1\,([0,T]\,, \mathbb{R})$ because $x_1(t), x_2(t) \in AC\,([0,T]\,, \mathbb{R}^n)$. Let $r_0(x) = x$. By Theorem 2.3.3, there is a continuous mapping $r : M \to S_{[0,T]}\,(F, x_0)$. Because $x_1(t), x_2(t) \in S_{[0,T]}\,(F, x_0)$, by Theorem 2.3.3 again, $r\,(x_1(t)) = x_1(t),\;\; r\,(x_2(t)) = x_2(t)$. Now we define

$$\phi\,(\lambda) = r\,((1-\lambda)\,x_1(t) + \lambda x_2(t))$$

$\phi(\lambda)$ is continuous and $\phi\,(\lambda) \in S_{[0,T]}\,(F, x_0)$ for all $\lambda \in [0,1]$. It follows $\phi(0) = x_1(t)$ and $\phi(1) = x_2(t)$. We have complete the proof. □

We have no reason to assert that the set $S_{[0,T]}(F, x_0)$ is convex, consequently, $(1-\lambda)\,x_1(t) + \lambda x_2(t)$ may fail to be a solution of Inc. (2.4.1).

Theorem 2.4.1 is usually called as arc connectedness theorem. It illustrates that if we treat two solutions $x_1(t)$ and $x_2(t)$ as two points of the vector space of absolutely continuous functions, then there is an arc which connects the two points such that every point at the arc is also the solution in $S_{[0,T]}(F, x_0)$. In $\mathbb{R} \times \mathbb{R}^n$, every solution in $S_{[0,T]}(F, x_0)$ can be drawn as a curve, the conclusion illustrate that there exists a

[8] We only consider the case of $[0,\; T]$ below. Readers are suggested to give and prove the similar conclusions for the case of $S_{[-T,0]}\,(F, x_0)$ as exercises.

curved surface which links with $x_1(t)$ and $x_2(t)$, and every point in the surface has passed at least one solution in $S_{[0,T]}(F, x_0)$ (Fig. 2.12).

Definition 2.4.1 Let F be set-valued mapping, the following set

$$R_{[0,T]}(F, x_0) = \left\{x(T); x(t) \in S_{[0,T]}(F, x_0)\right\}$$

is the T-reachable set of Inc. (2.4.1) with the starting point x_0. When T and x_0 are in unambiguous, the set is called reachable set of Inc. (2.4.1). □

By Definition 2.4.1, we can restate Theorem 2.4.1 as follows: in $\mathbb{R}^n$, $R_{[0,t]}(F, x_0)$ is arc wise connected for every $t \in [0, T]$.

Theorem 2.4.2 Let $x(t) \in S_{[0,T]}(F, x_0)$ be a solution of Inc. (2.4.1), and $x(T) = x^*$. If $x^* \in \text{bd } R_{[0,T]}(F, x_0)$, then for every $t_1 \in [0, T]$, $x(t_1) \in \text{bd } R_{[0,t_1]}(F, x_0)$.

Proof It is proved by contradiction. If there is a $t_1 \in (0, T)$ such that $x(t_1) \notin \text{bd } R_{[0,t_1]}(F, x_0)$, then by the definition of boundary, there is an $\varepsilon > 0$ such that $B(x(t_1), \varepsilon) \subset S_{[0,t_1]}(F, x_0)$.

On the other hand, Because $x^* \in \text{bd } R_{[0,T]}(F, x_0)$, there is a $y \in \mathbb{R}^n$ such that $y \notin R_{[0,T]}(F, x_0)$ and $\|y - x^*\| \leq \varepsilon e^{-l|t_1-T|}$. To apply Theorem 2.3.3, let $M = \{x(t)\}$ and $r_0(x) = x - x^* + y$, then $\rho(t) = 0$ and $\delta = \varepsilon e^{-l|t_1-T|}$. Theorem 2.3.3 asserts there is a solution $y(t) = r(x(t)) \in S_{[t_1,T]}(F, y_1)$[9] where $y_1 = y(t_1)$ which satisfies $y(T) = y$ (Fig. 2.13).

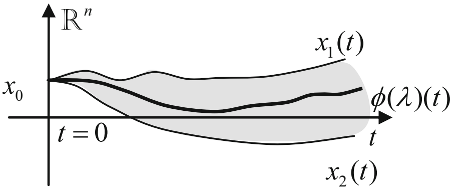


Fig. 2.12 The shadow part is contained in $S_{[0,T]}(F, x_0) \subset \mathrm{R} \times \mathrm{R}^n$

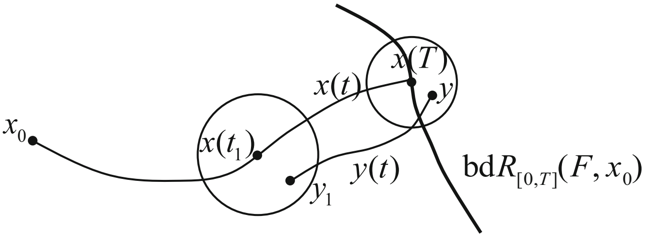


Fig. 2.13 Illustration of Theorem 2.4.2

[9] We apply Theorem 2.3.3 on the time interval $[T - t_1, T]$. The terminal is fixed at $y(T) = y$. The solution is $y(t) \in S_{[T-t_1,T]}(F, y_1)$, where $y_1 = y(T - t_1)$.

Using Inequality (2.3.14) in Theorem 2.3.3, we have $\|y(t_1) - x_1(t_1)\| \leq \|y - x^*\| e^{l(T-t_1)} < \varepsilon$. Hence, $y(t_1) \in R_{[0,t_1]}(F, x_0)$, i.e., there is a solution $x_2(t) \in S_{[0,t_1]}(F, x_0)$ such that $x_2(t_1) = y(t_1)$. We define

$$x_1(t) = \begin{cases} x_2(t) & 0 \leq t \leq t_1, \\ y(t) & t_1 \leq t \leq T. \end{cases}$$

Then $x_1(t) \in S_{[0,T]}(F, x_0)$, and $x_1(T) = y$. It is in conflict with the assumption $y \notin R_{[0,T]}(F, x_0)$. □

Theorem 2.4.2 illustrates that if $x(T)$ is on the boundary of $R_{[0,T]}(F, x_0)$, then the trajectory $x(t) \in S_{[0,T]}(F, x_0)$ has the property that for every $t_1 \in [0, T]$, $x(t_1)$ is at the boundary of $R_{[0,t_1]}(F, x_0)$. The readers are suggested to recall Example 3.2.1, then you can have a deeper understanding.

$S_{[0,T]}(F, x_0)$ is the solution set of Inc. (2.4.1), hence $S_{[0,T]}(F, x_0)$ is a set in normed space $AC([0, T], \mathbb{R}^n)$. We can define a set-valued mapping from $\mathbb{R}^n$ to $AC([0, T], \mathbb{R}^n)$ by $x \mapsto S_{[0,T]}(F, x)$. This is a mapping form the initial condition to the solution set. The following theorem is given for the continuity of the mapping.

Theorem 2.4.3 The set-valued mapping $x \mapsto S_{[0,T]}(F, x)$ is Lipschitzian. Moreover, if $x_1(t) \in S_{[0,T]}(F, x_0)$, then there is a continuous selection $\phi(x)$ of $S_{[0,T]}(F, x)$ such that $\phi(x_0) = x_1(t)$.

Proof Let $x_1, x_2 \in \mathbb{R}^n$. By Theorem 2.3.4, $S_{[0,T]}(F, x_1) \neq \varnothing$. $x_1(t)$ is an arbitrary solution in $S_{[0,T]}(F, x_1)$. Let $M = \{x_1(t)\}$ and $r_0(x) = x - x_1 + x_2$. Then, by Theorem 2.3.3, there is a mapping $r : M \to S_{[0,T]}(F, x_2)$. By using the notations defined in Theorem 2.3.3, we have $\rho(t) = 0$ and $\delta = \|x_2 - x_1\|_{\mathbb{R}^n}$. Let $r(x_1(t)) = x_2(t)$. Then by Inequality (2.3.14), we have

$$\left\|\dot{x}_1(t) - \dot{x}_2(t)\right\|_{\mathbb{R}^n} \leq \|x_1 - x_2\|_{\mathbb{R}^n} l e^{lt}$$

Using the norm defined in AC, we have

$$\|x_1(t) - x_2(t)\|_{AC} = \|x_1 - x_2\|_{\mathbb{R}^n} + \int_0^T \left\|\dot{x}_1(t) - \dot{x}_2(t)\right\| dt \leq e^{lT} \|x_1 - x_2\|_{\mathbb{R}^n}$$

The e^{lT} is independent of the selection of $x_1(t)$ and $x_2(t)$, hence $S_{[0,T]}(F, x)$ is Lipschitzian globally, i.e.,

$$S_{[0,T]}(F, x_1) \subset S_{[0,T]}(F, x_2) + e^{lT} \|x_1 - x_2\|_{\mathbb{R}^n} B.$$

For the second conclusion, let $M = \{x_1(t) + x - x_0\}$[10] and $r_0(x) = x$, then there is a continuous mapping $r : M \to S_{[0,T]}(F, x)$. Let $\phi(x) = r(x_1(t) + x - x_0)$. Then the $\phi(x)$ is continuous, $\phi(x) \in S_{[0,T]}(F, x)$ and $\phi(x_0) = r(x_1(t)) = x_1(t)$. □

2.4.2 *Qualitative Analysis for Upper Semi-continuous Differential Inclusions*

This subsection turns to upper semi-continuous differential inclusions and deals with the same problems as the last subsection. Because failure of Lipschhitz condition, we cannot apply Theorem 2.3.3, these conclusions are established by a different way. Readers are suggested to compare these conclusions as well as the proof procedures.

Consider the Cauchy problem of the following differential inclusion

$$\dot{x}(t) \in F(x(t)), \quad x(0) = x_0 \tag{2.4.2}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is an upper semi-continuous set-valued mapping and with closed and convex value. We further assume that the F is bounded, i.e., there is a $b \in \mathbb{R}\,(> 0)$ such that for every $x \in \mathbb{R}^n$, $F(x) \subset bB$. Comparing with Inc. (2.4.1), the set-valued mapping F in Inc. (2.4.2) is required to be bounded and with convex value.

Two lemmas are given firstly.

Lemma 2.4.1 Suppose $A \subset \mathbb{R}^n$ is a convex and compact set, $f(t) \in L_1([0,T], \mathbb{R}^n)$. If for almost all $t \in [0,T]$, $f(t) \in A$, then $\frac{1}{T}\int_0^T f(t)dt \in A$. □

The lemma can be treated as an extension of mean value theorem of Riemann integration. It can be proved by the definition of Lebesgue integration, and we omit it here. It is only pointed out that the condition that A is a convex and compact set is necessary.

Lemma 2.4.2 Suppose that $A \subset \mathbb{R}^n$ is a convex and compact set, $x_k(t) \in AC([0,T], \mathbb{R}^n)$, $k = 1, 2, \ldots$, and for almost all $t \in [0,T]$, $\dot{x}_k(t) \in A$. If $\lim_{k\to\infty} x_k(t) = x(t)$ for every $t \in [0,T]$, then $x(t) \in AC([0,T], \mathbb{R}^n)$ and $\dot{x}(t) \in A$ for almost all $t \in [0,T]$.

[10] This M is not a solution set.

Proof A is compact, so is bounded. Because $\dot{x}_k(t) \in A$, there exists a $l \in \mathbb{R}^+$ such that $\left\|\dot{x}_k(t)\right\| \leq l$ for almost all $t \in [0, T]$ and all $k = 1, 2, \ldots$. Consequently,

$$\|x_k(t_2) - x_k(t_1)\| \leq \int_{t_1}^{t_2} \left\|\dot{x}_k(\tau)\right\| d\tau \leq l\,|t_2 - t_1|,$$

for all $t_1,\ t_2 \in [0, T]$. It implies the sequence $\{x_k(t), t \in [0, T], k = 1, 2, \ldots\}$ is equicontinuous. Now let $k \to \infty$, we have that $x(t)$ is a Lipschitzian mapping, and $x(t) \in AC([0, T], \mathbb{R}^n)$.

Moreover, by Lemma 2.4.1,

$$\frac{x_k(t+h) - x_k(t)}{h} = \frac{1}{h}\int_{t}^{t+h} \dot{x}_k(\tau)\, d\tau \in A,$$

let $k \to \infty$, we obtain $\frac{x(t+h)-x(t)}{h} \in A$ for $t \in [0, T-h]$ because of the compactness of A. By the compactness again, $\dot{x}(t) = \lim\limits_{h\to 0} \frac{x(t+h)-x(t)}{h} \in A$ for all $t \in [0, T]$. □

The proof of Lemma 2.4.2 illustrates the compactness plays a critical role in the convergence. It also illustrates that if a sequence of absolutely continuous functions whose values are on a compact set is convergent, then the limitation is absolutely continuous.

We now give the main conclusion for upper semi-continuous differential inclusions which holds a similar position to that of Theorem 2.3.3.

Theorem 2.4.4 Let $F, F_k : \mathbb{R}^n \to \mathbb{R}^n$ are all upper semi-continuous and with convex and closed values. Moreover, they satisfy the following conditions:

There is a $b \in \mathbb{R}\ (> 0)$ such that

(1) $F(x) \subset \cdots \subset F_{k+1}(x) \subset F_k(x) \subset \cdots \subset F_0(x) \subset bB$ for every $x \in \mathbb{R}^n$;
(2) For every $\varepsilon > 0$ and $x \in \mathbb{R}^n$, there is a $k_0 = k_0(x, \varepsilon) \in \mathbb{N}$, when $k > k_0$, $F_k(x) \subset F(x) + \varepsilon B$;
(3) $x_k(t) \in S_{[0,T]}(F_k)$ and $x_k(t)$ converges to $x(t)$ uniformly.

Then $x(t) \in S_{[0,T]}(F)$.

Proof Because $\dot{x}_k(t) \in F_k(x_k(t)) \subset bB$ and $x_k(t)$ converges uniformly to $x(t)$, we conclude $x(t) \in AC$ by Lemma 2.4.2. Suppose $x(t)$ is derivable at $t_0 \in [0, T]$ and denote $x_0 = x(t_0)$. Using the second condition of the Theorem 2.4.4, for every $\varepsilon > 0$, we have a $\overline{k} > k_0(x_0, \varepsilon)$ such that $F_{\overline{k}}(x_0) \subset F(x_0) + \varepsilon B$. $F_{\overline{k}}(x_0)$ is an upper semi-continuous set-valued mapping, hence, there is an $\eta > 0$, when $x \in B(x_0, \eta)$,

$$F_{\overline{k}}(x) \subset F_{\overline{k}}(x_0) + \varepsilon B \subset F(x_0) + 2\varepsilon B. \tag{2.4.3}$$

On the other hand, $x_k(t)$ converges uniformly to $x(t)$, there exist $k_1 \geq \overline{k}$ and γ when $k > k_1$ and $|t - t_0| < \gamma$, we have

$$\|x_k(t) - x(t)\| < \frac{\eta}{2}, \;\; \|x(t) - x\left(t_0\right)\| < \frac{\eta}{2}.$$

By Relation (2.4.3),

$$\dot{x}_k(t) \in F_k\left(x_k(t)\right) \subset F_{k_0}\left(x_k(t)\right) \subset F_{k_0}\left(x\left(t_0\right)\right) + \varepsilon B \subset F\left(x_0\right) + 2\varepsilon B.$$

Using Lemma 2.4.2 again, $\dot{x}(t) \in F\left(x_0\right) + 2\varepsilon\overline{B}$ for all $|t - t_0| < \gamma$. Let $t = t_0$. Then $\dot{x}\left(t_0\right) \in F\left(x_0\right) + 2\varepsilon B$. The ε can be selected arbitrary and $F(x_0)$ is closed, consequently, $\dot{x}\left(t_0\right) \in F\left(x_0\right)$. □

Remark If the condition that $F_k : \mathbb{R}^n \to \mathbb{R}^n$ are upper semi-continuous is replaced by $F_k : \mathbb{R}^n \to \mathbb{R}^n$ are Lipschitzian mappings, then it can be proved that Relation (2.4.3) still holds, hence, the theorem is still valid.

By Theorem 2.2.4 and the above remark, Theorem 2.4.4 supports the following conclusion, the detailed proof is left to readers.

Theorem 2.4.5 Consider Inc. (2.4.2) i.e., $F : \mathbb{R}^n \to \mathbb{R}^n$ is an bounded, upper semi-continuous set-valued mapping and with closed and convex value. Then its Cauchy problem is solvable. □

Corollary 2.4.1 Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is upper semi-continuous (or locally Lipschitzian) and with closed and convex value, and F is bounded, i.e., there is a b such that for every $x \in \mathbb{R}^n F(x) \subset bB$. Then $S_{[0,T]}(F, x_0)$ is a compact set in AC.

Proof Because $F(x) \subset bB$, every sequence of $S_{[0,T]}(F, x_0)$ is equicontinuous. By Arzela-Ascili theorem (Sect. 1.1), the sequence holds a convergent subsequence. By Theorem 2.4.4, replacing all F_k by F, we conclude the limitation of the subsequence belongs also to $S_{[0,T]}(F, x_0)$. □

Corollary 2.4.1 can be extended to the set $S_{[0,T]}(F, x_0)$ provided that $C \subset \mathbb{R}^n$ is a compact set in $\mathbb{R}^n$ (Problem 1).

We now apply Theorem 2.4.4 to complete the corresponding qualitative analysis for Inc. (2.4.2).

A set A is said to be unconnected if there are two closed sets A_1 and A_2 which satisfy $A_1 \cap A_2 = \varnothing$, and $A_1 \cap A \neq \varnothing$ and $A_2 \cap A \neq \varnothing$ such that $A_1 \cup A_2 \supset A$. Otherwise, A is connected.

Theorem 2.4.6 Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is bounded, upper semi-continuous and with closed and convex value. Then for every $x_0 \in \mathbb{R}^n$, $S_{[0,T]}(F, x_0)$ is a connected set of the space $C\left([0, T], \mathbb{R}^n\right)$.[11]

[11] Note that the connectedness is considered in normed space C, and is not in $\mathbb{R} \times \mathbb{R}^n$. In $\mathbb{R} \times \mathbb{R}^n$, all solutions in $S_{[0,T]}(F, x_0)$ start at $(0, x_0)$, hence it has to be connected.

Proof The conclusion is proved by using contradiction. If there are two nonempty closed sets S_1 and S_2 which satisfy $S_1 \cap S_2 = \varnothing$, $S_1 \cup S_2 \supset S_{[0,T]}(F, x_0)$, and $S_i \cap S_{[0,T]}(F, x_0) \neq \varnothing$, for $i = 1, 2$. Denote $x_i(t) \in S_{[0,T]}(F, x_0) \cap S_i$, $i = 1, 2$. By Corollary 2.4.1, $S_{[0,T]}(F, x_0)$ is compact, hence S_1 and S_2 are all bounded. We have

$$\inf\left\{\|x_1(t)-x_2(t)\|\,; x_1(t) \in S_{[0,T]}(F, x_0) \bigcap S_1, x_2(t) \in S_{[0,T]}(F, x_0) \bigcap S_2\right\} = 2\delta > 0$$

A function is defined as follows:

$$\varphi : C([0, T], \mathbb{R}^n) \to \mathbb{R}, \ \ \varphi(x(t)) = d(x(t), S_1) - \delta$$

When $x(t) \in S_1$, $\varphi(x(t)) = -\delta$; and $x(t) \in S_2$, $\varphi(x(t)) \geq \delta$. We now fix a $x_1(t) \in S_1 \cap S_{[0,T]}(F, x_0)$ and a $x_2(t) \in S_2 \cap S_{[0,T]}(F, x_0)$.

By Theorem 2.2.4, there are $F_k : \mathbb{R}^n \to \mathbb{R}^n$, $k = 1, 2, \ldots$, which are locally Lipschitzian set-valued mappings and are with such that convex and closed values. Moreover, they satisfy

(1) For every $x \in \mathbb{R}^n$, $F(x) \subset \cdots \subset F_{k+1}(x) \subset F_k(x) \subset \cdots \subset F_0(x) \subset bB$;
(2) For every $\varepsilon > 0$, there exists a $k_0(x, \varepsilon)$, when $k > k(x_0, \varepsilon)$, we have $F_k(x) \subset F(x) + \varepsilon B$.

When $k > k_0(x_0, \varepsilon)$, we consider the Cauchy problem that $\dot{x} \in F_k(x)$, $x(0) = x_0$, its solution set $S_{[0,T]}(F_k, x_0) \neq \varnothing$ by Theorem 2.3.4. Using Theorem 2.4.1, there is a continuous mapping $\phi_k : [0, 1] \to S_{[0,T]}(F_k, x_0)$ such that $\phi_k(0) = x_1(t)$ and $\phi_k(1) = x_2(t)$. Consider the compound function: $\varphi \circ \phi_k : [0, 1] \to \mathbb{R}$, we have $\varphi \circ \phi_k(0) < 0$ and $\varphi \circ \phi_k(1) > 0$. By the property of continuous function, there is a $\lambda_k \in (0, 1)$ such that $\varphi \circ \phi_k(\lambda_k) = 0$. Denote $x_k(t) = \phi_k(\lambda_k)$, then $x_k(t) \in S_{[0,T]}(F_k, x_0)$. Because sequence $\left\{\dot{x}_k(t)\right\} \subset bB$, the sequence $\{x_k(t)\}$ has a convergent subsequence. Without loss of generality, we assume that $x_k(t) \to x(t)$ by the norm of the space of continuous functions. Theorem 2.4.4 then asserts that $x(t) \in S_{[0,T]}(F, x_0)$. From $\varphi(x_k(t)) = 0$ and the continuity of φ, we have $\varphi(x(t)) = 0$. Therefore, $x(t) \notin S_1 \cup S_2$, it is contradictive to the hypothesis of $S_1 \cup S_2 \supset S_{[0,T]}(F, x_0)$. Hence $S_{[0,T]}(F, x_0)$ is connected. □

Theorem 2.4.7 Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is bounded, upper semi-continuous and with closed and convex value. Let $x^* \in \text{bd } R_{[0,T]}(F, x_0)$. Then there is an $x_1(t) \in S_{[0,T]}(F, x_0)$ which holds the properties: (1) $x_1(T) = x^*$, (2) for every $t \in [0, T]$, $x_1(t) \in \text{bd } R_{[0,t]}(F, x_0)$.

Proof The proof consists of two parts. At first, we construct a solution $x(t) \in S_{[0,T]}(F, x_0)$, then we prove the solution satisfies the requirements of the theorem.

(1) By Theorem 2.2.4, there is a sequence $\{F_k(x)\}$ such that for every $x \in \mathbb{R}^n$

$$F(x) \subset \cdots \subset F_{k+1}(x) \subset F_k(x) \subset \cdots \subset F_0(x) \subset bB;$$

And for every $\varepsilon > 0$, there exists a $k_0(x, \varepsilon)$, when $k > k_0(x, \varepsilon)$, $F_k(x) \subset F(x) + \varepsilon B$.

By Corollary 2.4.1, $S_{[0,T]}(F_k, x_0)$ is compact. Hence, there exists a $x_k \in R_{[0,T]}(F_k, x_0)$ such that $d\left(x^*, R_{[0,T]}(F_k, x_0)\right) = \|x_k - x^*\|$. It is obvious that $x_k \in \mathrm{bd}R_{[0,T]}(F_k, x_0)$. By Theorem 2.4.2, there exists an $x_k(t) \in S_{[0,T]}(F_k, x_0)$ such that $x_k(T) = x_k$, moreover, for every $t \in [0,\ T]$, $x_k(t) \in \mathrm{bd}R_{[0,t]}(F_k, x_0)$. By the same reasoning as used in Theorem 2.4.6, we can assume that $x_k(t) \to x(t)$ by the norm of the space of continuous functions, and $x(t) \in S_{[0,T]}(F, x_0)$.

(2) We now prove that the $x(t)$ satisfies the requirements of theorem. At first, we assert that $x(T) = x^*$. The conclusion is verified by contradiction. If $\delta = \|x(T) - x^*\| > 0$, then there is a k_0 such that for every $k \geq k_0$, $\|x_k - x^*\| \geq \frac{\delta}{2}$ where $x_k = x_k(T) \in \mathrm{bd}R_{[0,T]}(F_k,\ x_0)$. By the condition of theorem $x^* \in R_{[0,T]}(F, x_0) \subset R_{[0,T]}(F_k, x_0)$, hence $x^* \notin \mathrm{bd}R_{[0,T]}(F_k, x_0)$ (recall the selection of x_k). It follows x^* is an inner point of $R_{[0,T]}(F_k, x_0)$, consequently, there is an $\varepsilon > 0$ such that $x^* + \varepsilon B \subset R_{[0,T]}(F_k, x_0)$ for all $\geq k_0$. If $y \in B(x^*, \varepsilon)$, then we have a $y_k(t) \in S_{[0,T]}(F_k, x_0)$ and $y = y_k(T)$. By the proof procedure applied above, we can assume $\{y_k(t)\}$ converges uniformly to $y(t) \in S_{[0,T]}(F, x_0)$. That $y \in R_{[0,T]}(F, x_0)$ implies that x^* is not in on the boundary of $R_{[0,T]}(F_k, x_0)$. It contradict condition of the theorem.

At last, we show that $x(t)$ has the property that $x(t) \in \mathrm{bd}R_{[0,t]}(F, x_0)$ for every $t \in [0, T]$. Otherwise, this conclusion is not true, i.e., there is a $t_1 \in [0, T]$, such that $x(t_1) + \varepsilon B \subset R_{[0,t_1]}(F, x_0)$. Then for k which is large enough, we have

$$x_k(t_1) + \frac{\varepsilon}{2}B \subset x(t_1) + \varepsilon B \subset R_{[0,t_1]}(F, x_0) \subset R_{[0,t_1]}(F_k, x_0)$$

It conflicts to Theorem 2.4.2 since F_k is a Lipschitzian set-valued mapping. □

Remark Theorem 2.4.6 and Theorem 2.4.7, in the theory of differential equations in manifold, are known as Kneser theorem and Hukuhara theorem, respectively.

Readers should distinguish Theorem 2.4.2 from Theorem 2.4.7. Theorem 2.4.2 asserts that if $x_1(T) = x^*$, then for every $t \in [0, T]$, $x_1(t) \in \mathrm{bd}\ R_{[0,t]}(F, x_0)$. But, Theorem 2.4.6 states that there may be several solutions in $S_{[0,T]}(F, x_0)$ whose terminals are all x^*. Among these solutions there is certainly a $x_1(t)$ such that $x_1(T) = x^*$ and $x_1(t) \in \mathrm{bd}\ R_{[0,t]}(F, x_0)$ for $t \in [0, T]$.

Example 2.4.1 A set-valued mapping $F(x)$ is defined as follows:

$$F(x) = \begin{cases} \overline{B}_2 & x_2 < 0, \\ 6\overline{B}_2 & x_2 = 0, \\ 0 & x_2 > 0. \end{cases} \tag{2.4.4}$$

where $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^2$[12] and $\overline{B}_2$ is the closed unit ball in $\mathbb{R}^2$. Let the initial condition be $x(0) = \begin{bmatrix} 0 & -4 \end{bmatrix}^{\mathrm{T}}$ and terminal be $x(T) = \begin{bmatrix} 3 & 0 \end{bmatrix}^{\mathrm{T}}$, and time interval is $[0, 5]$ (i.e., $T = 5$).

[12] The superscript T means transposition.

It is easy to check that the $F(x)$ is an upper semi-continuous and with convex and closed value. We firstly illustrate that $F(x)$ is not a Lipschitzian mapping. If $x = \begin{bmatrix} 1 & 0 \end{bmatrix}^{\mathrm{T}}$, then $F(x) = 6\overline{B}_2$; but if $y = \begin{bmatrix} 1 & \varepsilon \end{bmatrix}^{\mathrm{T}}$ for every $\varepsilon > 0$, then $F(y) \equiv 0$. If we require that

$$6\overline{B}_2 = F(x) \subset F(y) + l\,\|x-y\|\,B_2 = l\varepsilon B_2$$

then $l\varepsilon > 6$. When $\varepsilon \to 0$, then $l \to \infty$, hence $F(x)$ is not a Lipschitzian mapping.

Now

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 0 \\ -4 + \dfrac{8t}{9} \end{bmatrix}, t \in \left[0, \frac{9}{2}\right], \quad \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 6\left(t - \dfrac{9}{2}\right) \\ 0 \end{bmatrix}, \quad t \in \left[\frac{9}{2}, 5\right]$$

is a solution of differential inclusion $\dot{x} \in F(x)$ where $F(x)$ is given in Eq. (2.4.4). The solution satisfies that

$$\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ -4 \end{bmatrix}, \quad \begin{bmatrix} x_1(5) \\ x_2(5) \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$$

But when $t \in \left[0, \dfrac{9}{2}\right]$, $\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 \\ \dfrac{8}{9} \end{bmatrix}$ which does not lay on the boundary of B_2.

There is another solution

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \dfrac{3}{5}t \\ -4 + \dfrac{4}{5}t \end{bmatrix}, \quad t \in [0,5],$$

which also satisfies the boundary conditions. The solution is always on the boundary. □

Theorem 2.4.8 Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is a bounded, upper semi-continuous set-valued mapping, and with closed and convex value. If we treat $S_{[0,T]}(F, x_0)$ as a set-valued mapping which maps $\mathbb{R}^n$ to $AC\left([0,T], \mathbb{R}^n\right)$. Then the mapping is upper semi-continuous.

Proof We prove a more general conclusion that if $C \subset \mathbb{R}^n$ is a compact set, then $S_{[0,T]}(F, C)$ is a compact set of $AC\left([0,T], \mathbb{R}^n\right)$.

Suppose $\{x_k(x)\} \subset S_{[0,T]}(F, C)$ is a sequence. Because $F(C) \subset bB$ for some constant b, $\left\|\dot{x}_k(x)\right\| \leq b$ is true for all k and almost all x. The fact implies that $\{x_k(x)\}$ is equicontinuous. There exists a convergent subsequence by Arzela-Ascili theorem. Without loss of generality, we assume $\{x_k(x)\}$ is uniformly convergence, i.e., there is a function $x_0(t) \in AC\left([0,T], \mathbb{R}^n\right)$ such that $x_{k_j}(t) \to x_0(t)$, we now prove that $x_0(t) \in S_{[0,T]}(F, C)$.

Suppose $\dot{x}_k(t) \in F_k(x_k(t))$ where $F_k \equiv F$, $k = 1, 2, \ldots$. Then by Theorem 2.4.4, $x_0(t) \in S_{[0,T]}(F, C)$. $S_{[0,T]}(F, C)$ is compact by Corollary 2.4.1. Moreover, by the Problem 9 in Sect. 2.1, we conclude that the mapping $x_0 \mapsto S_{[0,T]}(F, x_0)$ is upper semi-continuous. □

Note that in Theorem 2.4.8, the condition that $F(x)$ is convex is necessary.

2.4.3 Convexification of Differential Inclusions and Relaxed Theorem

At last, we deal with the convexification of differential inclusions in this section. It have been found that convexity is a critical condition for the existence of the solutions for differential inclusions. One may ask such a question *if the set$F(x)$is not convex, what can we do?* The best answer currently is to make $F(x)$ convex.

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a set-valued mapping, and $\overline{\text{co}}\, F : \mathbb{R}^n \to \mathbb{R}^n$ be the closure of convex hull of F. Then the differential inclusion

$$\dot{x} \in \overline{\text{co}}\, F(x) \tag{2.4.5}$$

is the convexification of the original differential inclusion

$$\dot{x} \in F(x). \tag{2.4.6}$$

For the sake of convenience, the notation $\overline{\text{co}}\, F$ is simplified as $\overline{F}$, and the solution of Inc. (2.4.5) as $\overline{x}(t)$. If the initial conditions are equal, i.e., $\overline{x}(0) = x(0) = x_0$, then the solution of Inc. (2.4.6) is certainly the solution of Inc. (2.4.5). In the study of convexification, the most important problem is to present the conditions under which the solutions of Inc. (2.4.5) is dense at the solutions of Inc. (2.4.6).

The main target of this subsection is to prove Theorem 2.4.9. The proof is quite similar to that of the existence of solution for differential equations by using Eular polygonal lines. But the construction is much complicated.

Theorem 2.4.9 Let $x_0 \in \mathbb{R}^n$, $F : B(x_0, b) \to \mathbb{R}^n$ be a Lipschitzian set-valued mapping with Lipschitzian constant l. F is bounded and with closed value. Then there is a $T > 0$ such that the solutions of Inc. (2.4.5) is dense at the solutions of Inc. (2.4.6) on the time interval $I = [-T, T]$.

Proof The theorem is equivalent to prove that for every $\varepsilon > 0$ and every $\overline{x}(t, x_0)$ which is a solution of Inc. (2.4.5), then there is a solution $x(t, x_0)$ of Inc. (2.4.6) such that $\|x(t, x_0) - \overline{x}(t, x_0)\| < \varepsilon$ for every $t \in I$. Both $\overline{x}(t, x_0)$ and $x(t, x_0)$ are in $B(x_0, b)$ for each $t \in [-T, \ T]$.

Because F is bounded, there is a constant M such that $F(B(x_0, b)) \subset M\overline{B}$, where $\overline{B}$ is the closed unit ball of $\mathbb{R}^n$. It implies $\overline{F}(B(x_0, b)) \subset M\overline{B}$. F and $\overline{F}$ are all Lipschitzian mappings, hence, Inc. (2.4.5) and Inc. (2.4.6) exist solutions.

If for an $\varepsilon > 0$ given arbitrarily, we can find a $z(t) \in AC$, $z(0) = x_0$ (for convenience, it is denoted by $z(t, x_0)$) such that for $t \in I$,

$$\|\overline{x}(t, x_0) - z(t, x_0)\| < \frac{\varepsilon}{2} \tag{2.4.7}$$

and

$$d\left(\dot{z}(t, x_0), F(z(t, x_0))\right) \leq \frac{l\varepsilon}{2\left(e^{lT} - 1\right)} \tag{2.4.8}$$

then by the notations used in Theorem 2.3.3, we have $\delta = 0$ and $\rho(t) = \frac{l\varepsilon}{2\left(e^{lT}-1\right)}$. Theorem 2.3.3 asserts that there is a solution $x \in S_{[0,T]}(F, x_0)$ such that

$$\|x(t, x_0) - z(t, x_0)\| \leq \int\limits_0^t e^{l(t-s)} \frac{l\varepsilon}{2\left(e^{lT} - 1\right)} ds \leq \frac{\varepsilon}{2}$$

It follows $\|x(t, x_0) - \overline{x}(t, x_0)\| < \varepsilon$. The theorem is then verified. Hence, the key point is to construct a $z(t, x_0)$ which satisfies Inequalities (2.4.7) and (2.4.8).

(1) We now construct $z(t, x_0)$ by the following two steps.

(i) The time interval I is divided into $2q$ equal sections, i.e., the length of each section is $\frac{T}{q}$. $I_i = [t_i, t_{i+1}]$ is denoted for the ith sub-interval and $\Phi_i = \overline{F}(\overline{x}(t_i, x_0)) \subset \mathbb{R}^n$ is the image of $\overline{F}(\overline{x}(t_i, x_0))$ which is closed and convex. It is direct to show $\overline{F}(\overline{x}(t, x_0))$ is a Lipschitzian mapping and its Lipschitzian constant can be lM (to see Problem 8 of this section). It follows that for all $t \in I_i$, $\overline{F}(\overline{x}(t, x_0)) \subset \Phi_i + \frac{lMT}{q}\overline{B}$. We denote $S = \Phi_i + \frac{lMT}{q}\overline{B}$, S is bounded, closed and convex. The S can be contained by a union of finite lattice cubes[13], i.e., $S \subset \bigcup\limits_{j=1}^{J} S_j$ where S_j's are lattice cubes with equal side length. We can require that the diameter of the lattice cubes is less than a given positive number ζ. Figure 2.14 is given to explain the set S and these lattice cubes. The line in the lowest position is the segment I, where $\overline{x}(t, x_0)$ is defined. $\Phi_{\overline{x}}$ is the image of the set-valued mapping $\overline{F}(\overline{x}(t, x_0))$, $\Phi_{\overline{x}}$ and $\Phi_i = \overline{F}(\overline{x}(t_i, x_0))$ are surrounded by full line (both $\Phi_{\overline{x}}$ and Φ_i are slices in Fig. 2.14). S is a stereoscopic cube and is surrounded by dash lines. S contains $\Phi_{i.}$[14] These straight lines separate S into lattice cubes. S_j is one lattice cube, whose diameter is its diagonal line (no drawing). We can increase the density of these lattice vertexes so that the diameter is less than any pointed real number ζ. $\dot{\overline{x}}(t, x_0) \in \overline{F}(\overline{x}(t, x_0))$, hence $\dot{\overline{x}}(t, x_0)$ passes through $\Phi_{\overline{x}}$ in only one S_j.

[13] A lattice cube is a regular cube whose vertexes are rational numbers $\frac{q}{m}$, $q \in \{0, \pm 1, \pm 2, \ldots\}$ and length of side is $\frac{1}{m}$.

[14] S looks to like a sandwich and Φ_i likes a piece of ham in the sandwich.

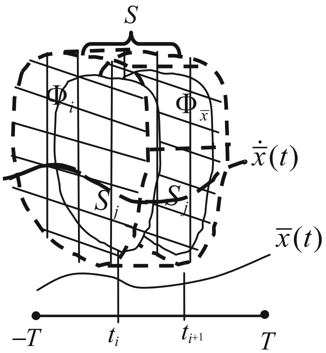


Fig. 2.14 Diagrammatic drawing for the proof of Theorem 2.4.9

A set $E_j = \left\{t; t \in I_i,\ \dot{\overline{x}}(t, x_0) \in S_j\right\}$ and a step function $\xi(t) = \sum_j \xi_j \chi_j(t)$ are defined where $\xi_j \in S_j \cap S$ is a determined vector and $\chi_j(t)$ is the characteristic function of E_j, i.e.,

$$\chi_j(t) = \begin{cases} 1, & t \in E_j, \\ 0, & t \notin E_j. \end{cases}$$

It is obvious that $d\left(\xi_j, \Phi_i\right) \le \frac{lMT}{q}$, $\left\|\dot{\overline{x}}(t, x_0) - \xi(t)\right\| \le \zeta, t \in I_i$

(ii) The function $\xi(t) = \sum_j \xi_j \chi_j(t)$ is a simple function of S_j, we now construct a simple function for $F(\overline{x}(t_i, x_0))$.

Recall $\xi_j \in S_j \cap S \subset S = \Phi_i + \frac{lMT}{q}\overline{B}$. $\Phi_i = \overline{F}(\overline{x}(t_i, x_0))$ is the closed convex hull of $F(\overline{x}(t_i, x_0))$, hence, there exist k vectors $z_{jh} \in F(\overline{x}(t_i, x_0))h = 1, 2, \ldots, k,$ and $k \le n + 1$, such that

$$\left\|\xi_j - \sum_{h=1}^{k} \alpha_{jh} z_{jh}\right\| \le \frac{2lMT}{q}, \ \sum_{h=1}^{k} \alpha_{jh} = 1, \alpha_{jh} > 0. \tag{2.4.9}$$

We now define $\psi(t) = \int_{t_i}^{t} \chi_j(s)ds$, the $\psi(t)$ is monotonous and $\psi(t_{i+1}) = m\left(E_j\right)$, where $m(E_j)$ is the Lebesgue measure of the set E_j. Real numbers $t_i < \tau_1 < \tau_2 < \cdots < \tau_k$ can be obtained

$$\tau_h = \sup\left\{t; \psi(t) \le m\left(E_j\right) \sum_{p=1}^{h} \alpha_{jp}\right\}$$

These τ_h's, $h = 1, 2, \ldots, k$ define k intervals $[\tau_0, \tau_1], (\tau_1, \tau_2], \ldots (\tau_{k-1}, \tau_k]$ with $\tau_0 = t_i$ and $\tau_k = t_{i+1}$. Set E_{jh} is then defined as $E_{jh} = E_j \cap (\tau_{h-1}, \tau_h], \; h = 1, 2, \ldots, k$, then these E_{jh}'s hold the properties that $E_{jh_1} \cap E_{jh_2} = \varnothing, \; h_1 \neq h_2$ and $\overset{k}{\underset{h=1}{\cup}} E_{jh} = E_j$. Let $\chi_{jh}(t)$ be the characteristic function of E_{jh}. Then on the set E_j a simple function $\zeta_j(t)$ is defined as follows

$$\zeta_j(t) = \sum_h z_{jh} \chi_{jh}(t)$$

where $z_{jh} \in F(\overline{x}(t_i, x_0))$ defined by Inequality (2.4.9). Define now $\zeta(t) = \sum_j \zeta_j(t)$ and finally

$$z(t, x_0) = x_0 + \int_0^t \zeta(s) ds \tag{2.4.10}$$

(2) We prove that $z(t, x_0)$ defined by Eq. (2.4.10) can meet the requirements of Inequalities (2.4.7) and (2.4.8).

At first, by the construction of $z(t, x_0)$, we have $\dot{z}(t, x_0) = \zeta(t) \in F(B(x_0, b)) \subset M\overline{B}$, and $z(t, x_0)$ is a Lipschitzian function and its Lipschitzian constant can be M, the same as $\overline{x}(t, x_0)$. For every $t \in [-T, T]$, there is an subinterval I_i, such that $t \in I_i$, we have

$$\begin{aligned}\|\overline{x}(t, x_0) - z(t, x_0)\| \le \|\overline{x}(t, x_0) - \overline{x}(t_i, x_0)\| + \|\overline{x}(t_i, x_0) - z(t_i, x_0)\| + \|z(t_i, x_0)\\ -z(t, x_0)\|.\end{aligned}$$

By the Lipschitzian condition, we have

$$\|\overline{x}(t, x_0) - \overline{x}(t_i, , x_0)\| \le \frac{MT}{q}, \quad \|z(t_i, x_0) - z(t, x_0)\| \le \frac{MT}{q}.$$

Recall the construction that $\xi(t) = \sum_j \xi_j \chi_j(t)$ and $\zeta(t) = \sum_j \sum_k z_{jk} \chi_{jk}(t)$, consequently, $\int_{I_i} \xi(s) ds = \sum_j m(E_j) \xi_j$, and

$$\int_{I_i} \zeta(s) ds = \int_{I_i} \sum_j \sum_k z_{jk} \chi_{jk}(s) ds = \sum_j \sum_k z_{jk} \int_{I_i} \chi_{jk}(s) ds = \sum_j \sum_k z_{jk} \alpha_{jk} m(E_j).$$

Therefore,

$$\begin{aligned}\left\| \int_{I_i} \xi(s)ds - \int_{I_i} \rho(s)ds \right\| &\leq \sum_j m\left(E_j\right) \left\| \xi_j - \sum_k \alpha_{jk} z_{jk} \right\| \\ &\leq \sum_j m\left(E_j\right) \frac{2lMT}{q} \\ &\leq m\left(I_i\right) \frac{2lMT}{q} \\ &= \frac{2lMT^2}{q^2}.\end{aligned}$$

At the time t_i, we have

$$\begin{aligned}\left\| z\left(t_i, x_0\right) - x_0 - \int_0^{t_i} \xi(s)ds \right\| &= \left\| \int_0^{t_i} \rho(s)ds - \int_0^{t_i} \xi(s)ds \right\| \\ &\leq \sum_{k=1}^{i} \int_{I_k} \left\| \rho(s) - \xi(s) \right\| ds \\ &\leq \frac{4lMT^2}{q} .\end{aligned}$$

On the other hand,

$$\left\| \overline{x}\left(t_i, x_0\right) - x_0 - \int_0^{t_i} \xi(s)ds \right\| = \left\| \int_0^{t_i} \dot{\overline{x}}\left(s, x_0\right) - \xi(s)ds \right\| \leq \int_0^{t_i} \left\| \dot{\overline{x}}\left(s, x_0\right) - \xi(s) \right\| ds \leq \zeta T.$$

The both inequalities lead to $\left\| \overline{x}\left(t_i, x_0\right) - z\left(t_i, x_0\right) \right\| \leq \zeta T + \frac{4lMT^2}{q}$.

Summing up the above procedure, we have

$$\left\| \overline{x}\left(t, x_0\right) - z\left(t, x_0\right) \right\| \leq \frac{2MT}{q} + \zeta T + \frac{4lMT^2}{q} \tag{2.4.11}$$

for every $t \in [-T, T]$.

Recall that S_j is a lattice cube. If its length of side is $\frac{1}{q}$, then its diameter is $\frac{\sqrt{n}}{q}$. Inequality (2.4.11) leads to

$$\left\| \overline{x}\left(t, x_0\right) - z\left(t, x_0\right) \right\| \leq \frac{\lambda_1}{q} \tag{2.4.12}$$

where $\lambda_1 = \left(2M + \sqrt{n} + 4lMT\right) T$.

Because $\dot{z}(t, x_0) \in F(\overline{x}(t_i, x_0))$ for $t \in I_i$, $d\Big(\dot{z}(t), F(\overline{x}(t, x_0))\Big) \leq \frac{lMT}{q}$. By Inequality (2.4.12), we have $F(x(t)) \subset F(z(t)) + \frac{\lambda_1 l}{q}B$. Using triangle inequality, at last

$$d\Big(\dot{z}(t), F(z(t))\Big) \leq \frac{lMT}{q} + \frac{\lambda_1 l}{q}$$

Therefore, when

$$q = \min\left\{\left\lfloor\frac{2\lambda_1}{\varepsilon}\right\rfloor, \left\lfloor\frac{2\left(e^{lT}-1\right)(MT+\lambda_1)}{\varepsilon l}\right\rfloor\right\} + 1$$

where $\lfloor\alpha\rfloor$ is used to express the largest integer which is less or equal to α. Inequalities (2.4.7) and (2.4.8) are both satisfied. We complete the proof. □

Theorem 2.4.9 is also named by relaxation theorem. We emphasize that the Lipschitzian condition for set-valued mapping F is necessary for the validation of the theorem.

Problems

1. Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is an upper semi-continuous and bounded set-valued mapping. If it is with convex and closed value, then prove $S_{[0,T]}(F, C)$ is a compact set in AC where $C \subset \mathbb{R}^n$ is a compact set. Furthermore, $R_{[0,T]}(F, x_0)$ is also compact.
2. Prove Lemma 2.4.1.
3. Prove Theorem 2.4.5.
4. Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitzian set-valued mapping, and is with closed value. By using Theorem 2.3.3, prove the former part of Theorem 2.4.3, i.e., if we treat $S_{[0,T]}(F, x_0)$ as a set-valued mapping from $\mathbb{R}^n$ to $AC([0,T], \mathbb{R}^n)$, then the mapping is also Lipschitzian.
5. Suppose $F(x) \equiv \{-1, 1\}$. Prove the F is upper semi-continuous and with closed value, but is not with convex value. Consider the differential inclusion $\dot{x} \in F(x), x(0) = 0$. Prove that a saw-toothed function

$$x_k(t) = \begin{cases} t - \frac{i}{2^k}, & \frac{i}{2^k} \leq t < \frac{i+1}{2^k}, \\ \frac{i+2}{2^k} - t, & \frac{i+1}{2^k} \leq t < \frac{i+2}{2^k}, \end{cases} \quad i = 0, 1, 2, \ldots$$

is its solution for every $k \in \mathbb{N}$. The limitation of $x_k(t)$, $x(t) \equiv 0$ when $k \to \infty$, is not a solution. Hence, Theorem 2.4.7 fails.
6. Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is a bounded and upper semi-continuous mapping, and with closed and convex value. $C \subset \mathbb{R}^n$ is a compact set. We construct a mapping $\text{gra}(F) \to \text{gra}S_{[0,T]}(F, C)$. Prove that the mapping is continuous, i.e., for every $\varepsilon > 0$, there exists a $\delta > 0$, if $\text{gra}G \subset \text{gra}F + \delta B$ then

$$\text{gra}S_{[0,T]}(G, C) \subset \text{gra}S_{[0,T]}(F, C) + \varepsilon B.$$

7. $F : \Lambda \times \mathbb{R}^n \to \mathbb{R}^n$, $(\lambda, x) \mapsto F(\lambda, x)$ is a set-valued mapping with parameter λ. Suppose F is an upper semi-continuous mapping, and with convex and closed value for every (λ, x). Then for every $\varepsilon > 0$, there is a $\delta > 0$ such that when $|\lambda - \lambda_0| < \delta$

$$\text{gra}S_{[0,T]}\left(F(\lambda, x(t)), x_0\right) \subset \text{gra}S_{[0,T]}\left(F(\lambda_0, x(t)), x_0\right) + \varepsilon B.$$

8. Suppose $F : \mathbb{R}^2 \to \mathbb{R}^2$ is as

$$F\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} \{-1, 1\} \\ \sqrt{|x_2|} + |x_1| \end{bmatrix}$$

Prove the following conclusions.

1) F is with compact value;
2) F is not a Lipschitzian mapping;
3) Consider the Cauchy problem for $\dot{x}(t) \in F(x(t))$, $x(0) = 0$. Prove that the relaxation theorem (Theorem 2.4.9) is not true for the example.

9. Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitzian mapping with Lipschitzian constant l, and $F(x) \subset MB$ for every $x \in \mathbb{R}^n$. If $\dot{x}(t) \in F(x(t))$ be a solution of the differential inclusion, then $t \mapsto F(x(t))$ is a Lipschitzian set-valued mapping whose Lipschitzian constant can be lM.

2.5 Stability of Differential Inclusions

In both theories of differential equations and control systems, stability is always a very fundamental issue. This section deals with the theory of stability for differential inclusions, and the elemental method is the Lyapunov direct method. To introduce the method, we define Dini derivatives firstly. Then the stabilities for differential inclusion are defined. At last, the theorems for stabilities are presented by using Lyapunov functions.

2.5.1 Dini Derivatives

This subsection deals with the concept of Dini derivatives and their main properties.

Suppose $f : [\alpha, \beta] \to \mathbb{R}$ is a single-valued function and $x_0 \in (\alpha, \beta)$, then

$$D^+ f(x_0) = \lim_{\Delta x \downarrow 0} \sup \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

$$D_+ f(x_0) = \lim_{\Delta x \downarrow 0} \inf \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

$$D^{-}f(x_0) = \lim_{\Delta x \uparrow 0} \sup \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

$$D_{-}f(x_0) = \lim_{\Delta x \uparrow 0} \inf \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

are defined to be the right upper Dini derivative, the right lower Dini derivative, the left upper Dini derivative and the left lower Dini derivative, of $f(x)$ at x_0, respectively.

By the decrease of Δx, $\sup \frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}$ decreases monotonously, the right upper Dini derivative is existed no matter what is the function $f(x)$. Of course, $D^{+}f(x_0)$ may be equal to infinite. By the same reason, we can conclude that $D_{+}f(x_0)$, $D^{-}f(x_0)$ and $D_{-}f(x_0)$ all exist, of course they may be equal to infinite. However, if $f(x)$ is a Lipschitzian function, then we can proof the four derivatives all have finite values.

Lemma 2.5.1 If $f : (\alpha, \beta) \to \mathbb{R}$ is a continuous function, then $f(x)$ is a monotonously increasing if and only if one of the following conditions holds at every $x_0 \in (\alpha, \beta)$:

(1) $D^{+}f(x_0) \geq 0$;
(2) $D_{+}f(x_0) \geq 0$;
(3) $D^{-}f(x_0) \geq 0$;
(4) $D_{-}f(x_0) \geq 0$.

Proof We only prove the first condition, the others are similar and left to readers.

If $f(x)$ is monotonously increasing on (α, β), then $f(x_0 + \Delta x) - f(x_0) \geq 0$ for every $x_0 \in (\alpha, \beta)$. Hence, when $\Delta x > 0$, $\sup \frac{f(x_0+\Delta x)-f(x_0)}{\Delta x} \geq 0$, it follows $D^{+}f(x_0) \geq 0$.

We now prove the sufficiency. At first, we consider the case that $D^{+}f(x_0) > 0$. If there are $x_1, x_2 \in (\alpha, \beta)$. $x_1 < x_2$, but $f(x_1) > f(x_2)$. Then there exists a constant M such that $f(x_1) > M > f(x_2)$. We define a set $S=\{x; x \in [x_1, x_2), f(x) \geq M\}$ which is a subset of (α, β).$S \neq \varnothing$ since $(x_1, x_1 + \delta_{x_1}) \subset S$ for some small $\delta_{x_1} > 0$. Let $\xi = \sup \ S$. Then $\xi < x_2 \leq \beta$ and $f(\xi) = M$ since $f(x)$ is continuous at $\xi \in (\alpha, \beta)$. For $x \in (\xi, \xi + \delta)$ where δ is a small positive number, $f(x) < M = f(\xi)$. Thus, $\frac{f(x)-f(\xi)}{x-\xi} < 0$, $x \in (\xi, \xi + \delta)$. It leads to $D^{+}f(\xi) \leq 0$. A contradiction appears.

If there is some $x_0 \in (\alpha, \beta)$, such that $D^{+}f(x_0) = 0$, then we consider a function $g(x) = f(x) + \varepsilon x$ with $\varepsilon > 0$. It is obvious that $D^{+}g(x_0) = D^{+}f(x_0) + \varepsilon > 0$. Then $g(x)$ is an increasing function, i.e., $f(x_1) + \varepsilon x_1 = g(x_1) \leq g(x_2) = f(x_2) + \varepsilon x_2$ if $x_1 < x_2$. Or equivalently, $f(x_1) \leq f(x_2) + \varepsilon(x_2 - x_1)$. Because $\varepsilon > 0$ can be selected arbitrarily, we have $f(x_1) \leq f(x_2)$. □

Remark 1 Theorem 2.5.1 is still valid if the domain $[\alpha, \beta]$ is replaced by (α, β) if at both terminals α and β only the single-side derivatives are considered. □

Remark 2 In general, if $f(x)$ is monotonously increasing on $B(x_0, \delta)$ we cannot conclude that $D^+f\,(x_0) > 0$. □

Remark 3 Theorem 2.5.1 illustrates that for the verification of monotonicity of a continuous function the four Dini derivatives have the same function. A further conclusion is if $f(x)$ is monotonous on a $(\alpha,\ \beta) \subset \mathbb{R}$, then for almost all point in $(\alpha,\ \beta)$ the four derivatives are equal and take finite values. □

Lemma 2.5.2 Suppose $f, g : [\alpha, \beta] \to \mathbb{R}$ are two continuous functions, if for every $x \in (\alpha, \beta) D^+f(x) \le g(x)$, then

$$f\,(\beta) - f\,(\alpha) \le \int\limits_{\alpha}^{\beta} g(x)dx$$

Proof Because $D^+f\,(\alpha) \le g\,(\alpha)$, for every $\varepsilon > 0$, there exists a $\delta_1 = \delta\,(\alpha, \varepsilon) > 0$ such that $\alpha + \delta_1 < \beta$, and for every $\Delta x \in [0, \delta_1]$,

$$\frac{f\,(\alpha + \Delta x) - f\,(\alpha)}{\Delta x} \le g\,(\alpha) + \varepsilon$$

Let $\Delta x = \delta_1$. Then the inequality leads to $f\,(\alpha + \delta_1) - f\,(\alpha) \le \delta_1 g\,(\alpha) + \delta_1 \varepsilon$.

Similarly, by $D^+f\,(\alpha + \delta_1) \le g\,(\alpha + \delta_1)$, there is a $\delta_2 = \delta\,(\alpha + \delta_1, \varepsilon) > 0$ such $\alpha + \delta_1 + \delta_2 < \beta$ and

$$f\,(\alpha + \delta_1 + \delta_2) - f\,(\alpha + \delta_1) \le \delta_2 g\,(\alpha + \delta_1) + \delta_2 \varepsilon$$

The procedure can be repeated before $\alpha + \sum\limits_{i=1}^{K} \sigma_i = \beta_.$[15] In the $(k+1)$th step, we obtain

$$f\left(\alpha + \sum_{i=1}^{k+1} \delta_i\right) - f\left(\alpha + \sum_{i=1}^{k} \delta_i\right) \le \delta_{k+1} g\left(\alpha + \sum_{i=1}^{k} \delta_i\right) + \delta_{k+1}\varepsilon, 1 \le k \le K.$$

Summing up these inequalities yields

$$f\,(\beta) - f\,(\alpha) \le \sum_{i=1}^{K} \delta_i g\left(\alpha + \sum_{k=0}^{i-1} \delta_k\right) + (\beta - \alpha)\,\varepsilon \tag{2.5.1}$$

[15] It is reasonable because on the interval $[\alpha, \beta], f(x)$ is equicontinuous.

where $\delta_0 = 0$. Because g is a continuous function, it is Lesbesgue integrable. When $\max \delta_i \to 0$, the right side of Inequality (2.5.1) converges to $\int_{\alpha}^{\beta} g(x)dx + (\beta - \alpha)\,\varepsilon$. The ε is selected arbitrarily, hence, the conclusion is verified. □

We now extend the definition of Dini derivatives to $\mathbb{R}^n$. Because there are infinite directions for a point in $\mathbb{R}^n$, we have to revise the definition.

Definition 2.5.1 Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a single-valued mapping, $x_0,\ \upsilon \in \mathbb{R}^n$, then

$$D^{+}f(x_0)(\upsilon) = \lim_{h\downarrow 0} \sup_{\upsilon' \to \upsilon} \frac{f(x_0 + h\upsilon') - f(x_0)}{h}$$

$$D^{-}f(x_0)(\upsilon) = \lim_{h\downarrow 0} \inf_{\upsilon' \to \upsilon} \frac{f(x_0 + h\upsilon') - f(x_0)}{h}$$

are called to be the upper Dini derivative and lower Dini derivative, respectively, of f at x_0 along with the vector υ. □

Readers may find that the definition of Dini derivatives is quite similar to the direction derivative of the convex function defined in Sect. 1.3 except the limitation of $\upsilon' \to \upsilon$ which leads to supremum and infimum. The introduction of supremum and infimum assures the existence of Dini derivatives in $\mathcal{R}$, the set of closed real numbers. However, if f is a Lipschitzian function, then Dini derivatives are all finite.

Readers may find that notations used for $n-$ dimensional function are quite different from those for $\mathbb{R}$. For example, by using the notating method for multivariable case, the upper Dini derivative $D^{+}f(x_0)$ given at the beginning of this subsection should be changed to $D^{+}f(x_0)(1)$. In the notation $D^{+}f(x_0)(1)$, the superscript "+" means supremum that differs from that in $D^{+}f(x_0)$ where means convergence from right side, and "1" means the right limitation. By such a notating method $D^{-}f(x_0)$ should be changed to $D^{+}f(x_0)(-1)$, $D_{+}f(x_0)$ should to $D^{-}f(x_0)(1)$, and $D_{-}f(x_0)$ to $D^{-}f(x_0)(-1)$.

The following theorem plays a basic role in the direct Lyapunov method.

Theorem 2.5.1 Suppose $x(t) : [0, T] \to \Omega \subset \mathbb{R}^n$ is a solution of n-dimensional differential equation $\dot{x} = f(x)$. Suppose $V : \mathbb{R}^n \to \mathbb{R}$ is a continuously single-valued and Lipschitzian function. Then the compounded function $V(x(t)) : [0, T] \subset \mathbb{R} \to \mathbb{R}$ satisfies that

$$D^{+}V(x(t))(1) = \lim_{h\downarrow 0} \sup \frac{V(x(t) + hf(x(t))) - V(x(t))}{h}$$

$$D^{-}V(x(t))(1) = \lim_{h\downarrow 0} \inf \frac{V(x(t) + hf(x(t))) - V(x(t))}{h}$$

Proof By definition of upper Dini derivative, we have

$$
\begin{aligned}
D^{+}V\left(x(t)\right)(1) &= \lim_{h\downarrow 0}\ \sup \frac{V\left(x\left(t+h\right)\right)-V\left(x(t)\right)}{h} \\
&= \lim_{h\downarrow 0}\ \sup_{o(h)} \frac{V\left(x(t)+hf\left(x(t)\right)+o(h)\right)-V\left(x(t)\right)}{h} \\
&\leq \lim_{h\downarrow 0}\ \sup_{o(h)} \frac{V\Big(x(t)+hf\left(x(t)\right)+L\left|o(h)\right|-V\left(x(t)\right)}{h} \\
&= \lim_{h\downarrow 0}\ \sup \frac{V\Big(x(t)+hf\left(x(t)\right)-V\left(x(t)\right)}{h},
\end{aligned}
$$

where $x\left(t+h\right)=x(t)+\dot{x}(t)h+o(h)=x(t)+hf(x)+o(h)$ is the Taylor expansion. $o(h)$ is an infinitesimal with higher order than h. At the third step, we apply mean value theorem where L is the Lipschitzian constant. On the other hand,

$$
\begin{aligned}
D^{+}V\left(x(t)\right)(1) &= \lim_{h\downarrow 0}\ \sup \frac{V\left(x(t)+hf\left(x(t)\right)+o(h)\right)-V\left(x(t)\right)}{h} \\
&\geq \lim_{h\downarrow 0}\ \sup \frac{V\Big(x(t)+hf\left(x(t)\right)-L\left|o(h)\right|-V\left(x(t)\right)}{h} \\
&= \lim_{h\downarrow 0}\ \sup \frac{V\Big(x(t)+hf\left(x(t)\right)-V\left(x(t)\right)}{h}.
\end{aligned}
$$

The first equation is verified. The second one can be verified by a similar way. It is omitted. □

Theorem 2.5.1 illustrates that Dini derivatives can be applied in direct Lyapunov method as the same as general derivative.

2.5.2 Definitions of Stability of Differential Inclusions

The stability of differential inclusions is defined by following the Lyapunov theory for differential equations. Hence, we start with the concept of equilibrium.

Suppose $F:\mathbb{R}^n\rightarrow\mathbb{R}^n$ is a set-valued mapping. If $0\in F\left(x_0\right)$ for an $x_0\in\mathbb{R}^n$, then x_0 is called as an equilibrium of the differential inclusion $\dot{x}(t)\in F\left(x(t)\right)$ since $x(t)\equiv x_0$ is a solution of the differential inclusion. It is clear that a differential inclusion may hold infinite equilibriums.

If $0 \in F(x_0)$, then we can make a coordinate transformation $y = x - x_0$ and denote $\widehat{F}(y) = F(y + x_0)$. By such a transformation, $0 \in \widehat{F}(0)$, i.e., $y(t) \equiv 0$ is an equilibrium of differential inclusion $\dot{y}(t) \in \widehat{F}(y(t))$. Therefore, we always assume in the discussion of stability that the equilibrium is the origin.

We now introduce the **K**-class functions and **L**-class functions.

$\alpha : \mathbb{R}(\geq 0) \to \mathbb{R}(\geq 0)$ is a function of **K**-class if $\alpha(0) = 0$ and it is continuous and strictly monotonously increasing; additionally, if $\alpha(t) \to \infty\ (t \to \infty)$, then it is a function of **KR**-class. The facts are denoted by $\alpha \in \mathbf{K}$ and $\alpha \in \mathbf{KR}$, respectively.

$\beta : \mathbb{R}(\geq 0) \to \mathbb{R}(\geq 0)$ is a function of **L**-class if it is continuous and strictly monotonously decreasing, and $\beta(\infty) = 0$. The fact is denoted by $\beta \in \mathbf{L}$.

The two concepts can be compounded. For example, $k : \mathbb{R}(\geq 0) \times \mathbb{R}(\geq 0) \to \mathbb{R}(\geq 0)$ is a **KL**-class function if we fix the second variable it is a function of **K**-class, and if we fix the first variable it is in class **L**. Similarly, we have classes **KKL, KLL** and **KRL,** etc.

We list several fundamental properties of **K**-class functions. If $\alpha_1,\ \alpha_2 \in \mathbf{K}$, then $\alpha_1 + \alpha_2 \in \mathbf{K}$, $\alpha_1(\alpha_2) \in \mathbf{K}$ and $k\alpha_1 \in \mathbf{K}$ for $k > 0$. If $\alpha \in \mathbf{K}$ then α is invertible and its inverse $\alpha^{-1} \in \mathbf{K}$. The discussion about **L**-class functions is left to readers.

$V : \mathbb{R}^n \to \mathbb{R}$ is said to be positively definite if it satisfies that $V(x) > 0$ for $x \neq 0$ and $V(0) = 0$. The $V(x)$ is said to be semi-positively definite, if $V(0) = 0$ and $V(x) \geq 0$ for all $x \in \mathbb{R}^n$. Additionally, if $V(x)$ satisfies $V(x) \to \infty\ (x \to \infty)$ then the function is called by positively (semi-positively) definite infinity. A function $W : \mathbb{R}^n \to \mathbb{R}$ is said to be negatively (semi-negatively) definite if $-W$ is positively (semi- positively) definite. We can also define negatively (semi-negatively) definite infinite. The positively definite and negatively definite functions play important roles in the study of stability by using Lyapunov method. Usually, a positively definite function $V(x)$ is also called V-function for simplicity.

Lemma 2.5.3 If $V(x)$ is continuous and positively definite, then there are $\alpha_1,\alpha_2 \in \mathbf{K}$ such that

$$\alpha_2(\|x\|) \leq V(x) \leq \alpha_1(\|x\|)$$

Additionally, if $V(x)$ is positively definite infinity, then $\alpha_1, \alpha_2 \in \mathbf{KR}$.

Proof We prove the first part, and the second part can be proved by a similar way. Hence, we omit it.

Define $\phi_m(\|x\|) = \min\limits_{\|r\| \geq \|x\|} V(r)$, then $\phi_m(\lambda) : \mathbb{R} \to \mathbb{R}$ is monotonously increasing and is continuous since $V(x)$ is continuous. It is obvious that $\phi_m(\|x\|) \leq V(x)$. Let $\alpha_2(\|x\|) = \frac{\|x\|}{\|x\|+1}\phi_m(\|x\|)$. Then $\alpha_2 \in \mathbf{K}$ and $\alpha_2(\|x\|) < \phi_m(\|x\|) \leq V(\|x\|)$.

Define $\phi_M(\|x\|) = \max\limits_{\|r\| \leq \|x\|} V(r)$, then $\phi_M(\lambda) : \mathbb{R} \to \mathbb{R}$ is also monotonously increasing and continuous. It is obvious that $\phi_M(\|x\|) \geq V(x)$. Let $\alpha_1(\|x\|) = \phi_M(\|x\|) + \|x\|$. Then $\alpha_1 \in \mathbf{K}$ and $\alpha_1(\|x\|) > \phi_M(\|x\|) \geq V(\|x\|)$. □

The following lemma is fundamental for the stability study by using **K**-class functions and **L**-class functions. The proof is referred to (Sontag 1989) and is omitted.

Lemma 2.5.4 Suppose $\alpha \in \mathbf{K}$, $y : \mathbb{R} \to \mathbb{R}$ is the solution of the following Cauchy problem

$$\dot{y} = -\alpha(y),\ y\,(t_0) = y_0 \tag{2.5.2}$$

where $y_0 > 0$, then there exists a $\beta \in \mathbf{KL}$ such that $|y(t)| \leq \beta\,(y_0, t - t_0)$. □

We now turn to deal with the stability of differential inclusions. Let $S(F, x_0)$[16] be the set of solutions of Cauchy problem of differential inclusion

$$\dot{x}(t) \in F\,(x(t))\, x(0) = x_0 \tag{2.5.3}$$

Note that we have assumed throughout at this chapter that $0 \in F(0)$, and the investigation of stability is carried out about the origin.

Definition 2.5.2 Consider the following differential inclusion

$$\dot{x}(t) \in F\,(x(t)) \tag{2.5.4}$$

If for every $\varepsilon > 0$, there exists a $\delta = \delta\,(\varepsilon)$ such that:

(1) When $\|x_0\| < \delta$, for every $x(t, x_0) \in S\,(F, x_0)$, $\|x\,(t, x_0)\| < \varepsilon$,[17] then Inc. (2.5.4) is strongly stable with regard to the equilibrium $x(t) \equiv 0$;
(2) If there is an $x(t, x_0) \in S\,(F, x_0)$ such that $\|x\,(t, x_0)\| < \varepsilon$, then Inc. (2.5.4) is weakly stable with regard to the equilibrium $x(t) \equiv 0$;
(3) With regard to the equilibrium $x(t) \equiv 0$, Inc. (2.5.4) is unstable if it is not weakly stable. □

Definition 2.5.3

(1) If Inc. (2.5.4) is strongly stable, additionally, for every $x(t, x_0) \in S\,(F, x_0)$ such that

$$\lim_{t\to\infty} x\,(t, x_0) = 0$$

then it is strongly asymptotically stable with regard to the equilibrium $x(t) \equiv 0$;
(2) If Inc. (2.5.4) is weakly stable, additionally, there is an $x(t, x_0) \in S\,(F, x_0)$ such that

$$\lim_{t\to\infty} x\,(t, x_0) = 0$$

[16] When we deal with the stability of differential equation or inclusion, the existence interval of a solution is always $[t_0, \infty)$, where t_0 is the initial time. Hence the subscript $[t_0, \infty)$ is omitted.

[17] The norm is $\|x\,(t, x_0)\|_{\mathbb{R}} = \sqrt{\max\limits_t x_1^2\,(t, x_0) + \max\limits_t x_2^2\,(t, x_0) + \cdots + \max\limits_t x_n^2\,(t, x_0)}$. It is a function of t. For the sake of convenience we omit the subscript $\mathbb{R}$.

then it is weakly asymptotically stable with regard to the equilibrium $x(t) \equiv 0$. □

Remark The stabilities of Inc. (2.5.4) can be equivalently defined by using **K**-class and **KL**-class functions. For example, with regard to the equilibrium $x(t) \equiv 0$ Inc. (2.5.4) is strongly stable if there is an $\alpha \in \mathbf{K}$, such that for every $x(t, x_0) \in S(F, x_0) \|x(t, x_0)\| \leq \alpha(\|x_0\|)$ in a neighborhood of the origin of $\mathbb{R}^n$. With regard to the equilibrium $x(t) \equiv 0$ Inc. (2.5.4) is strongly asymptotically stable if there is a $\beta \in \mathbf{KL}$, such that for every $x(t, x_0) \in S(F, x_0)$, $\|x(t, x_0)\| \leq \beta(\|x_0\|, t)$ in a neighborhood of the origin of $\mathbb{R}^n$. Readers can try to give the definitions for weak stability and weakly asymptotical stability for Inc. (2.5.4) by using **K**-class and **KL**-class functions. □

Usually, the word "strongly" in strongly stable and strongly asymptotically stable is omitted. Furthermore, we only consider the stability with regard to the equilibrium $x(t) \equiv 0$, hence the sentence "with regard to the equilibrium $x(t) \equiv 0$" is also omitted. We just call that Inc. (2.5.4) is stable, weakly stable, asymptotically stable, weakly asymptotically stable, and unstable, respectively.

2.5.3 Lyapunov-like Criteria for Stability of Differential Inclusions

The direct Lyapunov method is the uppermost method used in the investigation of differential equations as well as of differential inclusions. This subsection presents several Lyapunov-like criteria for differential inclusions.

We continue to consider Inc. (2.5.4). It is assumed that the set-valued mapping $F(x)$ is upper semi-continuous, and with convex and compact value.

Theorem 2.5.2 If there exist a positive constant η, a positively definite and continuous function $V : \mathbb{R}^n \to \mathbb{R}$, and a semi-negatively definite function $W : \mathbb{R}^n \to \mathbb{R}$ such that for every $\|x\| \leq \eta$ and $\upsilon \in F(x)$,

$$D^+V(x)(\upsilon) \leq W(x)$$

then Inc. (2.5.4) is stable.

Proof Let $x(t, x_0) \in S(F, x_0)$ be a solution. We now fix the time t and consider Dini derivative

$$D^+V(x(t, x_0))(1) = \lim\sup_{h \downarrow 0} \frac{V(x(t+h, x_0)) - V(x(t, x_0))}{h}$$

By the definition of supremum, there is sequence $\{h_n, n = 1, 2, \dots\}$ such that $h_n \downarrow 0 \quad (n \to \infty)$ and

$$D^+V(x(t, x_0))(1) = \lim_{n \to \infty} \frac{V(x(t+h_n, x_0)) - V(x(t, x_0))}{h_n}$$

Denote $\upsilon_n = \frac{x(t+h_n,x_0)-x(t,x_0)}{h_n}$. Then $x\left(t+h_n,x_0\right) = x\left(t,x_0\right) + h_n\upsilon_n$. Because $\dot{x}\left(t,x_0\right) \in F\left(x\left(t,x_0\right)\right)$, for $\gamma > 0$ given arbitrarily, there exists an N, when $n > N$

$$\upsilon_n = \frac{x\left(t+h_n,x_0\right)-x\left(t,x_0\right)}{h_n} \in F\left(x\left(t,x_0\right)\right) + \gamma\overline{B}$$

Because $F(x(t,x_0))$ is a compact set, $\{\upsilon_n\}$ holds a convergent subsequence. Without loss of generality, we can assume that $\upsilon_n \to \upsilon$, then $\upsilon \in F\left(x\left(t,x_0\right)\right)$. Thus,

$$\begin{aligned}D^+V\left(x\left(t,x_0\right)\right)(1) &= \limsup_{n\to\infty} \frac{V\left(x\left(t+h_n,x_0\right)\right)-V\left(x\left(t,x_0\right)\right)}{h_n} \\ &= \limsup_{n\to\infty} \frac{V\left(x\left(t,x_0\right)+h_n\upsilon_n\right)-V\left(x\left(t,x_0\right)\right)}{h_n} \\ &\le D^+V\left(x\left(t,x_0\right)\right)(\upsilon) \\ &\le W\left(x\left(t,x_0\right)\right) \ .\end{aligned}$$

This inequality is obtained from the condition that $D^+V(x)\,(\upsilon) \le W(x)$ for every $\upsilon \in F(x)$. This is a very useful conclusion.

By Lemma 2.5.2, for $t > 0$, we have

$$V\left(x\left(t,x_0\right)\right) - V\left(x(0)\right) \le \int_0^t W\left(x\left(s,x_0\right)\right) ds < 0 \tag{2.5.5}$$

or $V\left(x\left(t,x_0\right)\right) \le V\left(x_0\right)$. Using Lemma 2.5.3, two functions $\alpha_1, \alpha_2 \in \mathbf{K}$ can be found such that $\alpha_2\left(\|x\|\right) \le V(x) \le \alpha_1\left(\|x\|\right)$. For every $\eta \ge \varepsilon > 0$, we can find a $\delta > 0$ such that $\alpha_1(\delta) \le \alpha_2(\varepsilon)$. Now if $\|x_0\| < \delta$, then

$$\alpha_2\left(\left\|x\left(t,x_0\right)\right\|\right) \le V\left(x\left(t,x_0\right)\right) \le V\left(x_0\right) \le \alpha_1\left(\|x_0\|\right) < \alpha_1(\delta) \le \alpha_2(\varepsilon)$$

The last inequality implies $\left\|x\left(t,x_0\right)\right\| < \varepsilon$ for every $t \ge 0$, i.e., Inc. (2.5.4) is stable. □

In the proof of Theorem 2.5.2, we have established a useful inequality that

$$D^+V\left(x(t)\right)(1) \le D^+V\left(x(t)\right)(\upsilon) \tag{2.5.6}$$

where $\upsilon \in F\left(x(t)\right)$. The left side is $D^+V\left(x(t)\right)(1)$ whose argument is time t, but in $D^+V\left(x(t)\right)(\upsilon)$ the arguments are vector $x(t)$ and vector υ.

From the proof of Theorem 2.5.2, we can find that the upper semi-continuity and convexity compactness are only used to assure the existence of solutions. If we are sure that the solutions exist, then we only need the compactness to assure the convergence.

Theorem 2.5.3 If there exist a positive constant η, a positively definite and continuous function $V:\mathbb{R}^n \to \mathbb{R}$, and a negatively definite function $W:\mathbb{R}^n \to \mathbb{R}$ such that for every $\|x\| \le \eta$ and $\upsilon \in F(x)$,

$$D^+V(x)\,(\upsilon) \le W(x)$$

then Inc. (2.5.4) is asymptotically stable.

Proof The conditions of Theorem 2.5.3 implies those of Theorem 2.5.2, hence, for every $\varepsilon \in [0, \eta]$, there exists a δ>0 such that if $\|x_0\| < \delta$, then every $x\,(t, x_0) \in S(F, x_0)$ satisfies that $\|x\,(t, x_0)\| < \varepsilon$. We only need to prove that $x\,(t, x_0) \to 0 \quad (t \to \infty)$.

The conclusion is verified by reducing a contradiction. If there is a solution $\overline{x}\,(t, x_0) \in S(F, x_0)$ where $\|x_0\| < \delta$ and $\|\overline{x}\,(t, x_0)\| < \varepsilon$, but $\overline{x}\,(t, x_0)$ is not convergent to the origin. Therefore, there exist a constant l with $\varepsilon > l > 0$ and a sequence $\{t_n\} \subset \mathbb{R}\,(> 0)$ with $t_n \to \infty \quad (n \to \infty)$ such that $\|\overline{x}\,(t_n, x_0)\| \ge l$. Now we replace ε by this l, and repeat the proof of Theorem 2.5.2. Then we can obtain a $\overline{\delta} > 0$, when $\|\overline{x}_0\| < \overline{\delta}$, $\|x\,(t, \overline{x}_0)\| < l$ for every $x\,(t, \overline{x}_0) \in S\,(F, \overline{x}_0)$. Thus, in the solution $\overline{x}\,(t, x_0)$, x_0 has to satisfy that $\overline{\delta} \le \|x_0\| \le \delta$, and $\overline{x}\,(t, x_0)$ to satisfy $\overline{\delta} \le \|\overline{x}\,(t, x_0)\| \le \varepsilon$.

Let $-\lambda$ be the maximum of $W(x)$ in the area $\overline{\delta} \le \|x\| \le \eta$. Then $\lambda > 0$. Inequality (2.5.5) leads to

$$V\,(\overline{x}\,(t, x_0)) - V\,(\overline{x}\,(0, x_0)) \le \int_0^t W\,(\overline{x}\,(s, x_0))\,ds \le -\lambda t \tag{2.5.7}$$

It follows that $V\,(\overline{x}\,(t, x_0)) \le -\lambda t + V\,(\overline{x}\,(0, x_0))$. The left side will intend to negative infinite. This is impossible since $V(x)$ is positive definite. Consequently, every solution converges to the origin. □

The next theorem is about instability.

Theorem 2.5.4 If there exist a positive constant η, a continuous function $V:\mathbb{R}^n \to \mathbb{R}$ with $V(0) = 0$ and a negatively definite function $W:\mathbb{R}^n \to \mathbb{R}$ such that for every $\|x\| \le \eta$ and $\upsilon \in F(x)$,

$$D^+V(x)\,(\upsilon) \le W(x)$$

Moreover, for every $\delta > 0$, there is at least one $\|x\| < \delta$ such that $V(x) < 0$. Then Inc. (2.5.4) is unstable.

Proof The condition of Theorem 2.5.4 implies $\min\limits_{\|x\| \le \delta} V(x) < 0$ for $\delta > 0$ given arbitrarily. We denote $x_\delta = \arg \min\limits_{\|x\| \le \delta} V(x)$, and consider the solution $x(t, x_\delta)$. By the same procedure did in Theorem 2.5.2, when $\|x\,(t, x_\delta)\| \le \eta$, Inequality (2.5.5) holds, i.e.,

$$V\,(x\,(t, x_\delta)) - V\,(x_\delta) = V\,(x\,(t, x_\delta)) - V\,(x(0)) \le \int_0^t W\,(x\,(s, x_\delta))\,ds \le 0$$

hence, $V\left(x\left(t,x_\delta\right)\right) \leq V\left(x_\delta\right)$ for $t \in \left\{t \geq 0;\ \left\|x\left(t,x_\delta\right)\right\| \leq \eta\right\}$. Because $x_\delta = \arg\min\limits_{\|x\|\leq\delta} V(x)$, it is certain that $\left\|x\left(t,x_\delta\right)\right\| \geq \delta$. Now we denote $-\lambda = \max\limits_{\delta\leq\|x\|\leq\eta} W(x)$, then $\lambda > 0$. By a similar process of Inequality (2.5.7), we can obtain

$$V\left(x\left(t,x_\delta\right)\right) \leq V\left(x_\delta\right) - \lambda t \tag{2.5.8}$$

If for every t, $\left\|x\left(t,x_\delta\right)\right\| \leq \eta$, then Inequality (2.5.8) leads to $V\left(x\left(t,x_\delta\right)\right) \to -\infty$ $(t \to \infty)$. It contradicts to the fact $V(x)$ is continuous when $\|x\| \leq \eta$. Consequently, there is a T such that $\left\|x\left(T,x_\delta\right)\right\| = \eta$. It means the solution $x(t,x_\delta)$ is unstable. □

The last two theorems of this section are concerned with the weak stability. We prove firstly a lemma for $D^- V(x)\,(\upsilon)$.

Lemma 2.5.5 Suppose the set-valued mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is bounded and Lipschitzian, and with closed and convex value. Suppose $V : \mathbb{R}^n \to \mathbb{R}$ and $W : \mathbb{R}^n \to \mathbb{R}$ are two Lipschitzian mappings. Then the following two conclusions are equivalent.

(1) For every $x \in \mathbb{R}^n$, there is a $\upsilon \in F(x)$ such that $D^- V(x)\,(\upsilon) \leq W(x)$;
(2) For every $x_0 \in \mathbb{R}^n$, there is an $x\left(t,x_0\right) \in S_{[0,\infty)}\left(F,x_0\right)$ such that for $t_2 \geq t_1 \geq 0$,

$$V\left(x\left(t_2,x_0\right)\right) - V\left(x\left(t_1,x_0\right)\right) \leq \int\limits_{t_1}^{t_2} W\left(x\left(s,x_0\right)\right) ds$$

Proof We firstly prove that the first conclusion implies the second one.

We define a function $V_\varepsilon\left(x\left(t,x_0\right),t\right) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ for $x\left(t,x_0\right) \in S_{[0,\infty)}\left(F,x_0\right)$ as follows

$$V_\varepsilon\left(x\left(t,x_0\right),t\right) = V\left(x\left(t,x_0\right)\right) - V\left(x_0\right) - \int\limits_0^t W\left(x\left(s,x_0\right)\right) ds - \varepsilon t \tag{2.5.9}$$

where $\varepsilon > 0$ and $x_0 \in \mathbb{R}^n$ are treated as parameters. In any bonded area, $V_\varepsilon(x(t,x_0),t)$ is Lipschitzian for both variables $x(t,x_0)$ and t. It is direct to verify that $V_\varepsilon\left(x(0),0\right) = 0$. The following proof consists of two steps.

(i) For a given $T>0$, there is an $x(t) \in S_{[0,T]}\left(F,x_0\right)$ such that for every $t \in [0,T] V_\varepsilon\left(x(t),t\right) \leq \varepsilon$.

For every $x(t) \in S_{[0,T]}\left(F,x_0\right)$, we define a subset $\theta(x(t))$ of $[0,T]$ as follows

$$\theta\left(x(t)\right) = \left\{t;\ \ V_\varepsilon\left(x(t),t\right) \leq 0,\ \ \max_{s\in[0,t]} V_\varepsilon\left(x(t),t\right) \leq \varepsilon\right\}$$

Because $V_\varepsilon\left(x(0),0\right) = 0$, $0 \in \theta\left(x(t)\right)$, i.e., $\theta\left(x(t)\right) \neq \varnothing$. Denote $t_x = \sup\theta\left(x(t)\right)$, since $V_\varepsilon(x(t),t)$ is continuous for t, we have $V_\varepsilon\left(x\left(t_x\right),t_x\right) = 0$ if $t_x \neq T$. We further define

$$\Theta = \left\{t_x; x(t) \in S_{[0,T]}\left(F, x_0\right)\right\}$$

and $\widehat{t} = \sup\ \Theta$.

We now try to verify that there is an $\hat{x}(t) \in S_{[0,T]}\left(F, x_0\right)$ such that $\widehat{t} = t_{\hat{x}}$.

By the definition of supremum, there is $\{t\}_k \subset \Theta$ such that $t_k \uparrow \widehat{t}(k \to \infty)$. For every t_k, there is an $x_k(t) \in S_{[0,T]}\left(F, x_0\right)$ such that $t_k = t_{x_k} = \sup\theta\left(x_k(t)\right)$. $F(x)$ is bounded for every x, so is $\left\{\dot{x}_k(t)\right\}$. It implies the series $\{x_k(t)\}$ is equicontinuous, hence it holds a convergent subseries. Without loss of generality, we can assume $x_k(t) \to \hat{x}(t),\ \ k \to \infty$. By Corollary 2.4.1, we conclude that $\hat{x}(t) \in S_{[0,T]}\left(F, x_0\right)$. $V_\varepsilon(x,t)$ is continuous for x, hence, $\widehat{t} = t_{\hat{x}}$.

The fact $\widehat{t} = T$ is verified by contradiction. If $\widehat{t} < T$, then $V_\varepsilon\left(\hat{x}\left(\widehat{t}\right), \widehat{t}\right) = 0$. For this $\hat{x}(t)$, there is a $\upsilon \in F\left(\hat{x}\left(\widehat{t}\right)\right)$ such that $D^- V\left(\hat{x}\left(\widehat{t}\right)\right)(\upsilon) \le W\left(\hat{x}\left(\widehat{t}\right)\right)$ by the first condition of the lemma. Consequently,

$$\begin{aligned} D^- V_\varepsilon\left(\hat{x}\left(\widehat{t}, x_0\right), \widehat{t}\right)(\upsilon, 1) &= D^-\left[V\left(\hat{x}\left(\widehat{t}, x_0\right)\right) - V\left(x_0\right)\right](\upsilon) \\ &\quad - D^-\left[\int_0^{\widehat{t}} W\left(\hat{x}\left(s, x_0\right)\right) ds + \varepsilon t\right](1) \\ &= D^- V\left(\hat{x}\left(\widehat{t}, x_0\right)\right)(\upsilon) - \lim_{\Delta t \to 0} \int_{\widehat{t}}^{\widehat{t}+\Delta t} W\left(\hat{x}\left(s, x_0\right)\right) ds - \varepsilon \\ &\le -\varepsilon. \end{aligned} \tag{2.5.10}$$

A function $y(t)$ is now defined

$$y(t) = \hat{x}\left(\widehat{t}\right) + \left(t - \widehat{t}\right)\upsilon, \quad \widehat{t} \le t \le T$$

We now apply Theorem 2.3.3, and the initial time is $t_0 = \widehat{t}$. Using the notations of Theorem 2.3.3, $\delta = 0$ and

$$\begin{aligned} d\left(\dot{y}(t), F\left(y(t)\right)\right) &= d\Big(\upsilon, F\left(\hat{x}\left(\widehat{t}\right) + \left(t - \widehat{t}\right)\upsilon\right) \\ &\le d\Big(F\left(\hat{x}\left(\widehat{t}\right)\right), F\left(\hat{x}\left(\widehat{t}\right) + \left(t - \widehat{t}\right)\upsilon\right) \\ &\le l\,\|\upsilon\|\,\left\|t - \widehat{t}\right\|\ . \end{aligned}$$

It implies that $\rho(t)$ can be $l\,\|\upsilon\|\left(t - \widehat{t}\right)$. Theorem 2.3.3 concludes that there exists an $x(t) \in S_{\left[\widehat{t},T\right]}\left(F, \hat{x}\left(\widehat{t}\right)\right)$ such that

$$\|y(t) - x(t)\| \le \int_{\widehat{t}}^{t} e^{l(t-s)} \cdot l\,\|\upsilon\|\left(s - \widehat{t}\right) ds = \|\upsilon\|\left(\frac{1}{l}e^{l\left(t-\widehat{t}\right)} - \frac{1}{l} - \left(t - \widehat{t}\right)e^{l\left(t-\widehat{t}\right)}\right)$$

For convenience, we denote $\alpha\left(t-\widehat{t}\right) = \int\limits_{\widehat{t}}^{t} e^{l(t-s)}\cdot l\left\|\upsilon\right\|\left(s-\widehat{t}\right)ds$. It is direct to obtain

$$\lim_{t\downarrow\widehat{t}}\frac{\alpha\left(t-\widehat{t}\right)}{t-\widehat{t}}=0 \tag{2.5.11}$$

Let L be the Lipschitzian constant of V_ε. Then

$$V_\varepsilon\left(x(t),t\right)\le V_\varepsilon\left(y(t),t\right)+L\alpha\left(t-\widehat{t}\right) \tag{2.5.12}$$

By Inequality (2.5.10), we have

$$\lim_{t\downarrow\widehat{t}}\frac{V_\varepsilon\left(\hat{x}\left(\widehat{t}\right)+\left(t-\widehat{t}\right)\upsilon,t\right)-V_\varepsilon\left(\hat{x}\left(\widehat{t}\right),\widehat{t}\right)}{t-\widehat{t}}\le-\varepsilon \tag{2.5.13}$$

Because $V_\varepsilon\left(\hat{x}\left(\widehat{t}\right),\widehat{t}\right)=0$, Inequality (2.5.13) implies there is a $\tilde{t}>\widehat{t}$ such that for every $t\in\left[\widehat{t},\tilde{t}\right]$,

$$V_\varepsilon\left(y(t),t\right)=V_\varepsilon\left(\hat{x}\left(\widehat{t}\right)+\left(t-\widehat{t}\right)\upsilon,t\right)\le-\frac{\varepsilon}{2}\left(t-\widehat{t}\right)$$

We can further require that for this $\tilde{t}$

$$\alpha\left(t-\widehat{t}\right)\le\frac{\varepsilon}{2L}\left(t-\widehat{t}\right)$$

when $t\in\left[\widehat{t},\tilde{t}\right]$. Substituting these two inequalities to Inequality (2.5.12), we obtain $V_\varepsilon\left(x\left(\tilde{t}\right),\tilde{t}\right)\le 0$.

On the other hand,

$$\begin{aligned}V_\varepsilon\left(x(t),t\right)&\le V_\varepsilon\left(y(t),t\right)+L\alpha\left(t-\widehat{t}\right)\\&=V_\varepsilon\left(\hat{x}\left(\widehat{t}\right)+\left(t-\widehat{t}\right)\upsilon,t\right)+L\alpha\left(t-\widehat{t}\right)\\&\le V_\varepsilon\left(\hat{x}\left(\widehat{t}\right),\widehat{t}\right)+L\left\|\upsilon\right\|\left(t-\widehat{t}\right)+L\left(t-\widehat{t}\right)+L\alpha\left(t-\widehat{t}\right)\\&=L\left\|\upsilon\right\|\left(t-\widehat{t}\right)+L\alpha\left(t-\widehat{t}\right)+L\left(t-\widehat{t}\right)\quad.\end{aligned}$$

If $t-\widehat{t}$ is small enough, we can require $V_\varepsilon\left(x(t),t\right)\le\varepsilon$. The above discussion implies that for the function defined by

$$\tilde{x}(t)=\begin{cases}\hat{x}(t) & t\le\widehat{t},\\ x(t) & t\ge\widehat{t},\end{cases}$$

$\tilde{x}(t) \in S_{[0,T]}(F, x_0)$, and $\sup \theta(\tilde{x}(t)) > \widehat{t}$. It contradicts to $\widehat{t} = \sup \Theta$. Hence $\widehat{t} = T$.

(ii) There is an $x(t) \in S_{[0,T]}(F, x_0)$ such that Conclusion (2) of theorem is true.

Let $\{\varepsilon_n\}$ be a monotone series such that $\varepsilon_n \downarrow 0 \quad (n \to \infty)$. By the conclusion of (i), for every ε_n, there is an $x_n(t) \in S_{[0,T]}(F, x_0)$ such that $\max\limits_{s \in [0,T]} V_{\varepsilon_n}(x_n(s), s) \leq \varepsilon_n$. By the reason applied in the proof of (i), we can assume $x_n(t) \to x(t) \quad (n \to \infty)$ and then $x(t) \in S_{[0,T]}(F, x_0)$. By this $x(t)$, $\max\limits_{s \in [0,T]} V_{\varepsilon_n}(x_n(s), s) \leq \varepsilon_n$ leads to $V_0(x(t), t) \leq 0$, i.e.,

$$V(x(t)) - V(x_0) - \int_0^t W(x(s))\, ds \leq 0$$

If we replace the initial time by t_0, Then conclusion (2) is verified.

We now prove the second conclusion implies the first one.

Because F is a Lipschitzian set-valued mapping, it is also $\varepsilon-$ semi-upper continuous. Let $x(t) \in S_{[0,T]}(F, x_0)$. Then, for every integer k, we can find a $t_k > 0$ such that for $t \in [0, \ t_k]$

$$F(x(t)) \subset F(x(0)) + \frac{1}{k}B$$

We can further require $t_k \downarrow 0 \quad (k \to \infty)$. Consider now

$$\frac{x(t_k) - x(0)}{t_k} = \frac{1}{t_k}\int_0^{t_k} \dot{x}(s) ds \in F(x(0)) + \frac{1}{k}B$$

$F(x(0))$ is a bounded and closed set, hence $\left\{\frac{x(t_k)-x(0)}{t_k}\right\}$ holds a convergent subseries. Without loss of generality, we can assume $\frac{x(t_k)-x(0)}{t_k} \to \upsilon \quad (k \to \infty)$. Then $\upsilon \in F(x(0))$. The first conclusion of the lemma illustrates that

$$V(x(t_k)) - V(x_0) = V\left(x_0 + t_k \frac{x(t_k) - x(0)}{t_k}\right) - V(x(0)) \leq \int_0^{t_k} W(x(s))\, ds$$

By dividing both sides with t_k, and letting $k \to \infty$, the definition of Dini derivative leads to

$$D^- V(x_0)(\upsilon) \leq W(x_0)$$

The lemma is completely verified. □

We apply Lemma 2.5.5 to prove the following theorem.

Theorem 2.5.5 Suppose the set-valued mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is bounded and Lipschitzian, and is with closed and convex value. If there exist a positive constant η, a positive definite and Lipschitzian function $V : \mathbb{R}^n \to \mathbb{R}$ and a semi-negatively definite function $W : \mathbb{R}^n \to \mathbb{R}$ such that for every $\|x\| \le \eta$ there exists a $\upsilon \in F(x)$ such that

$$D^- V(x)\,(\upsilon) \le W(x)$$

Then Inc. (2.5.4) is weakly stable.

Proof By Lemma 2.5.5, for every $x_0 \in \mathbb{R}^n$ with $\|x_0\| \le \eta$, there is an $x(t) \in S_{[0,T]}\,(F, x_0)$ such that

$$V\,(x(t)) - V\,(x_0) \le \int\limits_0^t W\,(x(s))\,ds \le 0$$

V is positive definite, hence, there are $\alpha_1, \alpha_2 \in \mathbf{K}$ such that $\alpha_2\,(\|x\|) \le V(x) \le \alpha_1\,(\|x\|)$. It follows that

$$\alpha_2\,(\|x(t)\|) \le V\,(x(t)) \le V\,(x_0) \le \alpha_1\,(\|x_0\|)$$

For every $\varepsilon > 0$, if $\|x_0\| < \min\left\{\alpha_1^{-1}\alpha_2\,(\varepsilon)\,, \eta\right\}$, then $\alpha_2\,(\|x(t)\|) < \alpha_2\,(\varepsilon)$, or $\|x(t)\| < \varepsilon$. □

At last, we give a theorem about weakly asymptotical stability. It can be proved by a similar way to Theorem 2.5.3. But we prefer to provide a new method by using Lemma 2.5.4.

Theorem 2.5.6 Suppose the set-valued mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is bounded and Lipshitzian, and is with closed and convex value. If there exist a positive constant η, a positive definite and Lipschitzian function $V : \mathbb{R}^n \to \mathbb{R}$ and a negatively definite function $W : \mathbb{R}^n \to \mathbb{R}$ such that for every $\|x\| \le \eta$ there exists a $\upsilon \in F(x)$ such that

$$D^- V(x)\,(\upsilon) \le W(x)$$

Then Inc. (2.5.4) is weakly asymptotically stable.

Proof By Theorem 2.5.5, the conditions of Theorem 2.5.6 imply that for every $x_0 \in \mathbb{R}^n$ with $\|x_0\| \le \eta$, there is a stable solution $x(t) \in S_{[0,T]}\,(F, x_0)$. We now verify the solution is also asymptotically stable.

By Lemma 2.5.5, we have

$$V\,(x(t)) \le V\,(x_0) + \int\limits_0^t W\,(x(s))\,ds \tag{2.5.14}$$

By Lemma 2.5.3, there are $\alpha_1, \alpha_2, \alpha_3 \in \mathbf{K}$ such that $\alpha_2(\|x\|) \le V(x) \le \alpha_1(\|x\|)$ and $W(x) \le -\alpha_3(\|x\|)$.

Inequality (2.5.14) is equivalent to the following Cauchy problem

$$\frac{d\Big(V(x(t))\Big)}{dt} \le W(x(t)), \quad V(x(0)) = V(x_0)$$

We also have

$$W(x(t)) \le -\alpha_3(\|x(t)\|) \le -\alpha_3\Big(\alpha_1^{-1}(V(x(t)))\Big)$$

$\alpha_3\alpha_1^{-1}(\cdot)$ is a function in $\mathbf{K}$-class, by Lemma 2.5.4, there is $\beta \in \mathbf{KL}$ such that $V(x(t)) \le \beta(t, V(x_0))$, or $\|x(t)\| \le \alpha_2^{-1}(\beta(t, V(x_0)))$. The proof is completed. □

Problems

1. $f(x)$, $g(x)$: $\mathbb{R} \to \mathbb{R}$ are single-valued functions. Prove the following relations

   (1) $D^+(f(x) + g(x)) \le D^+f(x) + D^+g(x)$
   (2) $D^+(f(x) + g(x)) \ge D^+f(x) + D_+g(x)$

   By the above conclusions, establish the corresponding conclusions for $D_+(f(x) + g(x))$, $D^-(f(x) + g(x))$ and $D_-(f(x) + g(x))$.
2. $f(x)$: $\mathbb{R} \to \mathbb{R}$ is a single-valued function. Prove the following conclusions.

   (1) If $f(x_0) = \max\{f(x),\ x \in B(x_0, \delta)\}$, then $D^+(f(x_0)) \le 0 \le D_-(f(x_0))$,
   (2) If $f(x_0) = \min\{f(x),\ x \in B(x_0, \delta)\}$, then $D^-(f(x_0)) \le 0 \le D_+(f(x_0))$,

3. If $g(x)$ is differentiable, then the following equations hold.
   $D^+(f(x)g(x)) = f(x)\dot{g}(x) + g(x)D^+f(x)$, if $g(x) \ge 0$;
   $D^+(f(x)g(x)) = f(x)\dot{g}(x) + g(x)D_+f(x)$, if $g(x) \le 0$.
   For $D_+(f(x)g(x))$, $D^-(f(x)g(x))$ and $D_-(f(x)g(x))$, present similar conclusions.
4. Let $f(x)$, $g(x)$ be continuous functions, and $h(x) = \max(f(x),\ g(x))$. Prove that if $D^+f(x) \le 0$ and $D^+g(x) \le 0$, then $D^+h(x) \le 0$.
5. Prove the following conclusion given in Theorem 2.5.1 that

$$D^-f(x(t))(1) = \lim_{h\downarrow 0}\inf \frac{f(x(t) + hf(x(t))) - f(x(t))}{h}$$

6. If β_1 and β_2 are both $\mathbf{L}$-class functions, which of the following functions are in $\mathbf{L}$-class?

   (1) $\beta_1 + \beta_2$,
   (2) $a\beta_1$ $(a > 0)$,
   (3) $\beta_1(\beta_2(\cdot))$

7. If $\alpha(t)$ is a **K**-class function, and $\beta(t)$ is a **L**-class function, then $\beta(\alpha(t))$ and $\alpha(\beta(t))$ are all **L**-class functions.
8. The following statement is an alternative version of Theorem 2.5.3. Prove the statement.

   "If there are a constant $\eta > 0$, a positive and continuous function $V : \mathbb{R}^n \to \mathbb{R}$, and a **K**-class function α such that for every $\|x\| \leq \eta$ and $\upsilon \in F(x)$ the inequality $D^+V(x)(\upsilon) \leq -\alpha(\|x\|)$ holds, then Inc. (2.5.4) is asymptotically stable."
9. Try to use **K**-class function to alter Theorem 2.5.6 by imitating Problem 8.
10. Prove Theorem 2.5.6.
11. Consider discrete differential inclusion

$$x_{k+1} \in F(x_k), \quad k = 0, 1, 2, \ldots$$

   where $F : \mathbb{R}^n \to \mathbb{R}^n$ is a set-valued mapping. Suppose $0 \in F(0)$.

   The discrete differential inclusion is said to (strong) stable, if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $\|x_0\| < \delta\{x_k,\ k = 1, 2, \ldots\} \subset \varepsilon B$ where B is the unit ball in $\mathbb{R}^n$. Additionally, if $x_k \to 0$, then it is asymptotically stable.

   Give a condition under which the discrete differential inclusion is stable, and prove your condition is sufficient.
12. Consider the discrete differential inclusion given in Problem 11. Prove that if there are a positive and continuous function $V : \mathbb{R}^n \to \mathbb{R}$ and $\eta > 0, \lambda \in (0, 1)$, when $\|x\| \leq \eta$,

$$\inf_{\upsilon \in F(x)} V(\upsilon) \leq \lambda V(\upsilon)$$

   then the discrete differential inclusion is weakly asymptotically stable.

   Try to give a condition for weakly stable and prove it.
13. If the set-valued mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is bounded and Lipschitzian, and with closed and convex value. Suppose $V : \mathbb{R}^n \to \mathbb{R}$ is a Lipschitzian function. If for every $x \in \mathbb{R}^n$, there is a $\upsilon \in F(x)$ such that $D^-V(x)(\upsilon) \leq W(x)$, then for every $T > 0$, there is a solution $x(t) \in S_{[0,T]}(F, x_0)$ such that $V(x(t_1)) \leq V(x(t_2))$ provided that $T \geq t_1 \geq t_2 \geq 0$.

2.6 Monotonous Differential Inclusions

Monotonicity is a simple variation in the natural world. In mathematics, many investigations start with monotonous phenomenon, such as linear mappings, monotonous sequences, monotonous functions. This section deals with monotonous differential inclusions. The most interesting conclusion is that the solution of Cauchy problem of a maximal monotonous differential inclusion is unique. From this viewpoint the differential inclusion is quite similar to the differential equation. The authors thought that the most significance is the procedure of investigation. It shows a special project always needs some special techniques. We will start with Minty theorem, a precise proof is presented. Then we consider the Yosida approximation. At last we prove the main conclusion for the maximal monotonous differential inclusion.

2.6.1 Monotonous Set-valued Mappings and Their Properties

A differential inclusion is monotonous if the set-valued mapping in the right side of the inclusion is monotonous. Hence this subsection deals with monotonous set-valued mappings.

$f : \mathbb{R} \to \mathbb{R}$ is a single-valued mapping, f is said to be monotonous if $x_1 > x_2$ then $f(x_1) \geq f(x_2)$.[18] In the classical control theory, most nonlinear phenomena considered are monotonous, such as step nonlinearity, dead band nonlinearity, saturation nonlinearity. In order to extend the concept to the n-dimensional space or general Hilbert spaces,[19] we rewrite the requirement $x_1 > x_2 \Rightarrow f(x_1) \geq f(x_2)$ by $(x_1 - x_2)(f(x_1) - f(x_2)) \geq 0$, the restriction $x_1 > x_2$ is removed.

A single-valued mapping $f : X \to X$, where X is a Hilbert space, is monotonous if $\langle (x_1 - x_2), (f(x_1) - f(x_2)) \rangle \geq 0$ for arbitrary two vectors $x_1, x_2 \in X$ where $\langle \cdot, \cdot \rangle$ is the inner product. Sometimes, the inner product is also denoted by $(x_1 - x_2)^{\mathrm{T}} (f(x_1) - f(x_2))$, the superscript "T" means transposition.

Definition 2.6.1 $F : X \to X$ is a monotonous set-valued mapping, if for arbitrary two vectors $x_1, x_2 \in X$, $y_1 \in F(x_1)$ and $y_2 \in F(x_2)$ then $\langle (x_1 - x_2), (y_1 - y_2) \rangle \geq 0$;

If $\langle (x_1 - x_2), (y_1 - y_2) \rangle > 0$ for arbitrary $x_1 \neq x_2$, $y_1 \in F(x_1)$ and $y_2 \in F(x_2)$, the set-valued F is strictly monotonous;

F is a maximally monotonous set-valued mapping, if there is another set-valued mapping $F_1 : X \to X$ which holds the property that $F_1(x) \supset F(x)$ for every $x \in X$, then $F_1 = F$. □

Remark 1 In Definition 2.6.1, the space X can be replaced by a set $A \subset X$. we can say a set-valued mapping is monotonous at set A and maximally monotonous at set A. □

Remark 2 If $F : X \to X$ is a monotonous set-valued mapping, then $F^{-1} : X \to X$ is also a monotonous set-valued mapping. □

Remark 3 If F, G are two monotonous set-valued mappings, $\lambda, \mu \geq 0$ are two real numbers, then $\lambda F + \mu G$ is also a monotonous set-valued mapping. □

[18] The terminology used is somewhat different from the common meaning. In calculus, the monotonicity defined is called monotonous increase.

[19] In this section we do not restrict ourselves in a finite dimensional space, $\mathbb{R}^n$ can be replaced by a Hilbert space X.

Lemma 2.6.1 Let $F : X \to X$ be a set-valued mapping. Then F is monotonous if and only if for every positive real number $\lambda \geq 0$ and any $y_1 \in F(x_1)$, $y_2 \in F(x_2)$

$$\|x_1 - x_2\| \leq \|(x_1 - x_2) + \lambda (y_1 - y_2)\|$$

Proof Because

$$\|(x_1 - x_2) + \lambda (y_1 - y_2)\|^2 = \|x_1 - x_2\|^2 + \lambda^2 \|y_1 - y_2\|^2 + 2\lambda \langle (x_1 - x_2), (y_1 - y_2) \rangle$$

If the condition of Lemma 2.6.1 holds, then

$$\lambda^2 \|y_1 - y_2\|^2 + 2\lambda \langle (x_1 - x_2), (y_1 - y_2) \rangle \geq 0 \tag{2.6.1}$$

We assert Inequality (2.6.1) implies $\langle (x_1 - x_2), (y_1 - y_2) \rangle \geq 0$, i.e., F is monotonous. Otherwise, $\langle (x_1 - x_2), (y_1 - y_2) \rangle < 0$, then there exists a $\lambda > 0$, such that

$$\lambda \|y_1 - y_2\|^2 + \langle (x_1 - x_2), (y_1 - y_2) \rangle < 0$$

It conflicts to Inequality (2.6.1).

If f is monotonous, then $\langle (x_1 - x_2), (y_1 - y_2) \rangle \geq 0$. It implies Inequality (2.6.1), so is

$$\|(x_1 - x_2) + \lambda (y_1 - y_2)\| \geq \|x_1 - x_2\|$$

□

The advantage of Lemma 2.6.1 is that it does not involve inner product. Thus, some property of monotonous mappings may be extended to the Banach space.

The following example gives the difference of monotonous mapping and the maximal monotonous mapping.

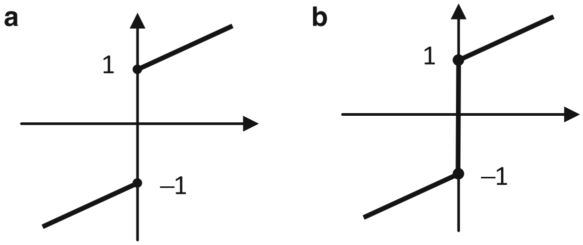
Example 2.6.1 Consider the following two set valued mappings

$$F_1(x) = \begin{cases} 0.5x + 1, & x > 0, \\ \{-1, 1\}, & x = 0, \\ 0.5x - 1, & x < 0; \end{cases} \quad F_2(x) = \begin{cases} 0.5x + 1, & x > 0, \\ [-1, \ 1], & x = 0, \\ 0.5x - 1, & x < 0. \end{cases}$$

Their graphs are given in Fig. 2.15.

$F_1(x)$ is not maximal since $F_2(0) \supset F_1(0)$ and $F_2(0) \neq F_1(0)$. But $F_2(x)$ is maximal since adding any a point in the $\mathbb{R}^2$ (Fig. 2.15b) will destroy the monotonicity. □

Fig. 2.15 Two monotonous set-valued mappings (**a**) Graph of $F_1(x)$ (**b**) Graph of $F_2(x)$



The above example directly leads to the following lemma, we omit its proof.

Lemma 2.6.2 Let $F : X \to X$ be a monotonous set-valued mapping. It is maximal if the conditions that $\langle u - x, v - y\rangle \geq 0$ and some $(x, y) \in \text{gra}\, F$ imply $v \in F(u)$ □

The following theorem presents the fundamental properties of maximally monotonous set-valued mappings.

Theorem 2.6.1 If $F : X \to X$ is a maximally monotonous set-valued mapping, then

(1) F is with closed and convex value;
(2) Suppose $x_n \to x$ $(n \to \infty)$, and $y_n \in F(x_n)$ such that $\langle y_n, z\rangle \to \langle y, z\rangle$ for every $z \in X$, then $y \in F(x)$.

Proof (1) By Lemma 2.6.2, we have

$$F(x) = \bigcap_{(u,v)\in \text{gra}\, F} \{y \in X, \langle x - u, y - v\rangle \geq 0\}$$

Let $(u, v) \in \text{gra}F$. Then the set $\{y; \langle x - u, y - v\rangle \geq 0\}$ is a closed and convex for any $x \in X$. Therefore as an intersection, $F(x)$ is closed and convex.

(2) By the conditions of theorem we have $\langle x_n - u, y_n - v\rangle \to \langle x - u, y - v\rangle$. Because $F : X \to X$ is a monotonous set-valued mapping and $y_n \in F(x_n)$, for every $(u, v) \in \text{gra}\, F \langle x_n - u, y_n - v\rangle \geq 0$. It follows $\langle x - u, y - v\rangle \geq 0$. F is maximal, by Lemma 2.6.1, $y \in F(x)$. The theorem is then verified. □

In functional analysis, the fact that $\langle y_n, z\rangle \to \langle y, z\rangle$ for every $z \in X$ is called that y_n weakly converges to y. The conclusion of (2) can also be stated as "if $x_n \to x$ strongly, $y_n \to y$ weakly, then $y \in F(x)$". The conclusion is then called "strong-weak compactness".

2.6.2 Minty Theorem

This subsection proves an important result of maximally monotonous set-valued mappings. It is called Minty theorem.

Suppose $\{\Omega_\lambda, \lambda \in \Lambda\}$ is a class of sets[20] where Λ is set of indexes. $\{\Omega_\lambda, \lambda \in \Lambda\}$ is said to hold the property of finite intersection: If for any finite subset $\{\Omega_{\lambda_1}, \ldots, \Omega_{\lambda_n}\}$ of $\{\Omega_\lambda, \lambda \in \Lambda\}$, their intersection $\bigcap\limits_{k=1}^{n} \Omega_{\lambda_k}$ is nonempty, then the intersection $\bigcap\limits_{\lambda \in \Lambda} \Omega_\lambda \neq \varnothing$. In topology there is a theorem as follows: Suppose $\{\Omega_\lambda, \lambda \in \Lambda\}$ is a class and Ω_λ's all are closed and contained in a compact set K. Then the class $\{\Omega_\lambda, \lambda \in \Lambda\}$ holds finite intersection property. The property is equivalent to finite covering theorem. The detailed proof will not be given in this book.

Before presenting Minty theorem, we need two lemmas.

Lemma 2.6.3 Suppose K and S are two sets of X, and K is compact. $\phi : K \times S \to \mathbb{R}$ is a single-valued functional. If for every $y_0 \in S$, $\phi(x, y_0)$ is lower semi-continuous. Suppose M is the class consisting of all finite subsets of S. Then there is a $\bar{x} \in K$ such that

$$\sup_{y \in S} \phi\left(\bar{x}, y\right) \leq \upsilon = \sup_{E \in M} \inf_{x \in K} \max_{y \in E} \phi\left(x, y\right)$$

Proof For a fixed $y \in S$, we define a set $\Omega_y = \{x \in K, \phi\left(x, y\right) \leq \upsilon\}$ where υ is real number defined by the theorem. From the definition of υ, we conclude that $\Omega_y \neq \varnothing$. The functional $\phi(x, y)$ is lower semi-continuous, hence, Ω_y is closed. $\Omega_y \subset K$, K is compact, so is the Ω_y.

Denote $\phi\left(x, E\right) = \max\limits_{y \in E} \phi\left(x, y\right)$. If E is fixed, then $\phi(x, E)$ is a functional of $x \in K$. It can be verified that $\phi(x, E)$ is a lower semi-continuous functional (to see Sect. 1.4). The functional $\phi(x, E)$ possesses such a property that for $x \in K$ and $y_i \in E$, $\phi(x, E) \geq \phi\left(x, y_i\right)$. K is compact, there is a $\bar{x}_E \in K$ such that $\phi\left(\bar{x}_E, E\right) = \min\limits_{x \in K} \phi\left(x, E\right) = \inf\limits_{x \in K} \phi\left(x, E\right) \leq \upsilon$ for every E. Combining the above reasoning, $\phi\left(\bar{x}_E, y_i\right) \leq \phi\left(\bar{x}_E, E\right) \leq \upsilon$. Furthermore, $\bar{x}_E \in \bigcap\limits_{y \in E} \Omega_y$. The fact supports that $\bigcap\limits_{y \in E} \Omega_y \neq \varnothing$ for all finite set E, i.e., it holds finite intersection property. K is compact, therefore, $\bigcap\limits_{y \in S} \Omega_y \neq \varnothing$. There is a $\bar{x} \in \bigcap\limits_{y \in S} \Omega_y$. By the definition of Ω_y, this $\bar{x}$ has the property that $\phi\left(\bar{x}, y\right) \leq \upsilon, \quad y \in S$. The conclusion is proved. □

The statement of Lemma 2.6.3 is quite complicated, Fig. 2.16 is used to explain the conclusion. In Fig. 2.16, $E = \{y_1, y_2, y_3\}$, three slimsy lines are $\phi(x, y_1)$, $\phi(x, y_2)$ and $\phi(x, y_3)$, the thick line is $\phi(x, E)$. If $S = E$, then the υ and $\bar{x}$ given in the figure meet the requirement of lemma.

[20] In this book we call the set of sets to be class in order to distinguish the set. For example the power set can call as the class of subsets.

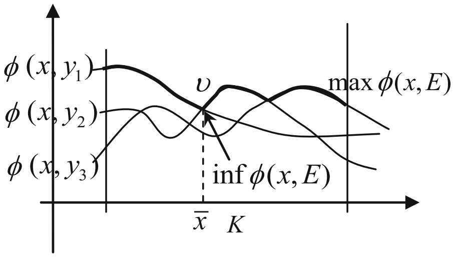


Fig. 2.16 Schematic diagram of inf max $\phi(x, y_i)$, $(\overline{x}, \upsilon)$ holds the min-max property that in $\phi(x, E)$ it is minimum, in $\phi\ (\overline{x}, S)$ is maximum

Lemma 2.6.4 Suppose that K is a convex and compact set in X. If $\phi : X \times X \to \mathbb{R}$ satisfies that:

(1) For every $y \in K$, the functional $\phi(\cdot, y)$ is lower semi-continuous;
(2) For every $x \in K$, the functional $\phi(x, \cdot)$ is a concave functional (i.e., $-\phi\ (x, \cdot\ )$ is a convex functional);
(3) For every $y \in K$, $\phi\ (y, y) \leq 0$.

Then there is a $\overline{x} \in K$ c $\phi\ (\overline{x}, y) \leq 0$ holds for every $y \in K$. □

Lemma 2.6.4 is known as Fan Ky inequality. To prove this inequality, we should apply the unit decomposition and fixed point theorem. This book does not try to prove the inequality, readers are referred to (Filippov 1988). It would like to point out the inequality holds in Banach space.

We are ready to prove Minty theorem. In the theorem I is used to denote the identical mapping in X. In order to make the proof convenience, we restrict ourselves in $\mathbb{R}^n$. A remark is given behind the proof to point out the difference if we deal with the theorem in a general Hilbert space.

Theorem 2.6.2 $F : \mathbb{R}^n \to \mathbb{R}^n$ is a monotonous set-valued mapping. Then the F is maximal if and only if $I + F$ is a surjective, i.e., the range of $I + F$ is $\mathbb{R}^n$.

Proof The proof for necessity consists of two steps.

(1) An alternative statement

The conclusion that $I + F$ is a surjective is equivalent to that for every $y \in \mathbb{R}^n$, there is an $\overline{x} \in \mathbb{R}^n$ such that $y \in \overline{x} + F\ (\overline{x})$. Let $A = -y + F$. Then the conclusion is further equivalent that $0 \in \overline{x} + A\ (\overline{x})$. By Problem 2 of this section, we conclude the F is maximal is equivalent to that A is maximal. Now, it is sufficient for a maximally monotonous set-valued mapping A to prove that there is an $\overline{x} \in \mathbb{R}^n$ such that $-\overline{x} \in A\ (\overline{x})$. Using Lemma 2.6.1, it is sufficient to verify $(\overline{x} - u, -\overline{x} - v) \geq 0$, or $(\overline{x} - u, \overline{x} + v) \leq 0$ for every $(u, v) \in$gra$A$.

Define a continuous functional $\phi(x, (u, v))$[21] as follows:

$$\phi\ (x, (u, v)) = \langle x - u, x + v \rangle = \|x\|^2 + \langle x, v - u \rangle - \langle u, v \rangle \tag{2.6.2}$$

[21] In $\phi(x, (u, v))$, v is not a dependent argument, it depends on u by $v \in A(u)$. Hence in $\phi(x, (u, v))$ the dependent variables are u and x, v can be treated as a parameter.

Then, we are only needed to prove there is an $\overline{x} \in X$, such that $\phi\left(\overline{x}, (u, v)\right) \leq 0$ for every $(u, v) \in$gra$A$.

(2) The existence of $\overline{x}$

We apply Lemma 2.6.4 to verify the existence of $\overline{x}$. Let $u_0 \in \mathbb{R}^n$ and $v_0 \in A\left(u_0\right)$. Then define a set $\Omega_{(u_0,v_0)} = \{x;\ \phi\left(x, (u_0, v_0)\right) \leq 0\}$. The set $\Omega_{(u_0,v_0)}$ holds the following properties.

(i) As a quadratic function, $\phi(x, (u_0, v_0))$ has its minimum, hence $\Omega_{(u_0,v_0)}$ is a closed set;

(ii) $\phi\left(u_0, (u_0, v_0)\right) = 0$, $u_0 \in \Omega_{(u_0,v_0)}$; and $\phi\left(x, (u_0, v_0)\right) \to +\infty \quad (x \to \infty)$, hence $\Omega_{(u_0,v_0)}$ is a nonempty and bounded set;

(iii) By the discussion in Sect. 1.3, $\phi(x, (u_0, v_0))$ is a convex function. Its epigraph is convex, so is the $\Omega_{(u_0,v_0)}$.

We now construct a compact set K. Let $u_i \in \mathbb{R}^n$, $i = 1, 2, \ldots, m$ be m vectors, and we take arbitrarily $v_i \in A\left(u_i\right)$. Denote $K = \overline{\text{co}}\{u_i;\ i = 1, 2, \ldots, m\}$.[22] Then K is compact and convex. Fix $x = x_0 \in K$ and prove that $\phi(x_0, (u, v))$ is concave. In order to use Theorem 1.3.3 (2), calculating

$$\nabla_{(u,v)}\left(-\phi\right) = \begin{pmatrix} -\dfrac{\partial \phi}{\partial u} & -\dfrac{\partial \phi}{\partial v} \end{pmatrix} = \left(\left(x^T + v^T\right) \quad \left(-x^T + u^T\right)\right)$$

Hence

$$\begin{aligned} &\left\langle \left(u_2 - u_1\right), \nabla_u\left(-\phi\right)\left(u_2\right) - \nabla\left(-\phi\right)_u\left(u_1\right)\right\rangle + \left\langle \left(v_2 - v_1\right), \nabla_v\left(-\phi\right)\left(u_2\right) - \nabla_v\left(-\phi\right)\left(u_1\right)\right\rangle \\ &\quad = \left\langle \left(u_2 - u_1\right), \left(v_2 - v_1\right)\right\rangle + \left\langle \left(v_2 - v_1\right), \left(u_2 - u_1\right)\right\rangle \\ &\quad \geq 0 \ . \end{aligned} \tag{2.6.3}$$

The last inequality comes from the fact of monotonicity.

To apply Lemma 2.6.4, we are lastly required to verify $\phi\left(u, (u, v)\right) \leq 0$. We now use Lemma 2.6.3 to verify the conclusion. Denote $M = \left\{\left(u_j, v_j\right), j = 1, 2, \ldots, m\right\}$, M is a finite set contained by gra A. By Lemma 2.6.3, we have

$$\begin{aligned} \sup_{(u,v)\in \text{gra } A} \phi\left(\overline{x}, (u, v)\right) &\leq \sup_{M \subset \text{gra} A} \inf_{x \in K} \max_{(u,v)\in M} \phi\left(x, (u, v)\right) \\ &\leq \sup_{M \subset \text{gra} A} \inf_{x \in K} \max_{\lambda \in \Lambda} \sum_j \lambda_j \phi\left(x, \left(u_j, v_j\right)\right), \end{aligned} \tag{2.6.4}$$

[22] Without loss of generality, we can select $u_i, i = 1, 2, \ldots, n$ such that they are linear independent. Therefore, Int $K \neq \varnothing$.

where $\Lambda = \left\{ \lambda = (\lambda_1\ \lambda_2 \ldots\ \lambda_m)^T; \lambda_i \geq 0, \sum_{i=1}^{m} \lambda_i = 1 \right\} \subset \mathbb{R}^m$. Inequality (2.6.4) is valid since

$$\phi\left(x, (u_i, v_i)\right) \in \left\{ \sum_j \lambda_j \phi\left(x, \left(u_j, v_j\right)\right), \lambda \in \Lambda \right\}$$

In Inequality (2.6.4), $x \in K$, hence $x = \sum_{k=1}^{m} \mu_k u_k$, with $\mu = (\mu_1\ \mu_2 \ldots \mu_m)^T \in \Lambda$. Thus,

$$\sum_j \lambda_j \phi\left(x, \left(u_j, v_j\right)\right) = \sum_j \lambda_j \phi\left(\sum_k \mu_k u_k, \left(u_j, v_j\right)\right)$$

For simplicity, we denote $\phi_M(\mu, \lambda) = \sum_j \lambda_j \phi\left(\sum_k \mu_k u_k, \left(u_j, v_j\right)\right)$. Using the notation, $\phi\left(u, (u, v)\right) \leq 0$ becomes $\phi_M(\mu, \mu) \leq 0$.

Because

$$\begin{aligned}
\phi\left(\sum_k \mu_k u_k, \left(u_j, v_j\right)\right) &= \left\langle \sum_k \mu_k u_k - u_j, \sum_k \mu_k u_k + v_j \right\rangle \\
&= \sum_k \mu_k \left\langle u_k - u_j, \sum_k \mu_k u_k + v_j \right\rangle \\
&= \sum_k \mu_k \left(\left\langle u_k - u_j, v_j \right\rangle + \left\langle u_k - u_j, \sum_k \mu_k u_k \right\rangle \right),
\end{aligned}$$

it follows

$$\begin{aligned}
\phi_M(\mu, \lambda) &= \sum_j \lambda_j \phi\left(\sum_k \mu_k u_k, \left(u_j, v_j\right)\right) \\
&= \sum_j \lambda_j \sum_k \mu_k \left\langle u_k - u_j, v_j \right\rangle + \sum_j \lambda_j \sum_k \mu_k \left\langle u_k - u_j, \sum_k \mu_k u_k \right\rangle.
\end{aligned}$$

Let $\lambda = \mu$. Then the second term of the above equation

$$\sum_{j,k=1}^{m} \mu_j \mu_k \left\langle u_k - u_j, \sum_k \mu_k u_k \right\rangle = 0$$

since both $\mu_j\mu_k\left\langle u_k - u_j, \sum_k \mu_k u_k\right\rangle$ and $\mu_k\mu_j\left\langle u_j - u_k, \sum_k \mu_k u_k\right\rangle$ appear in the summation, and they hold identical absolute values but opposite signs. We now turn to the first term

$$\begin{aligned}\sum_{j,k=1}^{m} \mu_j\mu_k\left\langle u_k - u_j, v_j\right\rangle &= \frac{1}{2}\sum_{j,k=1}^{m} \mu_j\mu_k\left\langle u_k - u_j, v_j\right\rangle + \frac{1}{2}\sum_{j,k=1}^{m} \mu_j\mu_k\left\langle u_k - u_j, v_j\right\rangle \\ &= \frac{1}{2}\sum_{j,k=1}^{m} \mu_j\mu_k\left\langle u_k - u_j, v_j\right\rangle + \frac{1}{2}\sum_{j,k=1}^{m} \mu_k\mu_j\left\langle u_j - u_k, v_k\right\rangle \\ &= \frac{1}{2}\sum_{j,k=1}^{m} \mu_j\mu_k\left\langle u_k - u_j, v_j - v_k\right\rangle \le 0.\end{aligned}$$

The last inequality comes from the monotonicity. Therefore, $\phi_M(\mu,\mu) \le 0$. By Lemma 2.6.4, there is a $\overline{\mu} \in \Lambda$ such that for every $\lambda \in \Lambda\phi_M(\overline{\mu},\lambda) \le 0$.

Furthermore,

$$\inf_{\mu\in\Lambda}\max_{\lambda\in\Lambda}\sum_j \phi_M(\mu,\lambda) \le \max_{\lambda\in\Lambda}\sum_j \phi_M(\overline{\mu},\lambda) \le 0$$

consequently, $\phi(\overline{x},(u,v)) \le 0$ for every $(u,v) \in \operatorname{gra} A$.

Sufficiency. Consider two arbitrary vectors $x,\ y \in \mathbb{R}^n$. Because $I+F$ is surjective, there is a $x_0 \in \mathbb{R}^n$ such that $y + x \in x_0 + F(x_0)$. Hence, there is a $y_0 \in F(x_0)$, such that $y + x = x_0 + y_0$. Thus, we have

$$\|y - y_0\|^2 = \langle y - y_0, y - y_0\rangle = -\langle y - y_0, x - x_0\rangle \tag{2.6.5}$$

If $\langle y - y_0,\ x - x_0\rangle \ge 0$, then Eq. (2.6.5) leads to $\|y - y_0\| \le 0$. We obtain $y = y_0$. By using the equation $y + x = x_0 + y_0$, we have $x = x_0$. Therefore $y \in F(x)$. By Lemma 2.6.2, F is maximal. □

Remark If X is a Hilbert space, the proof of Theorem 2.6.2 is similar except the concavity of $\phi(x_0,(u,v))$. In $\mathbb{R}^n$, we can apply partial derivative to verify the concavity of $\phi(x_0,(u,v))$. However, in a Hilbert space we have no such an operation. This conclusion is proved by using weak topology in Hilbert space. To complete such a proof, we need introduce some concepts which are beyond the target of this book. □

2.6.3 Yosida Approximation

In this subsection, we try to construct a series of single-valued mappings to approximate a maximally monotonous set-valued mapping.

Let A be a maximally monotonous set-valued mapping, and let $\lambda \in \mathbb{R}\,(>0)$, then by Minty theorem (Theorem 2.6.2) $I + \lambda A$ is invertible, i.e., for every $x \in X$, $(I + \lambda A)^{-1}(x) \neq \varnothing$.

Definition 2.6.2 Suppose A is a maximally monotonous set-valued mapping, $\lambda \in \mathbb{R}\,(>0)$, then $J_\lambda = (I + \lambda A)^{-1}$ is the resolvent of $I +\lambda A$. $A_\lambda = \frac{1}{\lambda}\,(I - J_\lambda)$[23] is the Yosida $\lambda-$ approximation of A. □

The following theorem presents the fundamental property of J_λ.

Theorem 2.6.3 Suppose $A : X \to X$ is a maximally monotonous set-valued mapping and $\lambda \in \mathbb{R}\,(>0)$, then the resolvent $J_\lambda = (I + \lambda A)^{-1}$ is a single-valued mapping. It is Lipschitzian mapping and its Lipschitzian constant can be 1.

Proof Let $y \in X$. Then there is an $x \in X$ such that $y \in x + \lambda A(x)$ i.e., $x \in J_\lambda(y)$. The effective domain of J_λ is nonempty for every $\lambda \in \mathbb{R}\,(>0)$.

Now let $y_1, y_2 \in X$. Then there are $x_1, x_2 \in X$ such that $y_i \in x_i + \lambda A\,(x_i)$, $i = 1, 2$, i.e., there are $v_i \in A\,(x_i)$, $i = 1, 2$ such that $y_i = x_i + \lambda v_i$. Hence,

$$y_1 - y_2 = x_1 - x_2 + \lambda\,(v_1 - v_2)$$

It follows

$$\begin{aligned}\|y_1 - y_2\|^2 = \|x_1 - x_2\|^2 + \lambda^2\|v_1 - v_2\|^2 + 2\lambda\,\langle x_1 - x_2, v_1 - v_2\rangle \geq \|x_1 - x_2\|^2 \\ + \lambda^2\|v_1 - v_2\|^2\end{aligned} \tag{2.6.6}$$

Thus, $\|x_1 - x_2\| \leq \|y_1 - y_2\|$. If $y_1 = y_2$, then $x_1 = x_2$. It implies that J_λ is a single-valued mapping. It is Lipschitzian and its Lipschitzian constant can be 1. □

In terminology, if Lipschitzian constant is 1, then the mapping is said to be nonexpansive, i.e., the distance of two images is less than or equal to that of arguments. If it is less than 1, the mapping is a compressed mapping or contraction.

By the relation between J_λ and A_λ, the following corollary can be verified.

Corollary 2.6.1 A_λ is a Yosida approximation of A. Then A_λ holds the following properties.

(1) A_λ is a single-valued mapping;
(2) A_λ is a Lipschitzian mapping and its Lipschitzian constant can be $\frac{1}{\lambda}$
(3) For every $x \in X$, $A_\lambda(x) \in A\,(J_\lambda(x))$;
(4) A_λ is a maximally monotonous mapping.

[23] In order to distinguish the inverse of a mapping, we apply $\frac{a}{b}$ to express a fraction.

Proof (1) Because $A_\lambda = \frac{1}{\lambda}(I - J_\lambda)$, J_λ is single-valued mapping, so is A_λ.

(2) Applying the notations used in the proof of Theorem 2.6.3, for every $y \in X$, there are $x = J_\lambda(y)$, $v \in A(x)$ such that $v = \frac{1}{\lambda}(y - x)$. Then

$$A_\lambda(y) = \frac{1}{\lambda}(I - J_\lambda)\, y = \frac{1}{\lambda}(y - x) = v$$

Now let $v_i = A_\lambda\,(y_i)$, $i = 1{,}2$. By Inequality (2.6.6), we have

$$\|y_1 - y_2\|^2 \geq \lambda^2 \|v_1 - v_2\|^2 = \lambda^2 \|A_\lambda\,(y_1) - A_\lambda\,(y_2)\|^2$$

Therefore, the Lipschitzian constant of A_λ can be $\frac{1}{\lambda}$.

(3) Because $(I + \lambda A)\,(I + \lambda A)^{-1} = (I + \lambda A)\,J_\lambda$, for every $x \in X$, we have $x \in (I + \lambda A)\,J_\lambda(x) = J_\lambda(x) + \lambda A\,(J_\lambda(x))$. By the definition, $\lambda A_\lambda + J_\lambda = I$, i.e., $x = J_\lambda(x) + \lambda A_\lambda(x)$. Comparing these two results, it is clear $A_\lambda(x) \in A\,(J_\lambda(x))$.

(4) We only prove that A_λ is monotonous, then by Problem 1 of this section, it is maximal. Consider now

$\langle A_\lambda\,(y_1) - A_\lambda\,(y_2)\,, y_1 - y_2\rangle = \langle A_\lambda\,(y_1) - A_\lambda\,(y_2)\,, J_\lambda\,(y_1) - J_\lambda\,(y_2)\rangle + \lambda \|A_\lambda\,(y_1) - A_\lambda\,(y_2)\|^2$ where we have applied $y = \lambda A_\lambda(y) + J_\lambda(y)$. From conclusion (3) above, $A_\lambda(x) \in A\,(J_\lambda(x))$. A is monotonous, hence the first term is nonnegative. Thus, $\langle A_\lambda\,(y_1) - A_\lambda\,(y_2)\,, y_1 - y_2\rangle \geq 0$ for $y_1, y_2 \in X$. A_λ is monotonous by Definition 2.6.1. □

The further discussion needs the concept $m(A(x))$. Recall that $m(A(x))$ is the minimal norm element in $A(x)$. If $A(x)$ is convex, $m(A(x))$ is the unique projection of the origin on $A(x)$. The following theorem presents the relations between A_λ and $m(A(x))$.

Theorem 2.6.4 Suppose A is a maximally monotonous set-valued mapping. Then for every $x \in X$ the following conclusions hold.

(1) $\|m\,(A(x)) - A_\lambda(x)\|^2 \leq \|m\,(A(x))\|^2 - \|A_\lambda(x)\|^2$

(2) $\lim\limits_{\lambda \to 0} J_\lambda(x) = x$

(3) $\lim\limits_{\lambda \to 0} A_\lambda(x) = m\,(A(x))$

Proof (1) Because

$$\begin{aligned}\|m\,(A(x)) - A_\lambda(x)\|^2 &= \|m\,(A(x))\|^2 + \|A_\lambda(x)\|^2 - 2\,\langle A_\lambda(x), m\,(A(x))\rangle \\ &= \|m\,(A(x))\|^2 - \|A_\lambda(x)\|^2 - 2\,\langle A_\lambda(x), m\,(A(x)) - A_\lambda(x)\rangle\,,\end{aligned}$$

A is monotonous, and $m\,(A(x)) \in A(x)$, $A_\lambda(x) \in A\,(J_\lambda(x))$, hence

$$\langle A_\lambda(x), m\,(A(x)) - A_\lambda(x)\rangle = \frac{1}{\lambda}\,\langle x - J_\lambda(x), m\,(A(x)) - A_\lambda(x)\rangle \geq 0$$

The conclusion follows directly.

(2) By conclusion (1) above

$$\|x - J_\lambda(x)\| = \lambda \,\|A_\lambda(x)\| \le \lambda \,\|m\,(A(x))\|$$

It leads to

$$\lim_{\lambda \to 0} J_\lambda(x) = x \tag{2.6.7}$$

(3) The proof of this conclusion seems to be complicated, and is divided into three parts.

(i) For every $x \in X$, the relation $y \in A\,(x - \lambda y)$ has a unique solution $y = A_\lambda(x)$.

Let $y = A_\lambda(x)$. Then $x - \lambda y = J_\lambda(x)$. It follows $A\,(x - \lambda y) = A\,(J_\lambda(x))$. By Corollary 2.6.1 (3), $y = A_\lambda(x) \in A\,(J_\lambda(x)) = A\,(x - \lambda y)$, i.e., $A_\lambda(x)$ is a solution of relation $y \in A\,(x - \lambda y)$.

On the other hand, if y_0 is a solution of $y \in A\,(x - \lambda y)$. Denote $z = x - \lambda y_0$, then $y_0 \in A(z)$, i.e., $x \in z + \lambda A(z)$. By the definition of J_λ, it follows $z = J_\lambda(x)$. $y_0 \in A(z) = A\,(J_\lambda(x)) = A_\lambda(x)$. A_λ is a single-valued mapping, hence $y_0 = A_\lambda(x)$.

(ii) If $x \in X$ is fixed, then $\|A_\lambda(x)\|$ is a monotonous function of λ.

By the conclusion (2), if $x \in X$ is given, then $y_0 = A_{\lambda+\mu}(x)$ is the unique solution of relation $y \in A\,(x - (\lambda + \mu)\,y)$. Let us rewrite $y_0 \in A\,(x - (\lambda + \mu)\,y_0) = A\,((x - \mu y_0) - \lambda y_0)$, hence y_0 is also the solution of $y \in A_\lambda\,(x - \mu y)$. By Corollary 2.6.1 (4), $A_\lambda(x)$ is maximally monotonous. If we repeat the process of (i), then we can obtain $y_0 = (A_\lambda)_\mu(x)$ is the unique solution of relation $y \in A_\lambda\,(x - \mu y)$. Consequently $y_0 = A_{\lambda+\mu}(x) = (A_\lambda)_\mu(x)$ for every $x \in X$.

By conclusion (1) of this theorem, $\|m\,(A(x)) - A_\lambda(x)\|^2 \le \|m\,(A(x))\|^2 - \|A_\lambda(x)\|^2$. Replacing $A(x)$ and $A_\lambda(x)$ by $A_\mu(x)$ and $A_{\lambda+\mu}(x)$, respectively, we obtain

$$\left\|A_\mu(x) - A_{\mu+\lambda}(x)\right\|^2 \le \left\|A_\mu(x)\right\|^2 - \left\|A_{\mu+\lambda}(x)\right\|^2 \tag{2.6.8}$$

where $m(A_\lambda(x))$ has been replaced by $A_\lambda(x)$. This is correct because $A_\lambda(x)$ is a single-valued mapping. Inequality (2.6.8) implies that the decreasing of λ leads to the monotonously increasing of $\|A_\lambda(x)\|^2$.

(iii) The convergence of $A_\lambda(x)$.

By conclusion (1) of this theorem, for every $x \in X$, $\{\|A_\lambda(x)\|\}$ holds an upper boundary $\|m(A(x))\|$. We now let $\lambda = \frac{1}{k}$, $k \in \mathbb{N}$, by Inequality (2.6.8), $\{A_\lambda(x)\}$ is a Cauchy series. Hence, there exists a $\upsilon_x \in X$ such that $\upsilon_x = \lim_{\lambda \to 0} A_\lambda(x)$. The υ_x has the following properties:

(i) $A_\lambda(x) \in A\,(J_\lambda(x))$ and $\lim_{\lambda \to 0} J_\lambda(x) = x$, $A(x)$ is closed (Theorem 2.6.1), hence $\upsilon_x \in A(x)$;

(ii) $\|A_\lambda(x)\| \le \|m\,(A(x))\|$ by conclusion (1), consequently $\|\upsilon_x\| \le \|m\,(A(x))\|$.

The two facts lead to $\upsilon_x = m\,(A(x))$. □

Actually, $A_\lambda(x)$ is a approximate selection of $A(x)$. Let us write

$$\begin{aligned}(x, A_\lambda(x)) &= (J_\lambda(x), A_\lambda(x)) + (x - J_\lambda(x), 0)\\ &\in (J_\lambda(x), A\,(J_\lambda(x))) + \|x - J_\lambda(x)\|\, B,\end{aligned}$$

In the proof of Theorem 2.6.4 (2), we have $\|x - J_\lambda(x)\| \leq \lambda\, \|m\,(A(x))\|$, hence

$$(x, A_\lambda(x)) \in \text{gra}\ A + \lambda\, \|mA(x)\|\, B$$

2.6.4 Maximally Monotonous Differential Inclusions

To end this section, we prove that the solution of a maximally monotonous differential inclusion is uniquely existed. The proof is completed by using Yosida approximation. We now give a lemma about weak convergence.

Lemma 2.6.5 Suppose X is a Hilbert space, and $x_n \in X,\ n = 1, 2, \ldots$. If x_n weakly converges to x_0, then

(1) $\|x_0\| \leq \lim\limits_{n\to\infty}\ \inf\ \|x_n\|$

(2) Additionally, if $\lim\limits_{n\to\infty} \|x_n\| = \|x_0\|$, then $\lim\limits_{n\to\infty} \|x_n - x_0\| = 0$, i.e., x_n strongly converges to x_0.

Proof (1) $\|x_0\|^2 = \lim\limits_{n\to\infty} \langle x_n, x_0\rangle = \lim\limits_{n\to\infty} \|x_0\|\, \|x_n\| \cos\theta_n$ where θ_n is the angle yielded by x_0 and x_n. Because $\|x_n\| \cos\theta_n \leq \|x_n\|$, the equation implies $\|x_0\| \leq \lim\limits_{n\to\infty}\ \inf\ \|x_n\|$.

(2) $\|x_n - x\|^2 = \langle x_n - x_0, x_n - x_0\rangle = \langle x_n, x_n - x_0\rangle - \langle x_0, x_n - x_0\rangle$. Its second term converges to zero by the weak convergence of x_n. Consider the first term

$$\langle x_n, x_n - x_0\rangle = \langle x_n, x_n\rangle - \langle x_n, x_0\rangle \to \|x_n\|^2 - \|x_0\|^2 = 0$$

The conclusion is verified. □

We now consider the following Cauchy problem

$$\dot{x}(t) = -A\,(x(t))\,,\quad x(0) = x_0 \tag{2.6.9}$$

where A is a maximally monotonous set-valued mapping with effective domain $\text{dom}(A) \subset X$, $x_0 \in \text{dom}(A)$.

Theorem 2.6.5 The solution of Cauchy problem (2.6.9) exists uniquely. Let $x(t, x_0)$ be the solution. Then it is also the solution of

$$\dot{x}(t) = -m\left(A\left(x(t)\right)\right), \quad x(0) = x_0 \tag{2.6.10}$$

for almost all $t \in [0, \infty)$.

Furthermore, the solution $x(t, x_0)$ holds the following properties

(1) $\left\|\dot{x}\left(t, x_0\right)\right\|$ is monotonously decreasing;
(2) The right derivative $\dot{x}^{+}\left(t, x_0\right) = \lim\limits_{h \downarrow 0} \frac{x(t+h,x_0)-x(t,x_0)}{h}$ is rightly continuous;
(3) Let $x_i\left(t, x_{i0}\right), \ i = 1, 2$ be two solutions with initial conditions x_{10} and x_{20}, respectively. Then

$$\left\|x_1\left(t, x_{10}\right) - x_2\left(t, x_{20}\right)\right\| \le \left\|x_{10} - x_{20}\right\| = \left\|x_1(0) - x_2(0)\right\| \tag{2.6.11}$$

for every $t \in [0, \infty)$.

Proof The proof of Theorem 2.6.5 consists of 6 parts.

(1) We start to prove the conclusion (3). For $t \in [0, \ \infty)$, we have

$$\frac{d}{dt}\frac{1}{2}\left\|x_1\left(t, x_{10}\right) - x_2\left(t, x_{20}\right)\right\|^2 = \left\langle \dot{x}_1\left(t, x_{10}\right) - \dot{x}_2\left(t, x_0\right), \ x_1\left(t, x_{10}\right) - x_2\left(t, x_{20}\right)\right\rangle \le 0$$

The last inequality holds because of monotonicity of A. Furthermore,

$$\begin{aligned} 0 &\ge \int_0^t \left\langle \dot{x}_1\left(\tau, x_{10}\right) - \dot{x}_2\left(\tau, x_{20}\right), \ x_1\left(\tau, x_{10}\right) - x_2\left(\tau, x_{20}\right)\right\rangle d\tau \\ &= \tfrac{1}{2}\left(\left\|x_1\left(t, x_{10}\right) - x_2\left(t, x_{20}\right)\right\|^2 - \left\|x_1(0) - x_2(0)\right\|^2\right) \ . \end{aligned}$$

Conclusion (3) is verified. It also implies the uniqueness of the solution.

(2) Denote $y(t) = x\left(t + h\right)$ where $x(t)$ is the solution of Cauchy problem (2.6.9), and consider the following Cauchy problem

$$\dot{y}(t) = -A\left(y(t)\right), \quad y(0) = x(h).$$

By Inequality (2.6.10), we obtain

$$\left\|\frac{y(t) - x(t)}{h}\right\| \le \left\|\frac{y(0) - x(0)}{h}\right\|.$$

It is equivalent to

$$\left\|\frac{x\left(t + h\right) - x(t)}{h}\right\| \le \left\|\frac{x(h) - x(0)}{h}\right\|,$$

and then let $h \to 0$. When their derivatives all exist, we have $\left\|\dot{x}(t)\right\| \le \left\|\dot{x}(0)\right\|$. If we replace t and 0 by $t_0 + h$ and t_0, then the above discussion leads to $\left\|\dot{x}\left(t_0 + h\right)\right\| \le \left\|\dot{x}\left(t_0\right)\right\|$. The conclusion (1) is verified.

(3) We now prove the existence of Inc. (2.6.8). A is maximally monotonous, hence $A(x)$ is with closed and convex value for every $x \in \operatorname{dom} A$ (Theorem 2.6.1). Consider Cauchy problem of the following differential equation

$$\dot{x}_\lambda(t) = -A_\lambda\left(x_\lambda(t)\right),\ x_\lambda(0) = x_0 \tag{2.6.12}$$

We have proved $-A_\lambda$ is a Lipschitzian mapping (Corollary 2.6.1 (2)), the above Cauchy problem exists unique solution. By the first conclusion of this theorem, $\left\|\dot{x}_\lambda(t)\right\|$ is monotonously decreasing by the t increasing. Hence,

$$\left\|\dot{x}_\lambda(t)\right\| \le \left\|\dot{x}_\lambda(0)\right\| = \left\|A_\lambda\left(x_0\right)\right\| \le \left\|m\left(A\left(x_0\right)\right)\right\| \tag{2.6.13}$$

the last inequality comes from the proof of Theorem 2.6.4 (3). $\{x_\lambda(t)\}$ is a set of equicontinuous functions for all $\lambda > 0$. Consequently, series $\left\{x_{\frac{1}{k}}(t); k \in \mathbb{N}\right\}$ holds a convergent subsequence. Without loss of generality, we assume $\lim\limits_{k\to\infty} x_{\frac{1}{k}}(t) = x(t)$ for $x(t) \in AC\left([0, \infty), \mathbb{R}^n\right)$ uniformly. Because $\left\{\dot{x}_\lambda(t)\right\}$ is contained in a bounded set, by Theorem 1.1.7, we can assume $\dot{x}_{\frac{1}{k}}(t) \to \dot{x}(t)(k \to \infty)$ weakly.

Define $\mathcal{A} : x_\lambda(t) \to -A_\lambda\left(x_\lambda(t)\right)$, A is an operator form $L^2[0, T]$ to $L^2[0, T]$. We can prove A is maximally monotonous.[24] Because $x_{\frac{1}{k}}(t) \to x(t)$ strongly and $A_{\frac{1}{k}}\left(x_{\frac{1}{k}}(t)\right) = \dot{x}_{\frac{1}{k}}(t) \to \dot{x}(t)$ weakly, by Theorem 2.6.1 (2), $\dot{x}(t) \in -A\left(x\left(t\right)\right) = -A\left(x\left(t\right)\right)$. The existence is verified.

(4) $\left\|m\Big(A\left(x(t)\right)\Big)\right\|$ is monotonously decrease as t increases.

At the Part (3) of this proof, we have proved that $\{A_\lambda(x_\lambda(t))\}$ (or $\left\{A_{\frac{1}{k}}\left(x_{\frac{1}{k}}(t)\right)\right\}$) weakly converges to $\dot{x}(t)$ as $\lambda \to 0$ $(k \to \infty)$. For convenience, we denote $\upsilon(t) = \dot{x}(t)$, then $\upsilon(t) \in A\left(x(t)\right)$. It follows $\|\upsilon(t)\| \ge \left\|m\left(A\left(x(t)\right)\right)\right\|$. On other hand, $\{A_\lambda(x_\lambda(t))\}$ is weakly convergent to $\upsilon(t)$, hence, by Lemma 2.6.5,

$$\left\|m\Big(A\left(x(t)\right)\Big)\right\| \le \|\upsilon(t)\| \le \liminf_{\lambda\to 0}\left\|A_\lambda\left(x_\lambda(t)\right)\right\| \le \left\|mA\left(x(0)\right)\right\| \tag{2.6.14}$$

the last inequality is obtained from (2.6.12). Consequently, $\left\|m\left(A\left(x(0)\right)\right)\right\| \ge \left\|m\left(A\left(x(t)\right)\right)\right\|$. The conclusion implies that $\left\|m\Big(A\left(x(t)\right)\Big)\right\|$ is monotonously decrease.

(5) $\left\|m\Big(A\left(x(t)\right)\Big)\right\|$ is continuous form right.

[24] Note A:$X \to X$, but the operator A: is $L^2[0, T] \to L^2[0, T]$. The norm and inner product considered are in space $L^2[0, T]$. It is direct to show the maximal monotonicity of A implies that of A:.

We now prove for any sequence $\{t_n\}$, $t_n \downarrow t_0$, then $m(A(x(t_n))) \to m(A(x(t_0)))$. By Part (4) of the proof, the series $\{m(A(x(t_n)))\}$ is monotonously increasing sine $\{t_n\}$ decreases. Furthermore, by Inequality (2.6.13),

$$\|m(A(x(t_n)))\| \le \|v(t_n)\| \le \|m(A(x(t_0)))\|$$

There is a subsequence of $m(A(x(t_n)))$ and the subsequence is weakly convergent to a vector v_0. We can assume that $m(A(x(t_n)))$ is weakly convergent to v_0. Repeating the proof of Inequality (2.6.12), inequalities can be obtained as follows

$$\|m(A(x_0))\| \le \|v_0\| \le \lim_{n\to\infty} \inf \left\| m\Big(A(x(t_n))\Big) \right\| \le \|m(A(x(t_0)))\| = \|m(A(x_0))\|$$

By Lemma 2.6.5, $m\Big(A(x(t_n))$ converges strongly to $m(A(x_0))$.

(6) At last, the solution of Inc. (2.6.8) is identical to the solution of Eq. (2.6.9)

For almost all $t \in [0, T]$ and $h > 0$, we have

$$\|x(t+h) - x(t)\| = \left\| \int_t^{t+h} \dot{x}(\tau)\, d\tau \right\| \le \int_t^{t+h} \left\|\dot{x}(\tau)\right\| d\tau \le h \|m(A(x(t)))\|,$$

the last inequality is obtained from the fact the $\|m(A(x(t)))\|$ decreases monotonously. $\left\|\dot{x}(t)\right\|$ is continuous from right, when $h \downarrow 0$, $\left\| \frac{x(t+h)-x(t)}{h} \right\| \to \left\|\dot{x}(t)\right\|$. The above inequality implies $\left\|\dot{x}(t)\right\| \le \|m(A(x(t)))\|$. Because $\dot{x}(t) \in -A(x(t))$ and $A(x)$ is closed, $\dot{x}(t) = -m(A(x(t)))$.

Thus, we complete proof for all conclusions of the theorem. □

Note that in the theorems given above, the norms $\|x(t)\|$, $\left\|\dot{x}(t)\right\|$ and others are functions of time t. The $x(t)$ is an trajectory in space X.

The solution of Inc. (2.6.8) which satisfies equation $\dot{x}(t) = -m(A(x(t)))$ is called slow solution or mild solution.

Problems

1. $f : \mathbb{R}^n \to \mathbb{R}^n$ is a monotonously single-valued mapping, if f is continuous then it is maximal.
2. $F : \mathbb{R}^n \to \mathbb{R}^n$ is maximally monotonous, prove the following conclusions:

   (1) $y_0 \in \mathbb{R}^n$ is a given vector, then $y_0 + F$ is maximally monotonous.
   (2) $\alpha > 0$ is a positive real number, then αF is maximally monotonous.

3. $F : \mathbb{R}^n \to \mathbb{R}^n$ is a monotonous mapping. Prove that F is monotonous if and only if the inverse $F^{-1} : \mathbb{R}^n \to \mathbb{R}^n$, which may be a set-valued mapping, is monotonous.
4. If $F : \mathbb{R}^n \to \mathbb{R}^n$ is a convex mapping then its subdifferential $\partial F(x)$ is monotonous.

5. Assume, in Theorem 2.6.2, $X \; = \mathbb{R}^n$, prove that the set $\{x;\ \phi(x,(u,v)) \le 0\}$ defined in the proof of Theorem 2.6.2 is compact and convex.
6. Let $x_\lambda(t)$ be the solution of Eq. (2.6.11). Then $J_\lambda(x_\lambda(t))$ converges to $x(t)$ uniformly in the proof of Theorem 2.6.5.
7. If $V : \mathbb{R}^n \to \mathbb{R}\,(\infty)$ is a lower semi-continuous and convex function, then $\partial V(x)$ is maximally monotonous. Moreover, the solution of Cauchy problem $\dot{x}(t) = -\partial V(x), x(0) = x_0$ satisfies the equation $\frac{d}{dt}V(x(t)) + \left\|\dot{x}(t)\right\|^2 = 0$.

References

Aubin J-P, Cellina A (1984) Differential inclusions – set-valued maps and viability theory [M]. Springer, Berlin

Cai X, Liu L, Zhang W (2009) Saturated control design for linear differential inclusions subject to disturbance. Nonlinear Dyn 58(3):487–496

Filippov AF (1988) Differential equations with discontinuous right-hand sides [M]. Kluwer Academic Publishers, Dordrecht

Sontag ED (1989) Smooth stabilization implies coprime factorization [J]. IEEE Trans Autom Control 34(4):435–442

Chapter 3
Convex Processes

Starting from this chapter, we deal with several kinds of differential inclusions and their control. We recall that mathematical investigation in many fields starts from the linear case; for example, at the beginning of control theory, we deal with the linear system, the functional analysis starts with the linear normed spaces and linear mappings. Consequently, we consider convex processes which can be treated as an extension of the linear single-valued mapping to set-valued mapping. The next two chapters will consider linear polytopic differential inclusions and the Luré systems. They can also be thought as extensions of linear control systems to differential inclusions and nonlinear differential inclusions.

The organization of this chapter is as follows. At the first section, we define the convex processes in Banach space and verify that the convex processes hold the same properties as linear mapping in Banach space. The second section deals with the convex processes in finite dimensional spaces and shows that they have similar construction as matrices. The third section considers the differential inclusion yielded by convex processes. The controllability is discussed. And the last section investigates the stability of convex process differential inclusions.

3.1 Convex Processes in Linear Normed Spaces

This section deals with convex processes in linear normed spaces, its target is to extend the fundamental theorems of linear single-valued mappings in Banach space to convex processes.

Z. Han et al., *Theory of Control Systems Described by Differential Inclusions*,
Springer Tracts in Mechanical Engineering, DOI 10.1007/978-3-662-49245-1_3

3.1.1 Convex Processes and Their Adjoint Processes

In order to illustrate that the convex process is an extension of linear single-valued mapping, we recall fundamental properties of linear single-valued bounded mapping.

Let X and Y be two normed spaces, $A : X \to Y$ is a linear (single-valued) mapping if it satisfies that (1) for $x_1, x_2 \in X$, $A(x_1 + x_2) = A(x_1) + A(x_2)$; (2) for $x \in X$ and $\alpha \in \mathbb{R}$, $A(\alpha x) = \alpha A(x)$. A linear mapping is bounded if there is a $M \in \mathbb{R}$ which is independent of x, such that $\|Ax\|_Y \leq M\|x\|_X$ for every $x \in X$. All linear bounded mappings consist of a linear normed space $\mathfrak{L}(X, Y)$, the norm in $\mathfrak{L}(X, Y)$ is defined as

$$\|A\| = \sup_{x \neq 0} \frac{\|A(x)\|_Y}{\|x\|_X}.$$

Especially, when $Y = \mathbb{R}$, $\mathfrak{L}(X, \mathbb{R})$ is called conjugate space of X and denoted by X^*. X^* is always complete no matter whether X is complete.

If $f \in X^*$ and $x \in X$, then $f(x)$ is a real number. Sometimes, we write $f(x)$ by $\langle f, x\rangle$. Because $f(x)$ is linear for both f and x, the notation $\langle f, x\rangle$ meets with the requirement of inner product. Moreover, it may provide lots convenience.

Let $A \in \mathfrak{L}(X, Y)$, then the A can induce a mapping $A^* : Y^* \to X^*$ by the definition that $(A^*f)(x) = f(Ax)$ for every $f \in Y^*$ and $x \in X$. Using the notation that $f(x) = \langle f, x\rangle$, we have $\langle f, Ax\rangle = \langle A^*f, x\rangle$ which is quite simple. A^* is single-valued and called as the adjoint mapping of A.

If $A \in \mathfrak{L}(X, Y)$, then for $x_1, x_2 \in X$ and $\lambda_1, \lambda_2 \in \mathbb{R}$, we have $A(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 A(x_1) + \lambda_2 A(x_2)$. It implies gra$A$ is a convex cone. The property is applied to define the convex process.

Definition 3.1.1 Let X and Y be two normed spaces, and $A : X \to Y$ be a set-valued mapping. If gra$A$ is convex cone, then A is a convex process. Moreover, if gra$A$ is a closed convex cone, then A is a closed convex process.

A convex process is strict if its effective domain is X, i.e., dom $A = X$. □

In history, economist and mathematician Rockafellar appeared that a kind of economic phenomena hold convex characteristics. It is common that one person had invested 100 dollars, he received 120 return on investment. But if he invested 1000 dollars his return was not sure to be 1200 dollars. It may be in the interval (900, 1300). Hence the economic process is more suitable to be described by a set-valued mapping which holds convex cone characteristics. Rockafellar called the economic phenomena as convex process. After that, he investigated the convex processes and found that convex process can be treated as an extension of linear mapping. At the same time, some researchers considered another extension of linear mapping. They required that for every $x \in$ dom A, $A(x)$ is a linear subspace of Y and called it as linear set-valued mapping, or linear mapping for simplicity. They called the convex process by “sublinear” set-valued mapping. But the custom is not popular at present.

Because gra$A$ is a convex cone, for $x_1, x_2 \in \operatorname{dom} A$ and $y_1 \in A(x_1)$, $y_2 \in A(x_2)$, we have $\lambda(x_1, y_1) + (1-\lambda)(x_2, y_2) \in \operatorname{gra} A$ for $\lambda \in [0,1]$. It follows that $\lambda x_1 + (1-\lambda)x_2 \in \operatorname{dom} A$ and $\lambda y_1 + (1-\lambda)y_2 \in A(\lambda x_1 + (1-\lambda)x_2)$, i.e., both domain and range of a convex process are convex. Furthermore, $A(x)$ is convex for every $x \in \operatorname{dom} A$.

The following discussion further shows that the convex process holds linear-like property.

Lemma 3.1.1 $A : X \to Y$ is a set-valued mapping. A is a convex process if and only if the following two conditions are satisfied simultaneously.

(1) If $x \in \operatorname{dom} A$, then $\lambda x \in \operatorname{dom} A$ and $A(\lambda x) = \lambda A(x)$ where $\lambda > 0$;
(2) If $x_1, x_2 \in \operatorname{dom} A$, then $x_1 + x_2 \in \operatorname{dom} A$ and $A(x_1) + A(x_2) \subset A(x_1 + x_2)$. □

The proof of Lemma is left to readers as an exercise.

The two conditions given in Lemma 3.1.1 can be wrapped up that: if $x_1, x_2 \in \operatorname{dom} A$ and $\lambda, \mu \in \mathbb{R}(\geq 0)$, then $\lambda x_1 + \mu x_2 \in \operatorname{dom} A$ and

$$\lambda A(x_1) + \mu A(x_2) \subset A(\lambda x_1 + \mu x_2). \tag{3.1.1}$$

Lemma 3.1.2 Let $A : X \to Y$ be a strictly convex process. If there exists an $x_1 \in X$ such that $A(x_1)$ holds only one element, then A is a single-valued linear mapping.

Proof Because A is a convex process, we have $0 \in A(0)$. Then

$$A(x_1) \subset A(x_1) + A(0) \subset A(x_1 + 0) = A(x_1).$$

We obtain $A(x_1) = A(x_1) + A(0)$. It implies that $A(0)$ has only one element, i.e., $A(0) = \{0\}$.

$A(x) + A(-x) \subset A(0) = \{0\}$. It implies that $A(x)$ has only one element, so is $A(-x)$. More $A(x) + A(-x) = \{0\}$ implies $A(-x) = -A(x)$.

Let $x \in X$. Then, by the relation (3.1.1), we have

$$\lambda A(x) = \lambda A(x) + A(0) \subset A(\lambda x + 0) = A(\lambda x)$$

for $\lambda \in \mathbb{R}(\geq 0)$. Both $A(x)$ and $A(\lambda x)$ have one element, respectively. Hence the “$\subset$” can be replaced by “$=$”. Combining with $A(-x) = -A(x)$, we conclude the homogeneity holds.

$A(x_2), A(x_3), A(x_2 + x_3)$ have one element, respectively. Thus, $A(x_2) + A(x_3) \subset A(x_2 + x_3)$ implies $A(x_2) + A(x_3) = A(x_2 + x_3)$. The additivity is verified. Therefore, A is a linear single-valued mapping. □

Lemma 3.1.2 illustrates that if a convex process degenerates to a single-valued mapping then A is a linear mapping. The explanation supports that the convex process is an extension of linear mapping in theory of set-valued mappings.

Corollary 3.1.1 Let $A : X \to Y$ be a convex process and $A(0) = \{0\}$. But A may not be strict.

(1) If $x, -x \in \operatorname{dom} A$, then $A(-x) = -A(x)$ and $A(-x), A(x)$ are all single-valued;
(2) $X_1 \subset X$ is a subspace, and $X_1 \subset \operatorname{dom} A$. If $x_1 \in X_1, x, x + x_1 \in \operatorname{dom} A$, then $A(x + x_1) = A(x) + A(x_1)$.

Proof (1) Because $0 \in \operatorname{dom} A, A(x_1) + A(-x_1) \subset A(0) = \{0\}$. It implies both $A(x)$ and $A(-x)$ have only one element, respectively, and $A(-x_1) = -A(x_1)$.

(2) X_1 is a subspace and $X_1 \subset \operatorname{dom} A$. By (1), restricting on X_1 A is a single-valued linear mapping. If $x \notin X_1$, $A(x)$ may not be single-valued. But we have

$$A(x) = A(x) + A(x_1) - A(x_1) \subset A(x + x_1) - A(x_1) = A(x + x_1) + A(-x_1) \subset A(x).$$

Hence, $A(x) = A(x + x_1) - A(x_1)$, i.e., $A(x) + A(x_1) = A(x + x_1)$. □

We now define adjoint process for convex process. Readers may find that it is really an extension of adjoint mapping for linear mapping.

Definition 3.1.2 Let $A : X \to Y$ be a convex process. Define a set-valued mapping $A^* : Y^* \to X^*$ as follows: If $y^* \in Y^*$ then

$$A^*\left(y^*\right) = \{x^*; \langle x^*, x\rangle \le \langle y^*, y\rangle, \quad \text{for} \quad (x, y) \in \operatorname{gra} A\}.$$

A^* is called by the adjoint process of A. □

We now give some fundamental properties for the adjoint mapping.

Lemma 3.1.3 Suppose $A : X \to Y$ is a closed and convex process. Then the following statements are valid.

(1) $(y^*, x^*) \in \operatorname{gra} A^*$ if and only if $(-x^*, y^*) \in (\operatorname{gra} A)^*$
(2) $(-A)^*(y^*) = A^*(-y^*)$
(3) $\left(A^{-1}\right)^*(x^*) = -(A^*)^{-1}(-x^*)$
(4) $A^{**}(x) = -A(-x)$
(5) $A(0) = (\operatorname{dom} A^*)^*$
(6) $A^{-1}(0) = (\operatorname{Im}(-A^*))^*$

Proof These conclusions can be verified from Definition 3.1.2.

(1) If $(y^*, x^*) \in \operatorname{gra} A^*$, i.e., $x^* \in A^*(y^*)$. Definition 3.1.2 illustrates $\langle x^*, x\rangle \le \langle y^*, y\rangle$ for every pair $(x, y) \in \operatorname{gra} A$. The inequality is equivalent to $\langle y^*, y\rangle + \langle -x^*, x\rangle \ge 0$, i.e., $\langle(-x^*, y^*), (x, y)\rangle \ge 0$, or, $(-x^*, y^*) \in (\operatorname{gra} A)^*$. If we reverse the process, we can obtain $(y^*, x^*) \in \operatorname{gra} A^*$ from $(-x^*, y^*) \in (\operatorname{gra} A)^*$. The first statement is verified.

(2) Suppose $x^* \in (-A)^*(y^*)$, i.e., $(y^*, x^*) \in \text{gra}(-A)^*$. By (1), $(-x^*, y^*) \in (\text{gra}\,(-A))^*$. If $(x, y) \in \text{gra}A$, then, $(x, -y) \in \text{gra}\ (-A)$. Hence $\langle(-x^*, y^*), (x, -y)\rangle \geq 0$, i.e., $\langle -x^*, x\rangle + \langle -y^*, y\rangle \geq 0$. By (1) again, $x^* \in A^*(-y^*)$. The process can be reversed. Thus, (2) is verified.

(3) Suppose $y^* \in \left(A^{-1}\right)^*(x^*)$, i.e., $(x^*, y^*) \in \text{gra}\ \left(A^{-1}\right)^*$. By (1), $(-y^*, x^*) \in \left(\text{gra}\ \left(A^{-1}\right)\right)^*$. For every $(x, y) \in \text{gra}\ A$, then $(y, x) \in \text{gra}\ \left(A^{-1}\right)$, it follows $\langle(-y^*, x^*), (y, x)\rangle \geq 0$, i.e., $\langle x^*, x\rangle + \langle -y^*, y\rangle \geq 0$. By (1) again, $-x^* \in A^*(-y^*)$, i.e., $-y^* \in (A^*)^{-1}(-x^*)$. The process can be reversed. Hence, (3) is verified.

(4) Suppose $y \in A^{**}(x)$, i.e., $(x, y) \in \text{gra}\ (A^{**})$. By (1), $(-y, x) \in (\text{gra}\,A^*)^*$. For every $(y^*, x^*) \in \text{gra}\,A^*$, $\langle(-y, x), (y^*, x^*)\rangle \geq 0$, i.e., $\langle x, x^*\rangle + \langle -y, y^*\rangle \geq 0$. By (1) again, $(y^*, x^*) \in \text{gra}\ A^*$ implies $(-x^*, y^*) \in (\text{gra}\,A)^*$. Hence $\langle -x, x^*\rangle + \langle -y, y^*\rangle \geq 0$. It implies $(-x, -y) \in \text{gra}\ A^{**}$. By Problem (8) in Sect. 2.1, we have $\text{gra}\ A^{**} = \text{gra}\ A$. Hence $-y \in A(-x)$. The process can be reversed. Hence, (4) is verified.

(5) Suppose $y \in A(0)$, then $(0, y) \in \text{gra}\ A$. For every $y^* \in \text{dom}\ A^*$, there is a x^* such that $(y^*, x^*) \in \text{gra}\,A^*$, By (1), $(-x^*, y^*) \in (\text{gra}\,A)^*$, and by the definition of conjugate cone, we have $\langle y^*, y\rangle + \langle -x^*, 0\rangle \geq 0$, then $\langle y^*, y\rangle \geq 0$. It implies $A(0) \subset (\text{dom}\,A^*)^*$.

On the other hand, suppose $y \in (\text{dom}\,A^*)^*$, then for every $y^* \in \text{dom}\ A^*$, $\langle y^*, y\rangle \geq 0$. It follows $\langle y^*, y\rangle + \langle -x^*, 0\rangle \geq 0$ for every $x^* \in A^*(y^*)$. Because $(-x^*, y^*) \in (\text{gra}\,A)^*$ is arbitrarily element, it leads to $(0, y) \in (\text{gra}\,A)^{**}$ which is equal to gra A by Problem (8) of Sect. 2.1. Thus, (5) is verified.

(6) By conclusion (5), $A^{-1}(0) = \left(\text{dom}\ \left(A^{-1}\right)^*\right)^*$. By conclusion (3), $\left(\text{dom}\ \left(A^{-1}\right)^*\right)^* = \left(-\text{dom}\left(-(A^*)^{-1}\right)\right)^* = \left(-\text{dom}(A^*)^{-1}\right)^*$. By the definition of inverse mapping, we have $\left(-\text{dom}\ (A^*)^{-1}\right)^* = (-\text{Im}\ A^*)^* = (\text{Im}\ (-A^*))^*$. The conclusion is implied. □

Some authors used statement (1) as an alternative method to define the adjoint process.

Lemma 3.1.3 gives some characteristics of the convex process which are different from the single-valued linear mapping. By the conclusion (4) in Lemma 3.1.3, we know that, usually, $A^{**} \neq A$ and $-Ax \neq A(-x)$. Furthermore, if A is strict, we still cannot conclude A^* is strict (to see Example 3.1.1). But, we can prove A^* is closed.

Theorem 3.1.1 If A is a convex process, then A^* is closed convex process.

Proof From the notation of $\langle x^*, x\rangle$, it is direct to show that A^* is a convex process by using Lemma 3.1.1. We only prove that it is closed.

Suppose $\left(y_n^*, x_n^*\right) \in \text{gra}\ A^*$ and $\left(y_n^*, x_n^*\right) \to \left(y_0^*, x_0^*\right)$, we now prove $\left(y_0^*, x_0^*\right) \in \text{gra}\ A^*$.

Because $\left(y_n^*, x_n^*\right) \in \text{gra } A^*$, $\left\langle x_n^*, x\right\rangle \leq \left\langle y_n^*, y\right\rangle$ by the definition of adjoint process. For every $x \in X$, $x_n^*(x) \to x_0^*(x)$,[1] i.e., $\left\langle x_n^*, x\right\rangle \to \left\langle x_0^*, x\right\rangle$. Similarly, $\left\langle y_n^*, y\right\rangle \to \left\langle y_0^*, y\right\rangle$. Hence, $\left\langle x_0^*, x\right\rangle \leq \left\langle y_0^*, y\right\rangle$. The conclusion is verified. □

Theorem 3.1.1 illustrates that dom A^* is a closed convex cone.

Example 3.1.1 Let $C \in \mathfrak{L}(X, Y)$ and $K \subset Y$ be a closed convex cone. Define a set-valued mapping A as $A(x) = Cx + K$. Then gra A is a closed convex cone, hence, A is a closed convex process.

In order to obtain the adjoint process A^*, let us prove that the conjugate cone $(\text{gra } A)^*$ can be expressed as

$$(\text{gra } A)^* = \left\{\left(x^*, y^*\right); y^* \in K^*, \;\; x^* = -C^* y^*\right\},$$

where C^* is the adjoint mapping of C, and K^* is the conjugate cone of K.[2] Suppose that $(x, y) \in \text{gra } A$, where $y = Cx + k$ for some $k \in K$. Consider

$$\begin{aligned}\langle (x^*, y^*), (x, y)\rangle &= \langle y^*, y\rangle + \langle x^*, x\rangle \\ &= \langle y^*, Cx + k\rangle + \langle -C^* y^*, x\rangle \\ &= \langle y^*, Cx\rangle + \langle -C^* y^*, x\rangle + \langle y^*, k\rangle \\ &= \langle y^*, k\rangle \\ &\geq 0.\end{aligned}$$

The last inequality is obtained from $y^* \in K^*$. It illustrates $(-C^* y^*, y^*) \in (\text{gra } A)^*$ for $y^* \in K^*$.

If $(x^*, y^*) \in (\text{gra } A)^*$, i.e., $\langle (x^*, y^*), (x, y)\rangle = \langle y^*, y\rangle + \langle x^*, x\rangle \geq 0$ for every pair $(x, y) \in \text{gra } A$, then $y = Cx + k$, we have

$$\begin{aligned}0 &\leq \langle y^*, y\rangle + \langle x^*, x\rangle \\ &= \langle y^*, Cx + k\rangle + \langle x^*, x\rangle \\ &= \langle y^*, Cx\rangle + \langle x^*, x\rangle + \langle y^*, k\rangle \\ &= \langle C^* y^* + x^*, x\rangle + \langle y^*, k\rangle .\end{aligned} \tag{3.1.2}$$

K is closed, hence $0 \in K$. Let $k = 0$. Then Inequality (3.1.2) leads to $\langle C^* y^* + x^*, x\rangle \geq 0$. Since A is strict, both $x, -x \in X$, the above inequality results in $\langle C^* y^* + x^*, x\rangle = 0$ for every $x \in X$. Consequently, $x^* = -C^* y^*$. On the other hand, let $x = 0$, then Inequality (3.1.2) leads to $\langle y^*, k\rangle \geq 0$. The inequality is valid for every $y \in K$. Consequently, $y^* \in K^*$. Thus, we conclude

[1] On X^*, $x_n^* \to x_0^*$ means unified convergence of functions. $x_n^*(x) \to x_0^*(x)$, i.e., $\left\langle x_n^*, x\right\rangle \to \left\langle x_0^*, x\right\rangle$ is convergent at every vector x.

[2] Recall that $K^* = \{x^*; \langle x^*, x\rangle \geq 0, x \in K\}$. K^* is always closed.

$$A^*y^* = \begin{cases} C^*y^*, & y^* \in K^*, \\ \varnothing, & y^* \notin K^*. \end{cases}$$

□

From this example, we can draw two conclusions. At first, A is a strict convex process but its adjoint process A^* may not be strict. Second, A is a set-valued mapping but its adjoint process A^* may be a single-valued mapping on its domain.

In Example 3.1.1, if $K = 0$, then $K^* = X$. Moreover, A reduces to a single-valued mapping and A^* defines on whole X and it is exactly the conjugate mapping of A. It also illustrates that convex process can be treated as an extension of a linear mapping.

3.1.2 The Norm of Convex Processes

This subsection defines the norm for convex processes. By the definition and the properties of norm, the later will be given in the next subsection, we can further understand that the convex process is an extension of the linear bounded mapping.

Definition 3.1.3 Suppose X and Y are two linear normed spaces, and $A : X \to Y$ is a convex process. The norm of A is defined as follows.

$$\|A\| = \sup_{x \in \mathrm{dom} A \cap \mathbf{B}} \inf_{y \in A(x)} \|y\|, \tag{3.1.3}$$

where $\mathbf{B} = \mathrm{bd}\, B$ is the shell of unit ball in X. If $\|A\| < \infty$ then the convex process A is said to be bounded. □

From the definition, if $A = 0$, then $\|A\| = 0$. Furthermore, if A is with closed value, then

$$\begin{aligned} \sup_{x \in \mathrm{dom} A \cap \mathbf{B}} \inf_{y \in A(x)} \|y\| &= \sup_{x \in \mathrm{dom} A} \inf_{y \in A(x)} \frac{\|y\|}{\|x\|} \\ &= \sup_{x \in \mathrm{dom} A} \frac{d\left(0, A(x)\right)}{\|x\|} \\ &= \sup_{x \in \mathrm{dom} A} \frac{m\left(A(x)\right)}{\|x\|} \\ &= \sup_{x \in \mathrm{dom} A \cap \mathbf{B}} m\left(A(x)\right), \end{aligned} \tag{3.1.4}$$

i.e., $\|A\| = \sup\limits_{x \in \mathrm{dom} A \cap \mathbf{B}} m\left(A(x)\right)$. It implies that $A \neq 0$, the equation $\|A\| = 0$ may hold, but if $\|A\| = 0$, then $0 \in A(x)$ for every $x \in \mathrm{dom}\, A$.

Because $A : X \to Y$ is a convex process, for every $\alpha \in \mathbb{R}\,(\geq 0)$, we have

$$\|\alpha A\| = \sup_{x \in \mathrm{dom}(\alpha A)} \inf_{y \in \alpha A(x)} \frac{\|y\|}{\|x\|} = \sup_{x \in \mathrm{dom}(A)} \inf_{\tilde{y} = \frac{y}{\alpha} \in A(x)} \alpha \frac{\|\tilde{y}\|}{\|x\|} = \alpha \sup_{x \in \mathrm{dom}(A)} \inf_{\tilde{y} \in A(x)} \frac{\|\tilde{y}\|}{\|x\|} = \alpha \, \|A\|\,.$$

The positive homogeneity holds for the norm of convex process.

We now consider the triangle inequality of the norm of convex process. Let A_1 and A_2 be two convex processes. By the distance definition, we have

$$d\left(0, A_1(x) + A_2(x)\right) \le d\left(0, A_1(x)\right) + d\left(A_1(x), A_1(x) + A_2(x)\right).$$

From the second term, we have

$$\begin{aligned} d\left(A_1(x), A_1(x) + A_2(x)\right) &= \inf_{y_1, y_2 \in A_1(x), y_3 \in A_2(x)} d\left(y_1, y_2 + y_3\right) \\ &\le \inf_{y_1 \in A_1(x), y_3 \in A_2(x)} d\left(y_1, y_1 + y_3\right) \\ &= \inf_{y_3 \in A_2(x)} d\left(0, y_3\right) \\ &= d\left(0, A_2(x)\right). \end{aligned}$$

Therefore,

$$\begin{aligned} \sup_x \frac{d\left(0, A_1(x) + A_2(x)\right)}{\|x\|} &\le \sup_x \frac{1}{\|x\|}\left(d\left(0, A_1(x)\right) + d\left(0, A_2(x)\right)\right) \\ &\le \sup_x \frac{d\left(0, A_1(x)\right)}{\|x\|} + \sup_x \frac{d\left(0, A_2(x)\right)}{\|x\|}. \end{aligned}$$

By the second equation in Eq. (3.1.4), $\|A_1 + A_2\| \le \|A_1\| + \|A_2\|$.

The above discussion shows that the norm defined in Definition 3.1.3 holds similar properties to the norm of linear mapping, hence it is reasonable.

3.1.3 Fundamental Properties of Convex Processes

The target of defining the norm for the convex process is to extend the fundamental theorems in Banach spaces to the convex processes. To prove these theorems, we have to apply the Robinson-Ursescu theorem, whose proof is quite troublesome and omitted. The readers who are interested in its proof are referred to Aubin and Cellina (1984).

Theorem 3.1.2 (Robinson-Ursescu) Suppose X and Y are two Banach spaces. $F : X \to Y$ is a set-valued mapping with closed and convex value. Suppose $y_0 \in \text{Int Im}\,(F)$, i.e., y_0 is an inner point of $\text{Im}(F)$ and $x_0 \in F^{-1}(y_0)$.[3] Then there is a constant $l > 0$ such that for every $y \in B(y_0, l)$ and $x \in \text{dom}\, A$, we have

$$d\left(x, F^{-1}(y)\right) \le \frac{1}{l} d\left(y, F(x)\right)\left(1 + \|x - x_0\|\right).$$

□

We give a remark to Theorem 3.1.2.

[3] $x_0 \in F^{-1}(y_0)$ means $y_0 \in F(x_0)$.

Remark If we take $x = x_0$, then under the conditions of Theorem 3.1.2, we have.

$$d\left(x_0, F^{-1}(y)\right) \leq \frac{1}{l} d\left(y, F\left(x_0\right)\right).$$

An alternative statement is that there exists an $x \in F^{-1}(y)$ such that

$$\left\|x - x_0\right\| \leq \frac{1}{l}\left\|y - y_0\right\|. \tag{3.1.5}$$

Inequality (3.1.5) implies that F^{-1} is lower semi-continuous at inner points of $\operatorname{Im} F$. □

We now apply Theorem 3.1.2 to verify properties of convex processes. We denote $L = 1/l$.

Theorem 3.1.3 Suppose X and Y are two Banach spaces, and $A : X \to Y$ is a closed and convex process. If $\operatorname{Im}(A) = Y$, then A^{-1} is a Lipschitzian mapping.

Proof We prove that for two $y_1, y_2 \in Y$ and $x_1 \in F^{-1}\left(y_1\right)$, which are selected arbitrarily, then there exists an $x_2 \in A^{-1}\left(y_2\right)$ such that $\left\|x_1 - x_2\right\| \leq L\left\|y_1 - y_2\right\|$.

Because $A : X \to Y$ is a closed and convex process, $0 \in A(0)$ or $0 \in A^{-1}(0)$. We now fix $y_0 = 0, x_0 = 0$ and apply Theorem 3.1.2, there is a neighborhood $B(0, \varepsilon)$ such that for every $y \in B(0, \varepsilon)$, there is an $x \in A^{-1}(y)$ such that

$$\|x\| \leq L\|y\|. \tag{3.1.6}$$

Because A is a convex process, Inequality (3.1.6) holds for every $y \in Y$.

Now for two $y_1, y_2 \in Y$, by Inequality (3.1.6), there is an $e \in A^{-1}\left(y_2 - y_1\right)$ such that $\|e\| \leq L\left\|y_1 - y_2\right\|$. Take arbitrarily an $x_1 \in A^{-1}\left(y_1\right)$ and denote $x_2 = x_1 + e$, then

$$y_2 = y_1 + \left(y_2 - y_1\right) \in A\left(x_1\right) + A(e) \subset A\left(x_1 + e\right) = A\left(x_2\right),$$

i.e., $x_2 \in A^{-1}\left(y_2\right)$. Thus, there exist $x_1 \in A^{-1}\left(y_1\right), x_2 \in A^{-1}\left(y_2\right)$ such that

$$\left\|x_1 - x_2\right\| \leq L\left\|y_1 - y_2\right\|. \tag{3.1.7}$$

□

We give two remarks for Theorem 3.1.3.

Remark 1 Inequality (3.1.7) implies that A^{-1} is lower semi-continuous. □

Remark 2 Theorem 3.1.3 has an alternative statement: Under the conditions of Theorem 3.1.3, for every open set $O \subset X$, $A(O)$ is an open set in Y. The conclusion can be proved as follows. For every $x_0 \in O$, there exists $\varepsilon > 0$ such that $B\left(x_0, \varepsilon\right) \subset O$. We fix a $y_0 \in A\left(x_0\right)$, and take $y \in B\left(y_0, \varepsilon l\right)$ arbitrarily. By Theorem 3.1.3, there is an $x \in A^{-1}(y)$ such that $\|x - x_0\| \leq L \|y - y_0\| < \varepsilon$. It implies $B(y_0, \varepsilon l) \subset A(B(x_0, \varepsilon)) \subset A(O)$, i.e., $A(O)$ is open. The alternative conclusion is sometimes called as open mapping theorem of set-valued mapping. □

Corollary 3.1.2 Suppose X and Y are two Banach spaces. $A : X \to Y$ is a strictly closed and convex process. Then there exists an $\overline{L}$ such that for arbitrary $x_1, x_2 \in X$, we have

$$A(x_1) \subset A(x_2) + \overline{L}\, \|x_1 - x_2\|\, \overline{B}, \tag{3.1.8}$$

where $\overline{B}$ is the closed unit ball in Y.

Proof $A : X \to Y$ is a strictly closed and convex process, by Problem 2 of this section, A^{-1} is a subjective, and closed and convex process. By Theorem 3.1.3, for two vectors $x_1, x_2 \in X$ and one vector $y_1 \in A(x_1)$ which are all selected arbitrarily, there exists $y_2 \in A(x_2)$ such that $\|y_1 - y_2\| \leq \overline{L}\, \|x_1 - x_2\|$. It is equivalent to $y_1 \in y_2 + \overline{L}\, \|x_1 - x_2\|\, \overline{B}$, or $A(x_1) \subset A(x_2) + \overline{L}\, \|x_1 - x_2\|\, \overline{B}$. □

Corollary 3.1.2 is also called as closed graph theorem. If $x_2 = 0$ and $\|x_1\| = 1$, then Relation (3.1.8) leads to $A(x_1) \subset A(0) + \overline{LB}$. Hence if $A(0)$ is bounded, so is $\|A\|$.

To end this section, we prove the uniform boundedness theorem. In functional analysis, the theorem is also called as resonance theorem.

Theorem 3.1.4 Suppose X and Y are two Banach spaces. $A_\alpha : X \to Y$ is a set of closed and convex processes where $\alpha \in A$ and $\boldsymbol{A}$ is the set of indexes. If for every $x \in X$, there is a $y_\alpha \in A_\alpha(x)$ such that $\sup\limits_{\alpha} \{\|y_\alpha\|\} \leq M_x$, there is a constant which may depend on x, then $\sup\limits_{\alpha} \|A_\alpha\| < \infty$.

Proof Define a single-valued mapping as follows:

$$\rho_\alpha(x) = \inf_{y \in A_\alpha(x)} \|y\| = d(0, A_\alpha(x)) = m(A_\alpha(x)).$$

It can easily prove that $\rho_\alpha(x)$ has the following properties:

(1) $\rho_\alpha(x)$ is positively homogeneous.

(2) $\rho_\alpha(x)$ is lower semi-continuous (in a space with finite dimension it is continuous).

We further define

$$\rho(x) = \sup_{\alpha} \rho_\alpha(x).$$

By the conditions of the theorem, $\rho(x) < M_x$ for every $x \in X$, $\rho(x)$ is also positive and positively homogeneous and lower semi-continuous (to see Sect. 1.4 of this book). Then $\rho(x)$ is continuous at the origin. Hence there is a $\delta > 0$, for every $x \in B(0, \delta)$, $\rho(x) \leq 1$. By the positive homogeneity, $\rho(x) \leq \delta^{-1} \|x\|$. Thus,

$$\delta^{-1} \|x\| \geq \rho(x) = \sup_{\alpha} \rho_{\alpha}(x) = \sup_{\alpha} m(A_{\alpha}(x)).$$

It implies $\|A_{\alpha}\| \leq \delta^{-1}$ for every $\alpha \in A$. □

In functional analysis, resonance theorem is a very useful conclusion. In the following, we prove a conclusion which is useful in the control system theory.

Theorem 3.1.5 Let U be a metric space, X and Y be two Banach spaces. For every $u \in U$, $A(u) : X \to Y$ is a closed and convex process where u is treated as a parameter. If for every $x \in X$, there is a $y_u \in A(u)(x)$ such that $\sup_{u} \|y_u\| < \infty$, then the following statements are equivalent.

(1) $u \mapsto \operatorname{gra} A(u)$ is lower semi-continuous.
(2) $(u, x) \mapsto A(u)(x)$ is lower semi-continuous.

Proof The proof of (2) $\Rightarrow$ (1) is direct. Indeed, if we fix $x \in X$, then (2) implies (1).

(1) $\Rightarrow$ (2). By the Problem 2 of Sect. 2.1, it is sufficient to verify that for every sequence $(u_n, x_n) \to (u, x)\,(n \to \infty)$, and every $y \in A(u)(x)$, there is $y_n \in A(u_n)(x_n)$ such that $y_n \to y\,(n \to \infty)$. We now prove the conclusion.

By the condition, $u \mapsto \operatorname{gra} A(u)$ is lower semi-continuous, hence, for $(x, y) \in \operatorname{gra} A(u)$ and $u_n \to u$, $n \to \infty$, there exist $(\widehat{x}_n, \widehat{y}_n) \in \operatorname{gra} A(u_n)$ such that $(\widehat{x}_n, \widehat{y}_n) \to (x, y)$. The condition that for every $x \in X$ there is a $y_u \in A(u)(x)$ such that $\sup_{u} \|y_u\| < \infty$ implies $\|A(u)\|$ is uniformly bounded by using Theorem 3.1.5, i.e., there exists L such that $\|A(u)\| \leq L$ for every $u \in U$. Thus, let us consider $\widehat{x}_n - x_n$, there is $\tilde{y}_n \in A(u_n)(\widehat{x}_n - x_n)$ with $\|\tilde{y}_n\| \leq L \|\widehat{x}_n - x_n\|$. Because $\widehat{x}_n - x_n \to x - x = 0$, we have $\tilde{y}_n \to 0$. Denote $y_n = \tilde{y}_n + \widehat{y}_n$, then

$$y_n = \tilde{y}_n + \widehat{y}_n \in A(u_n)(x_n - \widehat{x}_n) + A(u_n)(\widehat{x}_n) \subset A(u_n)(x_n),$$

and $y_n \to y$ since $\widehat{y}_n \to y$. □

Problems

1. From Definition 3.1.1, prove that $0 \in A(0)$ for convex process A.
2. Prove Lemma 3.1.1.
3. Prove that if $A(x)$ is convex process then $A^{-1}(x)$ is also convex process.
4. Suppose X and Y are two Banach spaces, $L \subset X$, $M \subset Y$ are two closed and convex sets, and $f \in L(X, Y)$. Define

$$F(x) = \begin{cases} f(x) - M, & x \in L, \\ \varnothing, & x \notin L. \end{cases}$$

Prove the following conclusion:

(1) $F^{-1}(y) = \{x;\ x \in y + M\}$;
(2) Suppose $y_0 \in \text{Int}\ (f(L) - M)$, then for every $x_0 \in F^{-1}(y_0)$, there exists $l > 0$ such that for every $y \in B(y_0, l)$, $d\big(x, F^{-1}(y)\big) \leq Ld\,(f(x) - y, M)\,(1 + \|x - x_0\|)$, where $L = l^{-1}$.

5. Suppose X and Y are two Banach spaces, $A : X \to Y$ is a closed and convex process. $K \subset X$ is a closed convex cone and $K - \text{dom}A = X$. Let $A|_K$ is the restriction of A on K, then

$$(A|_K)^*(y) = \begin{cases} A^*(y^*), & y^* \in \text{dom}(A^*), \\ \varnothing, & \text{otherwise.} \end{cases}$$

6. Suppose X and Y are two Banach spaces, $A : X \to Y$ is a closed and convex process. $K \subset X$ is a closed convex cone and $K - \text{dom}A = X$. Then

$$(A(K))^* = \left(A^*\right)^{-1}\left(K^*\right).$$

(The conclusion is also called dual polar theorem. It is a fundamental result of adjoint process.)

3.2 Convex Processes in Spaces with Finite Dimensions

In this section, we restrict ourselves in the n-dimensional space, i.e., we consider the convex processes in $\mathbb{R}^n$. The main target is to give the construction characteristics of a convex process and show that it can have a Jordan-like construction which is one of main properties of the matrix.

3.2.1 Adjoint Processes in n-Dimensional Space

Let us start to recall that definition of adjoint process of a convex process.

If $A : X \to Y$ is a convex process, then its adjoint process $A^* : Y \to X$ is defined by for every $y^* \in Y$,

$$A^*\left(y^*\right) = \{x^*; \langle x^*, x\rangle \leq \langle y^*, y\rangle, (x, y) \in \text{gra}\, A\}.$$

By Lemma 3.1.1, we have that $(y^*, x^*) \in$ gra A^* if and only if $(-x^*, y^*) \in (\text{gra}\, A)^*$.

Theorem 3.2.1 Suppose $A_1 : \mathbb{R}^n \to \mathbb{R}^p$; $A_2 : \mathbb{R}^q \to \mathbb{R}^k$ are two convex processes. Then we have the following conclusions:

(1) If $p = q$, then $A_2A_1 : \mathbb{R}^n \to \mathbb{R}^k$ is a convex process. Furthermore, if $\operatorname{ri\,int}\,(\operatorname{Im} A_1 \cap \operatorname{dom} A_2) \neq \varnothing$, then $(A_2A_1)^* = A_1^*A_2^*$;
(2) If $p = k,\ q = n$ and $\operatorname{int}\,(\operatorname{dom} A_1 \cap \operatorname{dom} A_2) \neq \varnothing$, then $A_1 + A_2$ is a convex process and $(A_1 + A_2)^* = A_1^* + A_2^*$.

Proof The proof of (1) consists of three steps.

(i) Define two sets $K_1, K_2 \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^k$ as follows:

$$K_1 = \{(x, y, z)\,;\,(x, y) \in \operatorname{gra} A_1\} \ \ \textit{and} \ \ K_2 = \{(x, y, z)\,;\,(y, z) \in \operatorname{gra} A_2\}.$$

Because both $\operatorname{gra}A_1$ and $\operatorname{gra}A_2$ are convex cones, K_1 and K_2 are also convex cones. From the definitions of K_1 and K_2, it is direct that $(x, z) \in \operatorname{gra}\ A_2A_1$ if and only if there is a $y \in \mathbb{R}^p$ such that $(x, y, z) \in K_1 \cap K_2$. From the condition $\operatorname{ri\,int}\,(\operatorname{Im} A_1 \cap \operatorname{dom} A_2) \neq \varnothing$, it can be proved that $\operatorname{re\,int}\ K_1 \cap \operatorname{re\,int}\ K_2 \neq \varnothing$. Then by the Problem 9 of Sect. 2.1, we have

$$(K_1 \cap K_2)^* = \operatorname{cl}\left(K_1^* + K_2^*\right).$$

(ii) We now prove that $K_1^* + K_2^*$ is a closed set by contradiction. If $K_1^* + K_2^*$ is not closed, there is a $c^* \in \operatorname{cl}\left(K_1^* + K_2^*\right)$ but $c^* \notin \left(K_1^* + K_2^*\right)$. There exist two sequences $\left\{a_n^*\right\} \subset K_1^*$ and $\left\{b_n^*\right\} \subset K_2^*$ such that $a_n^* + b_n^* \to c^*\ \ (n \to \infty)$.

We then conclude both $\{a_n^*\}$ and $\{b_n^*\}$ are bounded. If $\{a_n^*\}$ is not bounded, $\{a_n^*\}$ has a subsequence $\left\{a_{n_k}^*\right\} \to \infty$. Without lose of generality, we can assume $\left\{a_n^*\right\} \to \infty$. Since $a_n^*/\left\|a_n^*\right\| \in \operatorname{db}B$ where $\operatorname{db}B$ is the closed shell of unit ball of $\mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^k$. $\left\{a_n^*/\left\|a_n^*\right\|\right\}$ has a convergent subsequence, without loss of generality, we can assume $a_n^*/\left\|a_n^*\right\| \to a^* \in K_1^* \cap \operatorname{db}B$. Consider now $\left(a_n^* + b_n^*\right)/\left\|a_n^*\right\| \to c^*/\left\|a_n^*\right\| = 0\,(n \to \infty)$, hence $b_n^*/\left\|a_n^*\right\| \to -a^* \in K_2^*$.

Denote $a^* = \left(x_0^*, y_0^*, z_0^*\right)$. Because $a^* \in K_1^*$, for every $(x, y, z) \in K_1$,

$$\langle x_0^*, x\rangle + \langle y_0^*, y\rangle + \langle z_0^*, z\rangle \geq 0.$$

In K_1, z is not restricted, it leads to $z_0^* = 0$. From $-a^* \in K_2^*$, a similar discussion results in $x_0^* = 0$. Hence $a^* = \left(0, y_0^*, 0\right)$ where $y_0^* \in \mathbb{R}^p$, $\left\|y_0^*\right\| = 1$ since $a^* \in \operatorname{db} B$. Moreover, $\left\langle y_0^*, y\right\rangle \geq 0$ for every $y \in \operatorname{Im} A_1$; on the other hand, $-a^* \in K_2^*$, $\left\langle y_0^*, y\right\rangle \leq 0$ for every $y \in \operatorname{dom}\ A_2$. It follows $\left\langle y_0^*, y\right\rangle \equiv 0$ for $y \in \operatorname{Im}\ A_1 \cap \operatorname{dom}\ A_2$. By the condition that $\operatorname{ri\,int}\ (\operatorname{Im} A_1 \cap \operatorname{dom} A_2) \neq \varnothing$, we have $y_0^* = 0$. It contradicts to $\left\|y_0^*\right\| = 1$. It follows $c^* \in K_1^* + K_2^*$. $K_1^* + K_2^*$ is closed.

(iii) At last, we show $\operatorname{gra}(A_2A_1)^* = \operatorname{gra}\left(A_1^*A_2^*\right)$. Suppose $(z^*, x^*) \in \operatorname{gra}(A_2A_1)^*$. By Lemma 3.1.3 (1), it is equivalent to $(-x^*, z^*) \in (\operatorname{gra} A_2A_1)^*$. It is equivalent to

$$\langle -x^*, x\rangle + \langle 0, y\rangle + \langle z^*, z\rangle \geq 0 \quad \text{for} \quad (x, y, z) \in K_1 \cap K_2.$$

Hence $(-x^*, 0, z^*) \in (K_1 \cap K_2)^* = K_1^* + K_2^*$. There exists a $y^* \in \mathbb{R}^p$, such that $(-x^*, y^*, 0) \in K_1^*$ and $(0, -y^*, z^*) \in K_2^*$, i.e., $(y^*, x^*) \in$ gra A_1^* and $(z^*, y^*) \in$ gra A_2^*. It is just $(z^*, x^*) \in$ gra $\left(A_1^* A_2^*\right)$. The procedure can be deduced inversely. Thus, we verify the conclusion.

(2) The conclusion (2) is verified by using the result of (1). We define three mapping as follows:

$$\begin{aligned}
&M_1 : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n, \;\; M_1(x) = (x, x) \\
&M_2 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^k \times \mathbb{R}^k, \;\; M_2(x_1, x_2) = \{(y_1, y_2)\,; y_1 \in A_1(x_1), y_2 \in A_2(x_2)\} \\
&M_3 : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k, \;\; M_3(x_1, x_2) = x_1 + x_2
\end{aligned}$$

M_1 and M_3 are linear single-valued mappings, using Example 3.1.1, we have $M_1^*\left(y_1^*, y_2^*\right) = y_1^* + y_2^*$ and $M_3^*(y^*) = (y^*, y^*)$.

We now prove $M_2^*\left(y_1^*, y_2^*\right) = \left(A_1^*\left(y_1^*\right), A_2^*\left(y_2^*\right)\right)$. Suppose $\left(x_1^*, x_2^*\right) \in M_2^*\left(y_1^*, y_2^*\right)$. Then,

$$\langle y_1^*, y_1\rangle + \langle y_2^*, y_2\rangle \geq \langle x_1^*, x_1\rangle + \langle x_2^*, x_2\rangle,$$

where $y_1 \in A_1(x_1)$ and $y_2 \in A_2(x_2)$ all variables can be selected arbitrarily in their reasonable ranges. A_1 is a convex process, consequently, $\lambda y_1 \in A_1(\lambda x_1)$ for every $\lambda > 0$. The above inequality leads to

$$\langle y_1^*, \lambda y_1\rangle + \langle y_2^*, y_2\rangle \geq \langle x_1^*, \lambda x_1\rangle + \langle x_2^*, x_2\rangle.$$

Let $\lambda \downarrow 0$, it leads to $\left\langle y_2^*, y_2\right\rangle \geq \left\langle x_2^*, x_2\right\rangle$, or $x_2^* \in A_2^*\left(y_2^*\right)$. Similarly, $x_1^* \in A_1^*\left(y_1^*\right)$.

It is obvious that $A_1 + A_2 = M_3 M_2 M_1$. It is direct to show $A_1 + A_2, M_1, M_2$ and M_3 are all convex processes, and ri in$t$ (Im $M_1 \cap$ dom $M_2) \neq \varnothing$ and other conditions required by (1). By result (1), we have $(A_1 + A_2)^* = M_1^* M_2^* M_3^*$, and

$$M_1^* M_2^* M_3^*\left(y^*\right) = M_1^* M_2^*\left(y^*, y^*\right) = M_1^*\left(A_1^* y^*, A_2^* y^*\right) = A_1^* y^* + A_2^* y^*,$$

for every $y^* \in \mathrm{dom} A_1^* \cap \mathrm{dom} A_2^*$. □

We have mentioned that for a single-valued linear bounded mapping $A : X \to Y$ we have $\langle y^*, Ax\rangle = \langle A^* y^*, x\rangle$, where A^* is the adjoint mapping of A. When A is a set-valued mapping, both Ax and $A^* y^*$ may not be unique. Theorem 3.2.2 gives an extended result for the equation.

Theorem 3.2.2 Suppose $A : \mathbb{R}^n \to \mathbb{R}^m$ is a convex process, $x_0 \in$ int dom A and $y_0^* \in$ dom A^*. Then

$$\sup\left\langle A^*\left(y_0^*\right), x_0\right\rangle = \inf\langle y_0^*, A(x_0)\rangle,$$

where $\left\langle A^*\left(y_0^*\right), x_0\right\rangle = \left\{\langle x^*, x_0\rangle\,;\; x^* \in A^*\left(y_0^*\right)\right\}$, and $\left\langle y_0^*, A(x_0)\right\rangle$ has a similar definition.

Proof Define a single-valued function $\Phi : \mathbb{R}^n \to \mathbb{R}$ as follows

$$\Phi(x) = \inf_{y\in\mathbb{R}^m} \{\langle y_0^*, y\rangle + \delta\left((x, y), \mathrm{gra}\, A\right)\},$$

where $y_0^* \in \mathrm{dom}\ A^*$ is fixed, and $\delta((x, y), \mathrm{gra} A)$ is the indicator of $\mathrm{gra} A$. From Problem 6 of Sect. 1.3, $\Phi(x)$ is a convex function. The proof consists of three steps.

(1) For every $x \in \mathrm{dom}\, A$, $\Phi(x) > -\infty$.

This conclusion is verified by contradiction. If $\Phi(x) = -\infty$, then there is $y_k \in A(x)$, $k = 1, 2, \ldots$, such that $\langle y_0^*, y_k\rangle \to -\infty$. From the definition of adjoint process, we have $\langle y_0^*, y_k\rangle \geq \langle x^*, x\rangle$ where $x^* \in A^*y^*$ can be selected arbitrarily. After x^* is determined, the right side is a constant. Hence $\langle y_0^*, y_k\rangle \to -\infty$ is impossible. We have a contradiction.

(2) $\Phi^{**}(x) = \Phi(x)$ when $x \in \mathrm{int}\ \mathrm{dom}\, A$.

When $x_0 \in \mathrm{int}\ \ \mathrm{dom}\ \ A$, there are at most $n+1$ vectors $x_i \in \mathrm{dom}\ \ A$, $i = 1, 2, \ldots, n+1$, such that $x_0 = \sum_{i=1}^{n+1} \lambda_i x_i$. A is a convex process, hence, $\mathrm{co}\,(x_1, x_2, \ldots, x_{n+1}) \subset \mathrm{dom} A$. It follows that $\Phi\,(x_0) \leq \max\{\Phi\,(x_i)\,;\ i = 1, 2, \ldots, n+1\}$ since $\Phi(x)$ is a convex function. $\Phi(x_0)$ is bounded by conclusion (1), therefore, there is neighborhood of x_0, $\Phi(x)$ is bounded at the neighborhood. By Lemma 1.3.1, $\Phi(x)$ is a locally Lipschitzian function. At last we conclude that $\Phi(x)$ is continuous at x in the neighborhood. By Theorem 1.4.6, $\Phi^{**}(x) = \Phi(x)$ when $x \in \mathrm{int}\ \mathrm{dom}\, A$.

(3) $\sup\langle A^*\left(y_0^*\right), x_0\rangle = \inf\langle y_0^*, A\,(x_0)\rangle$.

By the definition of conjugate function, we have

$$\Phi\,(x_0) = \sup_{x^*}\left\{\langle x^*, x_0\rangle - \Phi^*\left(x^*\right)\right\}. \tag{3.2.1}$$

Similarly, from $\Phi^{**}(x) = \Phi(x)$, $\Phi^*\,(x^*)$ is

$$\begin{aligned}\Phi^*\,(x^*) &= \sup_x \{\langle x, x^*\rangle - \Phi(x)\}\\ &= \sup_x \left\{\langle x, x^*\rangle - \inf_{y\in\mathbb{R}^m}\left(\langle y_0^*, y\rangle + \delta\left((x, y), \mathrm{gra} A\right)\right)\right\}\\ &= \sup_{x,y}\left\{\langle x, x^*\rangle - \langle y_0^*, y\rangle - \delta\left((x, y), \mathrm{gra} A\right)\right\}\\ &= \sup_{x,y}\left\{\langle (x, y), \left(x^*, -y_0^*\right)\rangle - \delta\left((x, y), \mathrm{gra}\, A\right)\right\}\ .\end{aligned}$$

The equation shows that $\Phi^*\,(x^*)$ is just the conjugate function of indicator function $\delta((x, y), \mathrm{gra} A)$, i.e., $\Phi^*\,(x^*) = S\left(\left(x^*, -y_0^*\right), \mathrm{gra}\, A\right)$. By Eq. (2.1.4), we obtain

$$\begin{aligned}\Phi^*\,(x^*) &= S\left(\left(x^*, -y_0^*\right), \mathrm{gra}\, A\right)\\ &= \delta\left(\left(x^*, -y_0^*\right), -(\mathrm{gra}\, A)^*\right)\\ &= \delta\left(\left(-x^*, y_0^*\right), (\mathrm{gra}\, A)^*\right)\\ &= \delta\left(\left(y_0^*, x^*\right), \mathrm{gra}\, A^*\right),\end{aligned}$$

where the last equation obtains from Lemma 3.1.3 (1). Substituting it into Eq. (3.2.1), we have

$$\sup_{x^*}\left\{\langle x^*, x_0\rangle - \delta\left(\left(y_0^*, x^*\right), \operatorname{gra} A^*\right)\right\} = \Phi\left(x_0\right) = \inf_{y}\left\{\langle y_0^*, y\rangle + \delta\left(\left(y, x_0\right), \operatorname{gra} A\right)\right\}.$$

The conclusion is implied. □

Theorem 3.2.3 Suppose $A : \mathbb{R}^n \to \mathbb{R}^m$ is a strictly convex process. Then the following conclusions are valid.

(1) A is a Lipschitzian mapping and its Lipschitzian constant can be $\|A\|$
(2) If $y^* \in \operatorname{dom} A^*$ and $x^* \in A^*\left(y^*\right)$, then $\|x^*\| \leq \|A\| \, \|y^*\|$.
(3) $\operatorname{dom} A^*$ is a closed convex cone.
(4) On $\operatorname{dom} A^* \cap \left(-\operatorname{dom} A^*\right)$, A^* is a linear single-valued mapping.

Proof (1) By Inequality (3.1.5), A is a Lipschitzian set-valued mapping. We now verify its Lipschitzian can be $\|A\|$.

Suppose that $x_1, x_2 \in \mathbb{R}^n$ and $x_1 \neq x_2$. Denote $a = \|x_1 - x_2\|$ and $x_3 = a^{-1}\left(x_1 - x_2\right)$, then $\|x_3\| = 1$. By the definition of norm of convex process, for every $\varepsilon > 0$, there exists $y_3 = y_3(\varepsilon) \in A\left(x_3\right)$ such that $\|y_3\| < \|A\| + \varepsilon$. Let $y_2 \in A\left(x_2\right)$ and $y_1 = y_2 + a y_3$. Because $x_1 = x_2 + a x_3$, $A\left(x_2\right) + A\left(a x_3\right) \subset A\left(x_1\right)$. It implies $y_1 = y_2 + a y_3 \in A\left(x_1\right)$. Then

$$\|y_1 - y_2\| = a\,\|y_3\| < \left(\|A\| + \varepsilon\right) a = \left(\|A\| + \varepsilon\right) \|x_1 - x_2\|. \tag{3.2.2}$$

The above discussion illustrates that for every $y_2 \in A\left(x_2\right)$, there is a $y_1 \in A\left(x_1\right)$ such that Inequality (3.2.2) holds. It implies $A\left(x_2\right) \subset A\left(x_1\right) + \left(\|A\| + \varepsilon\right)\|x_2 - x_1\|\,\overline{B}$, where $\overline{B}$ is the closed ball of $\mathbb{R}^m$. The ε can be selected arbitrarily, thus, $A\left(x_2\right) \subset A\left(x_1\right) + \|A\|\,\|x_2 - x_1\|\,\overline{B}$. The conclusion (1) is verified.

(2) Suppose $y^* \in \operatorname{dom} A^*$ and $x^* \in A^*\left(y^*\right)$. If $x^* = 0$, then the conclusion is true. We now assume that $x^* \neq 0$. By the definition of adjoint process, $\langle y^*, y\rangle \geq \langle x^*, x\rangle$ for every $(x, y) \in \operatorname{gra} A$. If we take $x = \|x^*\|^{-1} x^*$, then $\langle x^*, x\rangle = \|x^*\| \leq \langle y^*, y\rangle$. On the other hand because $\|x\| = 1$, there is a $y \in A(x)$ such that $\|y\| < \|A\| + \varepsilon$ for arbitrary $\varepsilon > 0$. By this y, we apply Schwarz inequality and can obtain

$$\|x^*\| \leq \langle y^*, y\rangle \leq \|y^*\|\,\|y\| \leq \left(\|A\| + \varepsilon\right)\|y^*\|.$$

The ε can be selected arbitrarily, the conclusion is implied.

(3) It is sufficient to prove $\operatorname{dom} A^*$ is closed. Suppose $y_n^* \in \operatorname{dom} A^*$ and $y_n^* \to y_0^*$. For every y_n^*, there is an $x_n^* \in A^*\left(y_n^*\right)$ and $\left\|x_n^*\right\| \leq \|A\|\,\left\|y_n^*\right\|$. It follows that $\left\{x_n^*\right\}$ has a convergent subsequence. Without loss of generality, we assume $\left\{x_n^*\right\}$ is convergent, $x_n^* \to x_0^*$. Hence $\left(y_n^*, x_n^*\right) \to \left(y_0^*, x_0^*\right)$. $\operatorname{gra} A^*$ is closed, consequently, $x_0^* \in A^*\left(y_0^*\right)$, i.e., $y_0^* \in \operatorname{dom} A^*$.

(4) By $\|x^*\| \leq \|A\|\,\|y^*\|$, if $y^* = 0$, then $x^* = 0$, i.e., $A^*(0) = \{0\}$. Because $\operatorname{dom} A^* \cap \left(-\operatorname{dom} A^*\right)$ is a linear subspace of $\mathbb{R}^m$, by the illustration after Corollary

3.1.1, The restriction of A^* on dom $A^* \cap (-\text{dom}\, A^*)$ is a single-valued linear mapping. □

It is obvious that dom $A^* \cap (-\text{dom}\, A^*)$ is the largest subspace in dom A^*. The conclusion (4) illustrates that on this subspace A^* is a single-valued mapping. We note conclusion (1) can be extended to Banach spaces X and Y.

For Theorem 3.2.3, we have two remarks.

Remark 1 Let $x_2 = 0$ and $y_2 = 0$, Inequality (3.2.2) leads to $\|y_1(\varepsilon)\| \leq (\|A\| + \varepsilon)\|x_1\|$ where $y_1(\varepsilon) = y_2 + ay_3(\varepsilon) \in A(x_1)$. Hence, $\{y_1(\varepsilon)\}$ is bounded and has a convergent subsequence. Suppose $y_1(\varepsilon_n) \to y_1^0$, then $y_1^0 \in A(x_1)$ since $A(x_1)$ is closed. It further leads to $\left\|y_1^0\right\| \leq \|A\|\,\|x_1\|$. The conclusion can be stated as for every $x \in$ dom A, there is a $y \in A(x)$ such that $\|y\| \leq \|A\|\,\|x\|$. The property can be treated as an extension of norm of linear single-valued mapping. □

Remark 2 Using Remark 1 to A^*, for every $y^* \in$ dom A^*, we have $x^* \in A^*(y^*)$ such that $\|x^*\| \leq \|A^*\|\,\|y^*\|$. However, the conclusion (2) of Theorem 3.2.2 is $\|x^*\| \leq \|A\|\,\|y^*\|$ for every $x^* \in A^*(y^*)$. It implies $\|A^*\| \leq \|A\|$. □

3.2.2 Structure of Convex Processes

We, in this subsection, deal with the structure of convex process on finite spaces. We will show that convex processes have Jordan-like structure.

A set $A \subset \mathbb{R}^n$ is said to be proper if it does not contain a subspace of $\mathbb{R}^n$.

Definition 3.2.1 Suppose $A : \mathbb{R}^n \to \mathbb{R}^n$ is a convex process, $x \in \mathbb{R}^n$ is a nonzero vector. If there exists a $\lambda \in \mathbb{R}$ such that $\lambda x \in A(x)$, then λ is an eigenvalue of A, and x is an eigenvector of A with eigenvalue λ. □

Differing form the eigenvalues of a matrix, we require here that the eigenvalue is real. Hence a real matrix may have no eigenvalue.

Lemma 3.2.1 Suppose $K \subset \mathbb{R}^n$ is a nonzero and proper closed cone, and $A : \mathbb{R}^n \to \mathbb{R}^n$ is a bounded and closed convex process. K and A also satisfy the following conditions:

(1) $K \subset$ dom A
(2) For every $x \in K$, $A(x) \cap K \neq \varnothing$

Then there exists an $x \in K$, $x \neq 0$ and $\lambda \geq 0$ such that $\lambda x \in A(x)$.

Proof (1) At first, the conclusion is verified for a stronger condition that $A(x) \cap$ re int $K \neq \varnothing$ for every $x \in K$ and $x \neq 0$.

Define a set Ω as follows:

$$\Omega = \left\{\omega \in \mathbb{R}^+; \text{there exists an } x \in K \cap \mathbf{B} \text{ such that } (A(x) - \omega x) \cap K \neq \varnothing\right\}, \tag{3.2.3}$$

where $\mathbf{B}$ is the shell of the unit ball. The condition $A(x) \cap \operatorname{re\,int} K \neq \varnothing$ implies $\omega \in \Omega$ for sufficiently small $\omega > 0$, then $\Omega \neq \varnothing$.

We prove that Ω is bounded. If there is a series $\{\omega_i;\ i = 1, 2, \ldots\}$ such that $\omega_i \in \Omega$ and $\omega_i \to \infty$ $(i \to \infty)$. Then for every ω_i there is an x_i such that $x_i \in K \cap \mathbf{B}$ and $A(x_i) - \omega_i x_i \in K$. The sequence $\{x_i\}$ is bounded so that it has a convergent subsequence; without loss of generality, we assume $x_i \to \overline{x}$. K is closed, hence $\overline{x} \in K \cap \mathbf{B}$. On the other hand, $\omega_i^{-1} A(x_i) - x_i \in K$. Because the set $\{A(x_i), i = 1, 2, \ldots\}$ is bounded, when $i \to \infty$, we obtain $-\overline{x} \in K \cap \mathbf{B}$. But by the condition of theorem K is proper. The contradiction illustrates Ω is bounded.

Because K is closed and A is with closed value, by a similar discussion, we can conclude that Ω is closed. Thus, there is $\omega_0 = \max \Omega$ and $\omega_0 > 0$. For this ω_0, there is an $x_0 \in K \cap \mathbf{B}$ such that $(A(x_0) - \omega_0 x_0) \cap K \neq \varnothing$. We now prove $(A(x_0) - \omega_0 x_0) \cap K = \{0\}$. Let $z_0 \in (A(x_0) - \omega_0 x_0) \cap K$. Then

$$\begin{aligned} A\left(x_0 + \frac{z_0}{2\omega_0}\right) - \omega_0\left(x_0 + \frac{z_0}{2\omega_0}\right) &\supset A(x_0) + A\left(\frac{z_0}{2\omega_0}\right) - \omega_0 x_0 - \frac{z_0}{2} \\ &\supset A\left(\frac{z_0}{2\omega_0}\right) + \frac{z_0}{2}, \end{aligned} \tag{3.2.4}$$

where we apply $z_0 \in (A(x_0) - \omega_0 x_0)$ at the last relation. If $z_0 \neq 0$, then $z_0/2\omega_0 \neq 0$, and $A(z_0/2\omega_0) \cap \operatorname{re\,int} K \neq \varnothing$ by the condition (2) of the theorem. Because K is a convex cone and $z_0/2 \in K$, $(A(z_0/2\omega_0) + z_0/2) \cap \operatorname{re\,int} K \neq \varnothing$. By Relation (3.2.4),

$$\left(A\left(x_0 + \frac{z_0}{2\omega_0}\right) - \omega_0\left(x_0 + \frac{z_0}{2\omega_0}\right)\right) \cap \operatorname{re\,int} K \neq \varnothing. \tag{3.2.5}$$

Denote $y_0 = \left\| x_0 + \frac{z_0}{2\omega_0} \right\|^{-1} \left(x_0 + \frac{z_0}{2\omega_0} \right)$, then $y_0 \in K \cap \mathbf{B}$. Relation (3.2.5) implies

$$(A(y_0) - \omega_0 y_0) \cap \operatorname{re\,int} K \neq \varnothing.$$

It means that we can increase ω_0 to be $\omega_0 + \Delta\omega$ with a small $\Delta\omega$ such that $(A(y_0) - (\omega_0 + \Delta\omega) y_0) \cap K \neq \varnothing$. It contradicts that $\omega_0 = \max \Omega$. Therefore, $z_0 = 0$, i.e., $\omega_0 x_0 \in A(x_0)$. It is equivalent that x_0 is an eigenvector and $\omega_0 > 0$ is its corresponding eigenvalue.

(2) We extend the conclusion to the case that $A(x) \cap K \neq \varnothing$.

K is a closed convex cone, there exist relative inner points. Hence we can fix an $x_c \in \operatorname{re\,int} K$ and define a linear single-valued mapping $A_k : \mathbb{R}^n \to \mathbb{R}^n$ as follows:

$$A_k(x) = k^{-1} \langle x_c, x \rangle x_c.$$

It is easy to verify that foe every $k > 0$ and $x \in \mathbb{R}^n$ $A_k(x) \cap \operatorname{re\,int} K \neq \varnothing$.

We now consider the set-valued mapping $(A + A_k)(x) = A(x) + k^{-1} \langle x_c, x \rangle x_c$. The mapping holds the property that for every $x \in K,\ x \neq 0$, $(A + A_k)(x) \cap$

re int $K \neq \varnothing$. Then using the result of (1), we conclude that there is a $\omega_k > 0$ and $x_k \in K \cap \mathbf{B}$ such that

$$\omega_k x_k \in (A + A_k)(x_k) = A(x_k) + k^{-1} \langle x_c, x_k \rangle x_c.$$

Without loss of generality, we can assume $\omega_k \to \omega_0, \;\; x_k \to x_0$ as $k \to \infty$. It follows $\omega_0 \geq 0$ and $\omega_0 x_0 \in A(x_0)$. Thus, we end the proof. □

Let $K \subset \mathbb{R}^n$ be a convex cone, and $A : \mathbb{R}^n \to \mathbb{R}^n$ be a convex process. K is said to be an invariant cone of A, if for every $x \in K, \;\; A(x) \subset K$. If $K \subset \mathbb{R}^n$ is a subspace, then K is said to be an invariant subspace of A. The facts are denoted by A-invariant cone and A-invariant subspace, respectively.

Suppose $A : \mathbb{R}^n \to \mathbb{R}^n$ is a strict closed and convex process. By conclusion (4) of Theorem 3.2.3, the restriction of A^* on dom $A^* \cap (-\text{dom}\, A^*)$ is a single-valued mapping. Let S be the maximal A^*-invariant subspace contained in dom $A^* \cap (-\text{dom}\, A^*)$. We further denote $T = S^{\perp}$. In the following, we will prove that T is the minimal A-invariant subspace of A, i.e., if there is a subspace T_1 such that $AT_1 \subset T_1$ and $T_1 \subset T$, then $T_1 = T$.

We have mentioned that the restriction of A^* on S is a single-valued linear mapping and S is an A^*-invariant subspace. Consequently, the mapping $A^*|_S$ can be described by a square matrix. Let $S_1 \subset S$ be a minimal A^*-invariant subspace. Then, from the conclusion of linear algebra, $\dim S_1 = 1$ or 2 when we restrict ourselves on the real field. If $\dim S_1 = 1$, then S_1 is an eigen subspace with a real eigenvalue. If $\dim S_1 = 2$, then

$$A^*|_{S_1} = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}.$$

under a basis selected appropriately. In the complex field, $A^*|_{S_1}$ has two eigenvalues $\alpha + \mathrm{i}\beta$ and $\alpha - \mathrm{i}\beta$, and S_1 is spanned by real and imaginary parts of eigenvector of $\alpha + \mathrm{i}\beta$.

Lemma 3.2.2 *T is the minimal A-invariant subspace.*

Proof Let $x \in T$ and $y \in A(x)$. For every $y^* \in S$, the set $A^*(y^*)$ has only one element and $A^*(y^*) \in S$. Then by the definition of adjoint process, we have

$$\langle y^*, y \rangle \geq \left\langle A^*\left(y^*\right), x \right\rangle = 0.$$

The last equation comes from the fact that $x \in T = S^{\perp}$. S is a subspace, hence $-y^* \in S$, and we have also $\langle -y^*, y \rangle \geq \langle A^*(-y^*), x \rangle = 0$. It implies $\langle y^*, y \rangle = 0$, i.e., $y \in S^{\perp} = T$. T is A-invariant.

We now prove that T is a minimal A-invariant subspace by contradiction. If there is another subspace $T_1 \subset T$, T_1 is A-invariant. Because $0 \in T_1$, $A(0) \subset T_1$. By the conclusion (5) in Lemma 3.1.3, $(\text{dom}\, A^*)^* = A(0) \subset T_1$.

Let $v^* \in T_1^{\perp}$, then for every $x \in (\text{dom}\,A^*)^*$, $\langle x, v^* \rangle = 0$, i.e., $v^* \in \left((\text{dom}\,A^*)^*\right)^* = \text{dom } A^*$ by the property of conjugate cone. Consequently, dom $A^* \supset T_1^{\perp}$. By a similar discussion, we can have $-$dom $A^* \supset T_1^{\perp}$, Hence dom $A^* \cap (-\text{dom}\,A^*) \supset T_1^{\perp}$. Denote $S_1 = T_1^{\perp}$, then dom $A^* \cap (-\text{dom}\,A^*) \supset S_1 \supset S$.

We prove now S_1 is A^*-invariant. Let $y^* \in S_1$ and $x \in T_1$ be two arbitrary vectors, and $y \in A(x) \in T_1$ be selected arbitrary too. Then by the definition of adjoint process,

$$\left\langle A^* \left(y^*\right), x\right\rangle \leq \langle y^*, y \rangle = 0,$$

the last equation is valid since T_1 is A-invariant. Replacing x by $-x$, then $-x \in T_1$ and $\langle A^* (y^*), -x \rangle \leq 0$. It leads to $\langle A^* (y^*), x \rangle = 0$, i.e., $A^* (y^*) \in T_1^{\perp} = S_1$. S_1 is A^*-invariant. We have assume that S is the maximal A^*-invariant subspace in dom $A^* \cap (-\text{dom}\,A^*)$. Hence $S = S_1$, so $T_1 = T$. □

Suppose $A : \mathbb{R}^n \to \mathbb{R}^n$ is a convex process and I is the identical mapping on $\mathbb{R}^n$. Then for every $\lambda \in \mathbb{R}$, $A - \lambda I$ is also a convex process. Moreover, $(A - \lambda I)^{-1}$ is also a convex process. By Theorem 3.1.3, if A is a closed convex process and $\|A\| < \infty$, then $(A - \lambda I)^{-1}$ is a closed convex process and its norm is also finite. Moreover, $(A - \lambda I)^{-k}$ is a closed convex process with finite norm for every $k \in \mathbb{N}$. We now define a set $L_k(\lambda)$ as follows:

$$L_k (\lambda) = (A - \lambda I)^{-k}(0), \quad k = 1, 2, \ldots.$$

$L_k(\lambda)$ is a closed convex cone for every $k \in \mathbb{N}$. It is obvious that $L_m (\lambda) \supset L_k (\lambda)$ if $m \geq k$. We further denote

$$L (\lambda) = \underset{k \geq 1}{\cup} L_k (\lambda).$$

Let $\sigma (A^*)$ be the set of all eigenvalues of A^*. Then $\sigma (A^*)$ is a bounded set provided that $\|A\| < \infty$, and $\|A\|$ is a upper boundary. Let $\lambda_M (A^*)$ be the maximal eigenvalue of A^*. Particularly, if $\sigma (A^*) = \varnothing$, then we set $\lambda_M (A^*) = -\infty$.

For the sake of convenience, we denote $B_\lambda = A - \lambda I$, and if λ is given, we simplify B_λ by B. By Theorem 3.2.1, $B^* = A^* - \lambda I$. It is east to verify the following facts and all proofs are omitted.

(1) dom $B =$ dom A
(2) dom $B^* =$ dom A^*
(3) K is an A-invariant subspace (convex cone) if and only if it is B-invariant
(4) K is an A^*-invariant subspace (convex cone) if and only if it is B^*-invariant

By the definition of eigenvalue, if $\lambda > \lambda_M (A^*)$, then $(B^*)^{-k}(0) = \{0\}$ for every $k \in \mathbb{N}$.

We are ready to prove the main result of this subsection. From now on, S is denoted as the maximal A^*-invariant subspace in dom $A^* \cap (-\text{dom } A^*)$.

Theorem 3.2.4 If $S = \{0\}$ and $\lambda > \lambda_M(A^*)$, then $L(\lambda) = \mathbb{R}^n$.

Proof It is sufficient to show that $(L(\lambda))^* = \{0\}$ for the conclusion $L(\lambda) = \mathbb{R}^n$.

We have the following equations.

$$(L_k(\lambda))^* = \left(B^{-k}(0)\right)^* = \text{cl im } \left(-\left(B^k\right)^*\right) = \text{cl im } \left(-\left(B^*\right)^k\right),$$

where the second equation is valid by conclusion (6) of Lemma 3.1.3, we apply the closure since $K^{**} = \text{cl } K$ for every convex cone K; the third equation is due to the conclusion (1) of Theorem 3.2.1.

Denote

$$M = \bigcap_{k \geq 1} \text{clim } \left(B^*\right)^k, \tag{3.2.6}$$

M is a closed convex cone. To prove $(L(\lambda))^* = \{0\}$, it is sufficient to prove $M = \{0\}$. The fact is verified by contradiction. Hence, we assume $M \neq \{0\}$. The proof consists of four steps.

(1) For every $k \in \mathbb{N}$, im $(B^*)^k$ is a closed set.

We have

$$\left(B^*\right)^{-k}(0) = -\left(B^k\right)^*(0) = -\left(\text{dom } B^{-k}\right)^*,$$

where the first equation is the conclusion (2) of Lemma 3.1.3, and the second equation is obtained from conclusions (5) and (4) of Lemma 3.1.3. $\{0\} = (B^*)^{-k}(0)$ as mentioned before the theorem, hence dom$B^{-k} = \mathbb{R}^n$, i.e., B^{-k} is strict. By the remark given after Theorem 3.1.1, dom$\left(B^{-k}\right)^*$ is a closed convex cone. By the conclusion (2) of Lemma 3.1.3 again, $-$dom$\left(B^{-k}\right)^* =$ dom$(B^*)^{-k}$. The right side is exactly equal to im $(B^*)^k$. Hence, im $(B^*)^k$ is closed. Thus, Eq. (3.2.6) can be written as

$$M = \bigcap_{k \geq 1} \text{ im } \left(B^*\right)^k. \tag{3.2.7}$$

(2) $(B^*)^{-1}(x^*) \cap M \neq \varnothing$ for every $x^* \in M$.

By Eq. (3.2.7), $x^* \in M$ implies $x^* \in$ im $(B^*)^k$ for every $k \in \mathbb{N}$. Hence, there is a y_k^* such that $(B^*)^k\left(y_k^*\right) = x^*$. It means there exists a sequence $\left\{y_{k,0}^*, y_{k,1}^*, \ldots, y_{k,k}^*\right\}$ such that

$$y_{k,k}^* = y_k^*, y_{k,k-1}^* \in B^*\left(y_{k,k}^*\right), \ldots, y_{k,1}^* \in B^*\left(y_{k,2}^*\right), y_{k,0}^* = x^* \in B^*\left(y_{k,1}^*\right),$$

i.e., $y_{k,i}^* \in B^*\left(y_{k,i+1}^*\right)$ for $i = 1, 2, \ldots, k-1$. Thus, we obtain a two-dimensional sequence as follows:

$$\begin{aligned}&k = 1,\ x^*, y^*_{1,1};\\ &k = 2,\ x^*, y^*_{2,1}, y^*_{2,2};\\ &\cdots\\ &k = n,\ x^*, y^*_{n,1}, y^*_{n,2}, \ldots, y^*_{n,i}, \ldots, y^*_{n,n};\\ &\cdots \quad\ \cdots\end{aligned}$$

From the second column, we obtain $y_{k,1} \in (B^*)^{-1}(x^*)$ for $k = 1, 2, \ldots$; from the third column, we have $y_{k,2} \in (B^*)^{-2}(x^*)$ for $k = 2, 3, \ldots$; $\ldots\ldots$; from the ith column, we have $y_{k,i} \in (B^*)^{-i}(x^*)$ for $k = i, i+1, \ldots$; $\ldots\ldots$.

By Theorem 3.1.3, $\|B^{-k}\|$ is bounded for every $k \in \mathbb{N}$. Thus, from conclusion (2) of Theorem 3.2.3, for every $i \in \mathbb{N}$, the infinite sequence $\{y^*_{k,i}, k = i, i+1, \ldots\}$ is bounded and holds a convergent subsequence. Let $\{y^*_{k_1,1}\} \subset \{y^*_{k,1}\}$ be a convergent subsequence. Furthermore, from the subsequence $\{y^*_{k_1,2}\} \subset \{y^*_{k,2}\}$, we can obtain a convergent subsequence $\{y^*_{k_2,2}\} \subset \{y^*_{k_1,2}\}$. By this way, we can obtain $\{y^*_{k_i,i}\} \subset \{y^*_{k_{i-1},i}\} \subset \{y^*_{k,i}\}$, $\{y^*_{k_i,i}\}$ is a convergent sequence. Let y^*_i be the limitation of series $\{y^*_{k_i,i}\}$. Then $y^*_i \in B^*(y^*_{i+1})$ for every $i \in \mathbb{N}$. Thus, $y^*_i \in (B^*)^2(y^*_{i+2}), \ldots, y^*_i \in (B^*)^k(y^*_{i+k})$, $\ldots\ldots$. Therefore, $y^*_i \in \operatorname{im}(B^*)^k$ for $k \in \mathbb{N}$, i.e., $y^*_i \in M$. Particularly, for y^*_1, $y^*_1 \in (B^*)^{-1}(x^*) \cap M$.

(3) $M \cap (-M) = \{0\}$, i.e., M does not contain a subspace of $\mathbb{R}^n$.

Suppose $N = M \cap (-M)$. If there is a nonzero vector $q \in N \subset M$, then by the second part of the proof, $(B^*)^{-1}(q) \cap M \neq \varnothing$. Because $M \subset \operatorname{im} B^* = \operatorname{dom}(B^*)^{-1}$, we obtain $N \subset \operatorname{dom}(B^*)^{-1} \cap \left(-\operatorname{dom}(B^*)^{-1}\right)$. The $(B^*)^{-1}(q) \cap M \neq \varnothing$ implies $(B^*)^{-1}(q) \in M$. By the same reason, form $-q \in N$, we can obtain $(B^*)^{-1}(-q) \in M$. Restricted on N, $(B^*)^{-1}$ is a single-valued linear mapping. Hence, $(B^*)^{-1}(-q) = -(B^*)^{-1}(q)$, $(B^*)^{-1}(q) \in M \cap (-M) = N$.

We have mentioned that $(B^*)^{-1}(0) = \{0\}$. It implies that on subspace N, the linear mapping $(B^*)^{-1}$ is one-to-one. We obtain $B^*(N) \subset N$. S is the maximal B^*-invariant space in $\operatorname{dom} B^* \cap (-\operatorname{dom} B^*)$, hence $N \subset S$. The condition of theorem assumes that $S = \{0\}$. It contradicts with $q \in N$ and $q \neq 0$. We therefore conclude that $M \cap (-M) = \{0\}$.

(4) A^* holds an eigenvalue which is large than λ.

The above proofs in (2) and (3) illustrate that closed convex process $(B^*)^{-1}$ satisfies all conditions given in Lemma 3.2.1. Hence, there is an nonzero vector $x^* \in M$ which is a eigenvector of $(B^*)^{-1}$ and whose eigenvalue μ is nonnegative. Moreover, $\{0\} = (B^*)^{-1}(0)$, consequently, $\mu > 0$. Thus,

$$\mu x^* \in \left(B^*\right)^{-1}\left(x^*\right) = \left(A^* - \lambda I\right)^{-1}\left(x^*\right),$$

or

$$\left(\lambda + \frac{1}{\mu}\right) x^* \in A^*\left(x^*\right),$$

i.e., x^* is a eigenvector of A^* and its eigenvalue is $\lambda + (1/\mu)$.

It contradicts the assumption that $\lambda > \lambda_M(A^*)$. The contradiction implies $M = \{0\}$. □

Remark It is obvious that $L_k(\lambda) \subset L_{k+1}(\lambda)$. If $L_k(\lambda) = L_{k+1}(\lambda)$, then

$$L_{k+2}(\lambda) = (A - \lambda I)^{-1} L_{k+1}(\lambda) = (A - \lambda I)^{-1} L_k(\lambda) = L_{k+1}(\lambda).$$

Thus, we can conclude $L_{k+i}(\lambda) = L_k(\lambda)$ for all $i = 2, \ldots$.

We now assume k_0 is the least integer such that $L_{k_0}(\lambda) = L_{k_0+1}(\lambda)$ and $L_{k_0-1}(\lambda) \neq L_{k_0}(\lambda)$. When the conditions of Theorem 3.2.4 are satisfied, $L_{k_0}(\lambda) = \mathbb{R}^n$, i.e., $(A - \lambda I)^{-k_0}(0) = \mathbb{R}^n$, or $\operatorname{im}(A - \lambda I)^{k_0} = \{0\}$.

It implies that there exists a vector $x \in \mathbb{R}^n$, $x \neq 0$ such that $(A - \lambda I)^{k_0}(x) = 0$ but $(A - \lambda I)^{k_0-1}(x) \neq 0$. Then we can find nonzero vectors $y_1, y_2, \cdots, y_{k_0}$ such that

$$\begin{aligned}
&\lambda y_1 \in A(y_1), \\
&y_1 + \lambda y_2 \in A(y_2), \\
&\cdots \\
&y_{k_0-1} + \lambda y_{k_0} \in A(y_{k_0}), \\
&y_{k_0} = x.
\end{aligned} \tag{3.2.8}$$

The procedure illustrates under the conditions of Theorem 3.2.4, (1) y_1 is an eigenvector of A and its eigenvalue is λ; (2) A holds infinite eigenvalues; (3) Relation (3.2.8) provides a Jordan-like structure of A. □

Theorem 3.2.5 If the conditions of Theorem 3.2.4 are satisfied, then $k_0 < \infty$.

Proof There is a simplex co $(x_1, x_2, \ldots, x_{n+1})$ with vertexes $x_1, x_2, \ldots, x_{n+1}$ such that the origin is an inner point of the simplex, i.e., $0 \in \operatorname{int}$ co $(x_1, x_2, \ldots, x_{n+1})$. Because $L(\lambda) = \bigcup_{k \geq 1} L_k(\lambda) = \mathbb{R}^n$, there is a k_i such that $x_i \in L_{k_i}(\lambda)$ for $i = 1, 2, \ldots, n+1$. Let $k_0 = \max_i \{k_i\}$. Then $x_i \in L_{k_0}(\lambda)$ for $i = 1, 2, \ldots, n+1$. $L_{k_0}(\lambda)$ is a closed convex cone, $x \in L_{k_0}(\lambda)$ for every $x \in$ co $(x_1, x_2, \ldots, x_{n+1})$. It implies there is a $\varepsilon > 0$ such that $B(0, \varepsilon) \subset L_{k_0}(\lambda)$ because $0 \in \operatorname{int}$ co $(x_1, x_2, \ldots, x_{n+1})$. $L_{k_0}(\lambda)$ is a cone, consequently, for every $x \in \mathbb{R}^n$ $x \in L_{k_0}(\lambda)$. □

To state the last conclusion of this section, we recall prevalent conclusions of linear algebra. Let X be a linear space, $V \subset X$ be a subspace. For every $x \in X$, we can construct an equivalent class $\{y; y - x \in V\}$. The class is denoted by $\overline{x}$ or $x + V$. It is easy to verify that these $\overline{x}$'s construct a linear space with the operations $\overline{x}_1 + \overline{x}_2 = \overline{x_1 + x_2}$ and $a\overline{x} = \overline{ax}$. The linear space is called the quotient space of X related to V and denoted by X/V. $\overline{0} = V$ is the origin of X/V. $P : X \to X/V$ with $Px = \overline{x}$ is called the natural projection (or projection for simplicity) from X to X/V. $P^{-1} : X/V \to X$ is a set-valued mapping, for every $\overline{x} \in X/V$, $P^{-1}\overline{x} = x + V$.

Suppose $A : X \to X$ is a linear mapping, and V is an A-invariant subspace. Then A can induce an linear mapping $\overline{A} : X/V \to X/V$ defined by $\overline{A}\overline{x} = \overline{Ax}$. By the definition of projection, we have $\overline{A}Px = \overline{A}\overline{x} = \overline{Ax} = PAx$ for every $x \in X$. Hence

$$\begin{tikzcd} X \arrow[r, "A"] \arrow[d, "P"'] & X \arrow[d, "P"] \\ X/V \arrow[r, "\overline{A}"] & X/V \end{tikzcd}$$

Fig. 3.1 The interchanging of linear mapping and projection

$PA = \overline{A}P$. The fact is often described by an interchanging plot (Fig. 3.1). In Fig. 3.1, there are a starting point X at the upper left corner and a terminal point X/V at the lower right corner. From the starting point to terminal point, there are two ways. The plot is said to be interchangeable if gains of the two wags are equal, i.e., PA, the gain of $X \to X \to X/V$, is equal to the gain of $X \to X/V \to X/V$, $\bar{A}P$.

We now assume that X is an n-dimensional space $\mathbb{R}^n$, $\{x_{n-\nu+1}, x_{n-\nu+2}, \ldots, x_n\}$ is a set of basis of V, then we find $x_1, x_2, \ldots, x_{n-\nu}$ such that $\{x_1, x_2, \ldots, x_{n-\nu}; x_{n-\nu+1}, x_{n-\nu+2}, \ldots, x_n\}$ form a set of basis of X. By the basis, X/V is isomorphic to Span $\{x_1, x_2, \ldots, x_{n-\nu}\}$. V is A-invariant, the mapping A has its matrix form of

$$\begin{bmatrix} A_{11} & 0 \\ A_{21} & A_V \end{bmatrix},$$

$P = [I \quad 0]$, and $\overline{A} = A_{11}$. Let $x \in X$. Then $x = \sum_{i=1}^{n-\nu} a_i x_i + \sum_{j=n-\nu}^{n} b_j x_j$ where $a_i, b_j \in \mathbb{R}$, $i = 1, 2, \ldots, n-\nu; j = n-\nu+1, \ldots, n$. $\overline{x} = Px = \sum_{i=1}^{n-\nu} a_i x_i$. If $\overline{y} = \overline{x}$, then $y = \sum_{i=1}^{n-\nu} a_i x_i + \sum_{j=n-\nu}^{n} c_j x_j$, where $c_j \in \mathbb{R}, j = n-\nu+1, \ldots n$ can be selected arbitrarily. There is an $x_0 = \sum_{i=1}^{n-\nu} a_i x_i \in X$ such that $\overline{x}_0 = \overline{x}$. The x_0 is called exact representative of the set $\overline{x} = x + V$. Although x_0 and $\overline{x}$ belong to different spaces, their expressions are identical. Consequently, we often do not distinguish them. The topology of X/V is defined to be the topology of Span $\{x_1, x_2, \ldots, x_{n-\nu}\}$. It is also obvious that $x_0 = m\left(P^{-1}(\overline{x})\right)$ and $\|\overline{x}\|_{X/V} = \|x_0\|_X = \left\| m\left(P^{-1}(\overline{x})\right)\right\|_X$.

We return to consider the convex process. Suppose $A : \mathbb{R}^n \to \mathbb{R}^n$ is a convex process, V is an A-invariant subspace. By the discussion above, we can construct a quotient space $\mathbb{R}^n/V$ and an induced process $\overline{A} : \mathbb{R}^n/V \to \mathbb{R}^n/V$ where $\bar{A}$ is defined as $\overline{A}(\overline{x}) = \overline{A(x)}$.

By the definition, $PA(x) = \overline{A(x)} = \overline{A}(\overline{x}) = \overline{A}(Px) = \overline{A}P(x)$, i.e., $PA = \overline{A}P$. Figure 3.1 is still valid for convex process and its induced process.

The following lemma is fundamental for the induced mapping.

Lemma 3.2.3

(1) The mapping $\bar{A}$ is well-defined, i.e., if $\bar{x} = \bar{y}$ (i.e., $y \in x+V$), then $\overline{A}\,(\bar{x}) = \overline{A}\,(\bar{y})$.
(2) $\bar{y} \in \overline{A}\,(\bar{x})$ if and only if $(y + V) \cap A(x) \neq \varnothing$.
(3) $\overline{A} : \mathbb{R}^n/V \to \mathbb{R}^n/V$ is a convex process.
(4) If A is closed, then $\bar{A}$ is also closed.

Proof (1) If $\bar{x} = \bar{y}$, i.e., $y = x + v$ for some $v \in V$, then $A(y) = A\,(x + v) \supset A(x) + A(v)$. V is A-invariant, hence $A(y) + V \supset A(x) + V$. By the same procedure, we can obtain $A(x)+V \supset A(y)+V$. Thus, $A(x)+V = A(y)+V$, i.e., $PA(x) = PA(y)$, or $\overline{A(x)} = \overline{A(y)}$.

(2) $\bar{y} \in \overline{A}\,(\bar{x}) = \overline{A(x)}$. By the definition of quotient space, $y \in A(x) + V$. It is equivalent to $(y + V) \cap A(x) \neq \varnothing$. The procedure can be inversed. (2) is verified.

(3) A is a convex process, hence gra$A$ is a convex cone. For $(\bar{x}_1, \bar{y}_1)\,, (\bar{x}_2, \bar{y}_2) \in$ gra $\overline{A}$, there exist $v_i \in V,\ \ i = 1, 2, 3, 4$ such that $(x_1 + v_1, y_1 + v_2)\,, (x_2 + v_3, y_2 + v_4) \in$ gra A.

$$\begin{aligned} \lambda\,(y_1 + v_2) + (1-\lambda)\,(y_2 + v_4) &\in \lambda A\,(x_1 + v_1) + (1-\lambda)\,A\,(x_2 + v_3) \\ &\subset A\,(\lambda\,(x_1 + v_1) + (1-\lambda)\,(x_2 + v_3))\ \ . \end{aligned}$$

Then

$$P\,(\lambda\,(y_1 + v_3) + (1-\lambda)\,(y_2 + v_4)) \in PA\,(\lambda\,(x_1 + v_1) + (1-\lambda)\,(x_2 + v_3))\,.$$

The projection P is a linear mapping, consequently,

$$\begin{aligned} \lambda P\,(y_1 + v_3) + (1-\lambda)\,P\,(y_2 + v_4) &\in \overline{A}P\,(\lambda\,(x_1 + v_1) + (1-\lambda)\,(x_2 + v_3)) \\ &= \overline{A}\,(\lambda P\,(x_1 + v_1) + (1-\lambda)\,P\,(x_2 + v_3))\,, \end{aligned}$$

or

$$\lambda \bar{y}_1 + (1-\lambda)\,\bar{y}_2 \in \overline{A}\,(\lambda \bar{x}_1 + (1-\lambda)\,\bar{x}_2)\,.$$

(4) A is a closed convex process, i.e., gra$A$ is closed. gra $\overline{A} = \{(PA(x),\ Px)\}$, by the topology defined for $\mathbb{R}^n/V$ gra$\bar{A}$ is closed.

Thus, we complete the proof of Lemma 3.2.3. □

Recall the definitions of subspaces T and S. T is the least A-invariant subspace. $S = T^{\perp}$, S is the largest A^*-invariant subspace contained in dom $A^* \cap (-\text{dom } A^*)$. S is isomorphic to $\mathbb{R}^n/T$. If $\dim T = \nu$, we can select a set of orthogonal basis of R^n such that the vectors in S and T have the forms of $[x_1 \cdots x_{n-\nu}\ 0 \cdots 0]^T$ and $[0 \cdots 0\ x_{n-\nu+1} \cdots x_n]^T$, respectively. By this way, every vector $x \in \mathbb{R}^n$, $x = x_S + x_T$ where $x_S \in S$ and $x_T \in T$, x_S and x_T are determined uniquely by x. Because S is isomorphic to $\mathbb{R}^n/T$, the vector in $\mathbb{R}^n/T$ is also denoted by x_S.

Lemma 3.2.4 Suppose $A : \mathbb{R}^n \to \mathbb{R}^n$ is a strictly convex process. S and T are spaces defined before Lemma 3.2.2. $\bar{A}$ is the induced mapping of A on $\mathbb{R}^n/T$. Then we have

(1) $\overline{A} : \mathbb{R}^n/T \to \mathbb{R}^n/T$ is a single-valued linear mapping.
(2) $\left(\overline{A}\right)^* = A^*|_S$.

Proof (1) It is sufficient to verify that $\overline{A}\left(\overline{x}\right) = PA\left(\left[x_S^T \;\; 0\right]^T\right)$ has only one element where $[x_S^T \;\; 0]^T$ is the exact representative of $x + T$.

If there is an $\overline{x} \in \mathbb{R}^n/T$ such that $p, q \in \overline{A}\left(\overline{x}\right)$, $p \neq q$, then there is an $s^* \in S \sim \mathbb{R}^n/T$ such that $\langle p, s^* \rangle < \langle q, s^* \rangle$. By the definition of adjoint process, we have $\langle x, A^*\left(s^*\right)\rangle \leq \langle (p \;\; v), (s^* \;\; 0)\rangle = \langle p, s^* \rangle$ since $p \in \overline{A}\left(\overline{x}\right)$ then there is a v such that $[p \;\; v] \in A(x)$. We have mentioned after Lemma 3.2.1, on S A^* is a single-valued linear, replacing s^* by $-s^*$, we obtain $\langle x, A^*\left(-s^*\right)\rangle \leq \langle p, -s^* \rangle$, or $\langle x, A^*\left(s^*\right)\rangle \geq \langle p, s^* \rangle$. Hence $\langle x, A^*\left(s^*\right)\rangle = \langle p, s^* \rangle$. Similarly, we can obtain $\langle x, A^*\left(s^*\right)\rangle = \langle q, s^* \rangle$. A contradiction appears.

Thus, we conclude that $\overline{A}\left(\overline{x}\right)$ has only one element for every $\overline{x} \in \mathbb{R}^n/T$.

(2) Because $\overline{A}\left(\overline{x}\right) = PA\left(\left[x_S^T \;\; 0\right]^T\right)$, this means $PA : X \to X/V$, $(PA)^* : X/V \to X$. By Theorem 3.2.1, $(PA)^*\left(\overline{x}^*\right) = A^*P^*\left(\overline{x}^*\right) = A^*\left[\left(x_S^*\right)^T \;\;\; 0\right]^T = A^*|_S\left(x_S^*\right)$. Hence, $\left(\overline{A}\right)^* = A^*|_S$. □

The next corollary is the corresponding results for adjoint processes.

Corollary 3.2.1 Let $\overline{(A^*)}$ be the induced mapping of A^* on the quotient space R^n/S. Then we have the following conclusions.

(1) $\overline{(A^*)} : \mathbb{R}^n/S \to \mathbb{R}^n/S$ is a closed convex process.
(2) $x^* \in \overline{(A^*)}\left(\widehat{y}^*\right)$ if and only if $\left(x^* + S\right) \cap A^*\left(y^*\right) \neq \varnothing$.
(3) $\overline{(A^*)} = \left(A|_T\right)^*$.
(4) dom$\overline{(A^*)}$ does not contain any $\overline{(A^*)}$-invariant subspace.
(5) The eigenvalues of $\overline{(A^*)}$ are all those of A^*.

Proof The proofs of Conclusions (1), (2), and (3) are similar to those of corresponding conclusions given in Lemmas 3.2.3 and 3.2.4. Hence, we only verify Conclusions (4) and (5) below.

(4) Let P be the projection form $\mathbb{R}^n$ to $\mathbb{R}^n/S$. If $\widehat{N}$ is an $\overline{(A^*)}$-invariant subspace in dom$\overline{(A^*)}$, i.e., $\overline{(A^*)}\left(\widehat{N}\right) \subset \widehat{N}$. By the definition of quotient space, $\overline{(A^*)}P\left(\widehat{N} + S\right) \subset \widehat{N}$. From Fig. 3.1, $\overline{(A^*)}P = PA^*$, hence, $PA^*\left(\widehat{N} + S\right) \subset \widehat{N} = P\left(\widehat{N} + S\right)$. It implies $A^*\left(\widehat{N} + S\right) \subset \widehat{N} + S$. $\widehat{N} + S$ is an A^* -invariant subspace in dom $A^* \cap (-\text{dom } A^*)$. It contradicts to the condition that S is maximal. Therefore, $\widehat{N} = 0$.

(5) Suppose λ is an eigenvalue of $\overline{(A^*)}$, but not an eigenvalue of A^*. There is a $\overline{y}^* \in \mathbb{R}^n/S$ and $\overline{y}^* \neq 0$ such that $\lambda\overline{y}^* \in \overline{(A^*)}\left(\overline{y}^*\right)$, or $\lambda\overline{y}^* \in \overline{A^*\left(y^*\right)}$. By the definition of quotient space, it is equivalent to $\lambda y^* - s \in A^*\left(y^*\right)$ where $s \in S$. By

the assumption, λ is not an eigenvalue of A^*, hence $s \neq 0$. By Theorem 3.2.3, on S $(A^* - \lambda I)$ is an one-to-one mapping, there is an $s_0 \in S$ such that $(A^* - \lambda I)\, s_0 = s$. Thus,

$$\lambda y^* - \left(A^* - \lambda I\right) s_0 \in A^* \left(y^*\right),$$

or

$$\lambda \left(y^* + s_0\right) \in A^* \left(y^*\right) + A^* \left(s_0\right) \subset A^* \left(y^* + s_0\right).$$

$y^* \notin S$, $s_0 \in S$, hence $y^* + s_0 \neq 0$. λ is an eigenvalue of A^*. A contradiction appears. □

Summing up the conclusions obtained in Theorem 3.2.4, Lemma 3.2.4 and Corollary 3.2.1, we have the following theorem.

Theorem 3.2.6 Suppose $A : \mathbb{R}^n \to \mathbb{R}^n$ is a strictly closed and convex process, $\lambda > \lambda_M \left(A^*\right)$. The subspaces T and S are those defined in Lemma 3.2.2, Then, we have:

(1) On the quotient space $\mathbb{R}^n/T$, $\bar{A}$, the induced mapping of A, is a single-valued linear mapping. And the adjoint mapping of $\left(\overline{A}\right)^*$ equates to the restriction of $A^*|_S$;
(2) There is an integer k such that $T = \left(\left(A - \lambda I\right)|_T\right)^{-k}(0)$. □

The first conclusion of Theorem 3.2.6 is obtained from Lemma 3.2.4. The second conclusion is derived from Theorem 3.2.5. Hence the detailed proof is omitted. Theorem 3.2.6 illustrates that if we decompose $\mathbb{R}^n$ into $\mathbb{R}^n = T \oplus S$ where the notation $\oplus$ means direct union of subspaces, if A is a strictly closed and convex process, then on the subspace S is a single-valued linear mapping, and on T it is a direct union of some sets. In detail, if $x \in T$, then there exist $y_i \in T$, $i = 1, 2, \ldots, k$ such that

$$\begin{aligned} &\lambda y_1 \in A \left(y_1\right), \\ &y_1 + \lambda y_2 \in A \left(y_2\right), \\ &\cdots \\ &y_{k-1} + \lambda y_k \in A \left(y_k\right), \\ &y_k = x. \end{aligned}$$

The relation given above shows that on T the convex process has Jordan-like structure. It is similar to a linear mapping on finite space.

Problems

1. If $A : \mathbb{R}^n \to \mathbb{R}^m$ is a convex process, then $(\mathrm{Im}(A))^* = \left(A^*\right)^{-1}(0)$ and $\left(\mathrm{dom}\left(A^*\right)\right)^* = A(0)$. By using the conclusion prove that ImA is dense on $\mathbb{R}^m$ if and only if $\left(A^*\right)^{-1}(0) = \{0\}$.

2. Apply the definition of adjoint mapping to show the adjoint mappings of M_1 and M_3 defined in the proof of Theorem 3.2.1 are $M_1^*\left(y_1^* \ y_2^*\right) = y_1^* + y_2^*$ and $M_3^*\left(y^*\right) = \left(y^*, y^*\right)$.
3. Let $\mathbb{R}^n/V$ be a quotient space, $P : \mathbb{R}^n \to \mathbb{R}^n/V$ is the projection, then if C is a closed set in $\mathbb{R}^n$, then PC is a closed set of $\mathbb{R}^n/V$. The conclusion holds also for the open set O. But the inverse conclusion is not true.
4. Suppose $A : \mathbb{R}^n \to \mathbb{R}^n$ is bounded, closed, and convex process. $x_n, x_0 \in \operatorname{dom} A$ and $x_n \to x_0\ (n \to \infty)$. Then for a convergent sequence $\{y_n, y_n \in A(x_n)\}$, $y_n \to y_0\ (n \to \infty)$. then $y_0 \in A(x_0)$.
5. Prove that if A is convex process, then $A - \lambda I$ is also a convex process where $\lambda \in \mathbb{R}$ and I is the identical mapping.
6. Prove that the vectors ω_k defined in the proof of second part of Lemma 3.2.1 form a bounded set.
7. Denote $B_\lambda = A - \lambda I$ where A is a convex process, prove the following conclusions:

   (1) $\operatorname{dom} B = \operatorname{dom} A, \quad \operatorname{dom} B^* = \operatorname{dom} A^*$;
   (2) K is a convex cone or a subspace, K is A-invariant if and only if K is B-invariant;
   (3) K is a convex cone or a subspace, K is A^*-invariant if and only if K is B^*-invariant.

8. Let S and T be the subspaces defined before Lemma 3.2.2. For every $x \in \mathbb{R}^n$, there is a unique decomposition $x = x_S + x_T$ where $x_S \in S,\ \ x_T \in T$. Prove the following conclusions.

   (1) There are two mappings $A_S^* : S \to S$ is a linear single-valued mapping and $A_T^* : T \to \mathbb{R}^n$ is a convex process such that
   $$A^*y^* = A_S^*y_S^* + A_T^*y_T^* = \left[A_S^* \ A_T^*\right]\begin{bmatrix} y_S^* \\ y_T^* \end{bmatrix};$$
   (2) $A^{**}(x) = \begin{bmatrix} A_S x_S \\ A_T^{**} x \end{bmatrix} = \begin{bmatrix} A_S x_S \\ -A_T(-x) \end{bmatrix}.$

9. Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be linear mapping, $K \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$ be two closed and convex cones. A set-valued mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is defined as follows:
   $$F(x) = \begin{cases} Ax + S, & x \in K, \\ \varnothing, & x \notin K. \end{cases}$$
   Prove that: (1) F is a closed and convex process;
   (2) Find the adjoint process $F^* : \mathbb{R}^m \to \mathbb{R}^n$;
   (3) $A(K) + S = \mathbb{R}^m$.

3.3 Controllability of Convex Processes

This section deals with the differential inclusion described by convex process, i.e.,

$$\dot{x}(t) \in A\left(x(t)\right) \tag{3.3.1}$$

where $x(\cdot) \in \mathbb{R}^n, A : \mathbb{R}^n \to \mathbb{R}^n$ is a strictly closed and convex process. By Theorem 3.2.3, A is a Lipschitzian process, and its Lipschitzian constant can be $\|A\|$. In the following, we call Inclusion (3.3.1) to be convex process system for simplicity. $A^* : \mathbb{R}^n \to \mathbb{R}^n$ is the adjoint process of A. The following differential inclusion

$$\dot{y}^*(t) \in -A^*\left(y^*(t)\right) \tag{3.3.2}$$

is called by conjugate process system of Inclusion (3.3.1).

Theorem 3.1.1 illustrates A^* is a closed and convex process. By Inequality (3.1.5), Inclusion (3.3.2) is also a Lipschitzian differential inclusion. Moreover, by Theorem 2.3.3 and the subsequent corollary, solutions of Inclusions (3.3.1) and (3.3.2) exist. We have assumed that A and A^* are all closed processes, hence $0 \in A(0)$ and $0 \in A^*(0)$, i.e., the origin is the equilibrium of the both inclusions.

3.3.1 T-Controllability

In Sect. 2.4, we considered the reachable set of a differential inclusion. The concept of reachability is defined for the terminal. We have proved that the reachable set is an arc connected set. The concept defined in this section is about the starting point. Although we only consider the controllability of convex process system and its conjugate system, the concept can be extended to other differential inclusions.

Definition 3.3.1 Consider Inclusion (3.3.1), a set P_T is defined as follows

$$P_T = \left\{p \in \mathbb{R}^n ; \text{there exists an } x(t) \in S_{[0,T]}\left(A, p\right), \text{ such that } x(T) = 0\right\}.$$

P_T is the controllability set of Inclusion (3.3.1) at time T. For simplicity, it is called as T-controllability set. □

Every reader who is familiar with the theory of linear system can find the definition of controllability set is quite similar to the definition for linear system.

Lemma 3.3.1 Let $S(A, x_0)$ be the solution set of Inclusion (3.3.1) with initial condition $x\left(t_0\right) = x_0$. If $x(t) \in S\left(A, x_0\right)$, then for every $\alpha \in \mathbb{R}\left(\geq 0\right)$, $\alpha x(t) \in S\left(A, \alpha x_0\right)$; if $x_i(t) \in S\left(A, x_{i0}\right)$, $i = 1, 2$, then $\left(x_1\left(t\right) + x_2(t)\right) \in S\left(A, x_{10} + x_{20}\right)$.

Proof The proof is direct. $\dot{x}(t) \in A\left(x(t)\right)$, $\alpha\dot{x}(t) \in \alpha A\left(x(t)\right) = A\left(\alpha x(t)\right)$ since A is a convex process. And $\alpha x(0) = \alpha x_0$, i.e., $\alpha x(t) \in S\left(A, \alpha x_0\right)$.

Because

$$\dot{x}_1(t) + \dot{x}_2(t) \in A\,(x_1(t)) + A\,(x_2(t)) \subset A\,(x_1(t) + x_2(t))\,,$$

and $x_1(0) + x_2(0) = x_{10} + x_{20}$, $(x_1(t) + x_2(t)) \in S\,(A, x_{10} + x_{20})$. □

By Lemma 3.3.1, it is clear that P_T is a convex cone.

A set for the Inc. (3.2.2) is defined as follows

$$Q_T = \left\{q \in \mathbb{R}^n; \text{there exists a } y^*(t) \in S_{[0,T]}\left(-A^*, q\right)\right\}.$$

Note that the definition of Q_T has no any requirement for the terminal $y^*(T)$.With a similar proof of Lemma 3.3.1, it is direct to show Q_T is also a convex cone.

We have found that at many situations, A^* has more useful properties than A, for example $A^*(0) = \{0\}$ there is only one element, and if we restrict A^* on a subspace, then A^* is a linear single-valued mapping. These are reason why we prefer investigate A^* to A. In this section, we will apply A^* to deal with the controllability of A.

We start with a simple fact.

Lemma 3.3.2 Fro Inc. (3.3.2), $S_{[0,T]}\,(-A^*, 0) = \{0\}$.

Proof By conclusion (2) of Theorem 3.2.3, we have $\left\|\dot{y}^*(t)\right\| \leq \|A\|\,\|y^*(t)\|$. Using Lemma 2.3.1 (Gronwall inequality), we can obtain that if $y^*(0) = 0$, then $y^*(t) \equiv 0$. □

Because $0 \in A(0)$, when $T_1 \leq T_2$, we conclude that $P_{T_1} \subset P_{T_2}$. From the definition of Q_T, we have $Q_{T_1} \supset Q_{T_2}$ if $T_1 \leq T_2$.

The main conclusion of this section is the following theorem which gives the duality of P_T and Q_T.

Theorem 3.3.1 $P_T^* = -Q_T$.

Proof Suppose $q \in Q_T$, then there exists a $y^*(t) \in S_{[0,T]}\,(-A^*, q)$. Let $p \in P_T$ be selected arbitrarily. Then there is an $x(t) \in S_{[0,T]}\,(A, p)$ and $x(T) = 0$. We have

$$\begin{aligned}\langle p, q\rangle = \langle x(0), y^*(0)\rangle &= \langle x(0), y^*(0)\rangle - \langle x(T), y^*(T)\rangle \\ &= -\int_0^T \frac{d}{ds}\langle x(s), y^*(s)\rangle\, ds \\ &= -\int_0^T \left\langle \dot{x}(s), y^*(s)\right\rangle + \left\langle x(s), \dot{y}^*(s)\right\rangle ds \\ &\leq 0.\end{aligned}$$

It implies $-q \in P_T^*$, hence, $-Q_T \subset P_T^*$.

We now prove the opposite conclusion, i.e., $-Q_T \supset P_T^*$, or equivalently, we prove that for every $-q \in P_T^*$, then $q \in Q_T$, i.e., $S_{[0,T]}\left(-A^*, q\right) \neq \varnothing$. The proof consists of three steps.

(1) Constructing a finite series $\{\widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_N\}$ where $\widehat{v}_i \in \mathbb{R}^n$.

Let N be a positive integer. Then we define $\tau = T/N$. Take $v_N \in A(0)$, a finite series can be obtained as follows

$$\begin{aligned}
&v_N \in A(0),\\
&v_{N-1} \in A\left(-\tau v_N\right),\\
&\cdots\\
&v_1 \in A\left(-\tau\left(v_N + v_{N-1} + \cdots + v_2\right)\right).
\end{aligned} \tag{3.3.3}$$

Because dom $A = \mathbb{R}^n$, the $v_1, v_2, \ldots, v_N$ exist. Define a set $K \subset \mathbb{R}^{2Nn}$ as follows.

$$K = \left\{\left(u_N^T, \ldots, u_1^T, v_N^T, \ldots v_1^T\right)^T \in \mathbb{R}^{2Nn}; \quad (u_i, v_i) \in \operatorname{gra} A, i = 1, 2, \ldots, N\right\}.$$

By Relation (3.3.3), K is nonempty and is a closed and convex cone since A is a strict closed convex process. For convenience, denote $\overline{v} = \left[v_N^T v_{N-1}^T \cdots v_2^T v_1^T\right]^T \in \mathbb{R}^{Nn}$ and define a $2Nn \times Nn$ real matrix Λ as follows:

$$\Lambda = \begin{bmatrix}
0 & 0 & 0 & \cdots & 0\\
-\tau I & 0 & 0 & \cdots & 0\\
-\tau I & -\tau I & 0 & \cdots & 0\\
\cdots & \cdots & \cdots & \cdots & \cdots\\
-\tau I & -\tau I & -\tau I & \cdots & 0\\
I & 0 & 0 & \cdots & 0\\
\cdots & \cdots & \cdots & \cdots & \cdots\\
0 & 0 & 0 & \cdots & I
\end{bmatrix} \in \mathbb{R}^{2Nn \times Nn},$$

where $I \in \mathbb{R}^{n \times n}$ is the identical matrix and 0 is the $n \times n$ zero matrix, then we have

$$\begin{aligned}
\Lambda\overline{v} &= \Lambda\left[v_N^T v_{N-1}^T \cdots v_1^T\right]^T\\
&= \left[0^T - \tau v_N^T - \tau\left(v_N^T + v_{N-1}^T\right) \cdots - \tau\left(v_N^T + v_{N-1}^T + v_2^T\right) v_N^T v_{N-1}^T \cdots v_1^T\right]^T.
\end{aligned}$$

If $\Lambda\overline{v} \in K$, then $\overline{v} = \left[v_N^T v_{N-1}^T \cdots v_1^T\right]^T$ satisfies Relation (3.3.3).

We now consider the following optimization problem with constraint,

$$\begin{aligned}
\inf \Psi\left(\overline{v}\right) &= \tau\sum_{k=1}^{N}\langle q, v_k\rangle + \tau\sum_{k=1}^{N}\langle v_k, v_k\rangle\\
&= \tau\left\langle\overline{q}, \overline{v}\right\rangle + \tau\|\overline{v}\|^2,\\
&\text{s.t.} \quad \Lambda\overline{v} \in K,
\end{aligned}$$

where $-q \in P_T^*$ and $\overline{q} = \Big[\overbrace{q^T q^T \cdots q^T}^{N}\Big]^T \in \mathbb{R}^{Nn}$. An alternative expression of the optimization problem is

$$\text{Inf}\ \ \Phi\left(\overline{v}\right) = \Psi\left(\overline{v}\right) + \delta\left(\overline{v}, \Lambda^{-1}K\right) = \tau\left\langle \overline{q}, \overline{v}\right\rangle + \tau\|\overline{v}\|^2 + \delta\left(\overline{v}, \Lambda^{-1}K\right), \tag{3.3.4}$$

where Λ^{-1} is the inverse mapping of Λ hence $\Lambda^{-1}K$ is a set of $\mathbb{R}^{Nn}$. $\delta\left(\overline{v}, \Lambda^{-1}K\right)$ is the indicator function of $\Lambda^{-1}K$. A is a strict convex process, hence, it is direct to show

$$\Lambda\mathbb{R}^{Nn} - K = \mathbb{R}^{2Nn}. \tag{3.3.5}$$

By the definition of $\Psi\left(\overline{v}\right)$, it is true that $\Psi\left(\overline{v}\right) \to \infty$ as $\overline{v} \to \infty$. Hence, there is an $M \in \mathbb{R}$ such that

$$\inf_{\overline{v}\in\mathbb{R}^{Nn}} \Psi\left(\overline{v}\right) = \inf_{\|\overline{v}\|\leq M} \Psi\left(\overline{v}\right) = \min_{\|\overline{v}\|\leq M} \Psi\left(\overline{v}\right).$$

A is a closed convex process, $K \cap \Lambda\left\{\overline{v}; \|\overline{v}\| \leq M\right\}$ is a closed set of $\mathbb{R}^{2Nn}$; consequently, there is a $\widehat{v}$ with $\|\widehat{v}\| \leq M$ and $\widehat{v} \in \Lambda^{-1}K$ such that $\Psi\left(\widehat{v}\right) = \min\limits_{\|\overline{v}\|\leq M} \Psi\left(\overline{v}\right) = \inf \Psi\left(\overline{v}\right)$, or equivalently, $\Phi\left(\widehat{v}\right) = \min \Phi\left(\overline{v}\right)$. Since $\Psi(0) = 0$, $\Phi\left(\widehat{v}\right) = \Psi\left(\widehat{v}\right) \leq 0$.

(2) Estimating the boundary of $\widehat{v}$.

In order to obtain a solution of Inc. (3.3.1) by using Theorem 2.3.3, a function $\widehat{x}(t)$ is constructed as follows:

$$\widehat{x}(t) = \begin{cases} (t-k\tau)\,\widehat{v}_k - \tau\left(\widehat{v}_{k+1} + \cdots + \widehat{v}_N\right), & t \in \left[(k-1)\,\tau, k\tau\right], \\ (t-N\tau)\,\widehat{v}_N, & t \in \left[(N-1)\,\tau, N\tau\right]. \end{cases} \quad k = 1, 2, \ldots, N-1,$$

The $\widehat{x}(t)$ is a piecewise linear function, and $\widehat{x}(T) = \widehat{x}\left(N\tau\right) = 0$. Let $M = \left\{\widehat{x}(t)\right\}$ and $r_0(x) = x$. Then for every $p \in P_T$, there exists a $\tilde{x}(t) \in S_{[0,T]}\left(A, p\right)$ by Theorem 2.3.3[4] such that

$$\|\widehat{x}(t) - \tilde{x}(t)\| \leq \int_t^T e^{\|A\|(T-s)} d\Big(\dot{\widehat{x}}(s), A\left(\widehat{x}(s)\right) ds,$$

where $d(\dot{\widehat{x}}(s), A\left(\widehat{x}(s)\right))$ is the distance between $\dot{\widehat{x}}(s)$ and $A\left(\widehat{x}(s)\right)$. Especially,

[4] We apply Theorem 2.3.3 at the terminal time T, i.e., it requires $\widehat{x}(T) = \tilde{x}(T)$.

$$\|\widehat{x}(0)-\tilde{x}(0)\| \leq \int_0^T e^{\|A\|(T-s)} d\Big(\dot{\widehat{x}}(s), A\,(\widehat{x}(s))\, ds \leq e^{\|A\|T} \int_0^T d\Big(\dot{\widehat{x}}(s), A\,(\widehat{x}(s))\, ds.$$

By the definition of $\widehat{x}(s)$, we have

$$\begin{aligned}\int_0^T d\left(\dot{\widehat{x}}(s), A\,(\widehat{x}(s))\right) ds &= \sum_{k=1}^{N} \int_{(k-1)\tau}^{k\tau} d\left(\dot{\widehat{x}}(s), A\,(\widehat{x}(s))\right) ds \\ &\leq \sum_{k=1}^{N} \int_{(k-1)\tau}^{k\tau} \left[d\left(\widehat{v}_k, A\left(\widehat{x}\,(k\tau)\right)\right) + d\left(A\,(\widehat{x}(s))\,, A\left(\widehat{x}\,(k\tau)\right)\right)\right] ds,\end{aligned}$$

Because $\widehat{x}\,(k\tau) = -\tau\,(\widehat{v}_{k+1} + \cdots + \widehat{v}_N)$ and $\widehat{v}_k \in A\left(\widehat{x}\,(k\tau)\right)$ by Relation (3.3.3), $d\left(\widehat{v}_k, A\left(\widehat{x}\,(k\tau)\right)\right) = 0$ and $d\left(A\,(\widehat{x}(s))\,, A\left(\widehat{x}\,(k\tau)\right)\right) \leq \|A\|\,\|\widehat{x}(s) - \widehat{x}\,(k\tau)\| = \|A\|\,\|\widehat{v}_k\|\,|s-k\tau|$ since $s \in [(k-1)\,\tau, k\tau]$. We then obtain

$$\int_0^T d\left(\dot{\widehat{x}}(s), A\,(\widehat{x}(s))\right) ds \leq \|A\| \sum_{k=1}^{N} \|\widehat{v}_k\| \int_{(k-1)\tau}^{k\tau} |s-k\tau|\, ds = \frac{1}{2}\,\|A\|\,\tau^2 \sum_{k=1}^{N} \|\widehat{v}_k\|.$$

It follows that

$$\|\widehat{x}(0)-\tilde{x}(0)\| \leq \frac{1}{2} e^{T\|A\|}\,\|A\|\,\tau^2 \sum_{k=1}^{N} \|\widehat{v}_k\|. \tag{3.3.6}$$

At the end of the first step, we have mentioned that

$$0 \geq \Psi\,(\widehat{v}) = \tau \left\langle q, \sum_{k=1}^{N} \widehat{v}_k \right\rangle + \tau \sum_{k=1}^{N} \|\widehat{v}_k\|^2.$$

Because $\widehat{x}(0) = -\tau \sum_{k=1}^{N} \widehat{v}_k$, $\tau \left\langle q, \sum_{k=1}^{N} \widehat{v}_k \right\rangle = \langle -q, \widehat{x}(0)\rangle$. Moreover, we have

$$\begin{aligned} 0 &\geq \langle -q, \widehat{x}(0)\rangle + \tau \sum_{k=1}^{N} \|\widehat{v}_k\|^2 + \langle q, \tilde{x}(0)\rangle - \langle q, \tilde{x}(0)\rangle \\ &\geq \langle q, \tilde{x}(0) - \widehat{x}(0)\rangle + \tau \sum_{k=1}^{N} \|\widehat{v}_k\|^2 \\ &\geq -\|q\|\,\|\tilde{x}(0) - \widehat{x}(0)\| + \tau \sum_{k=1}^{N} \|\widehat{v}_k\|^2. \end{aligned}$$

In the second inequality, we have apply the fact that $\langle -q, \tilde{x}(0)\rangle \geq 0$ which can be proved by $-q \in P_T^*$ and $\tilde{x}(0) = p \in P_T$. From Inequality (3.3.6), we have

$$\tau \sum_{k=1}^{N} \|\widehat{v}_k\|^2 \leq \|q\| \, \|\tilde{x}(0) - \widehat{x}(0)\| \leq \frac{1}{2} \mathrm{e}^{T\|A\|} \, \|q\| \, \|A\| \, \tau^2 \sum_{k=1}^{N} \|\widehat{v}_k\| \leq c\tau^2 \sum_{k=1}^{N} \|\widehat{v}_k\|,$$

where $c = \frac{1}{2} e^{T\|A\|} \, \|q\| \, \|A\|$. Using the inequality $\sqrt{\frac{1}{n}\sum_{i=1}^{n} a_i^2} \geq \frac{1}{n}\sum_{i=1}^{n} a_i$, we obtain $\sqrt{N\sum_{k=1}^{N} \|\widehat{v}_k\|^2} \geq \sum_{k=1}^{N} \|\widehat{v}_k\|$. Substituting it to the above inequality leads to for $i = 1, 2, \ldots, N$,

$$\|\widehat{v}_i\| \leq \sqrt{\sum_{k=1}^{N} \|\widehat{v}_k\|^2} \leq \frac{cT}{\sqrt{N}}, \tag{3.3.7}$$

where we apply the equation $N\tau = T$.

(3) Constructing a solution $y^*(t) \in S_{[0,T]}(-A^*, q)$.

Recall the optimization (3.3.4), by Theorem 1.3.6, we have $0 \in \partial\Phi(\widehat{v})$, i.e.,

$$0 \in \partial\Phi(\widehat{v}) = \tau\overline{q} + 2\tau\widehat{v} + \partial\delta\left(\widehat{v}, \Lambda^{-1}K\right) = \tau\overline{q} + 2\tau\widehat{v} - \left(\Lambda^{-1}K\right)^*,$$

where we apply the fact that $\partial\delta\left(\widehat{v}, \Lambda^{-1}K\right) = -\left(\Lambda^{-1}K\right)^*$ by Problem 10 (4) of Sect. 2.1. Also from Problem 10 (3) of Sect. 2.1, Eq. (3.3.5) implies $\left(\Lambda^{-1}K\right)^* = \Lambda^* K^*$, therefore, $0 \in \tau\overline{q} + 2\tau\widehat{v} - \Lambda^* K^*$. There is an $(\overline{x}^*, \overline{v}^*) \in K^*$ such that

$$0 = \tau\overline{q} + 2\tau\widehat{v} - \Lambda^*\left(\overline{x}^*, \overline{v}^*\right).$$

The equation is equivalent to

$$\begin{aligned} &\tau\left[q^T q^T \cdots q^T\right]^T + 2\tau\left[\widehat{v}_N^T \widehat{v}_{N-1}^T \cdots \widehat{v}_1^T\right]^T \\ &= \left[\left(-\tau\sum_{k=1}^{N-1} x_k^* + v_N^*\right)^T \left(-\tau\sum_{k=1}^{N-2} x_k^* + v_{N-1}^*\right)^T \cdots \left(-\tau x_1^* + v_2^*\right)^T \left(v_1^*\right)^T\right]^T. \end{aligned} \tag{3.3.8}$$

Rewrite Eq. (3.3.8) in its component form, we obtain

$$\begin{aligned} &\tau q + 2\tau\widehat{v}_1 = v_1^*, \\ &\tau q + 2\tau\widehat{v}_2 = -\tau x_1^* + v_2^*, \\ &\cdots\cdots \\ &\tau q + 2\tau\widehat{v}_N = -\tau\left(x_{N-1}^* + \cdots + x_1^*\right) + v_N^*. \end{aligned} \tag{3.3.9}$$

Because $(\overline{x}^*, \overline{v}^*) \in K^*$, $\left(v_k^*, x_k^*\right) \in \operatorname{gra}(-A^*)$, $k = 1, 2, \ldots, N$. Let $\widehat{y}_k^* = \left(x_k^*/\tau\right)$, $k = 1, 2, \ldots, N$. Then by the first equation in (3.3.9), we obtain $\tau\widehat{y}_1^* \in -A^*(\tau q + 2\tau\widehat{v}_1)$, or $\widehat{y}_1^* \in -A^*(q + 2\widehat{v}_1)$. From the second equation of (3.3.9), similarly, we have $\widehat{y}_2^* \in -A^*\left(q + \tau y_1^* + 2\widehat{v}_2\right)$. Generally, we have

$$\widehat{y}_k^* \in -A^*\left(q + \tau\left(y_1^* + \cdots y_{k-1}^*\right) + 2\widehat{v}_k\right). \tag{3.3.10}$$

A function $y_N^*(t)$ is constructed by using $\widehat{y}_k^*$ as follows:

$$y_N^*(t) = \begin{cases} q + t\widehat{y}_1^*, & t \in [0, \tau], \\ \quad\cdots\cdots \\ q + \tau\left(\widehat{y}_1^* + \cdots \widehat{y}_k^*\right) + (t - k\tau)\widehat{y}_{k+1}^*, & t \in [k\tau, (k+1)\tau], \\ \quad\cdots\cdots \\ q + \tau\left(\widehat{y}_1^* + \cdots \widehat{y}_{N-1}^*\right) + (t - (N-1)\tau)\widehat{y}_N^*, & t \in [(N-1)\tau, N\tau]. \end{cases}$$

It is easy to see that this $y_N^*(t)$ holds the following properties:

(i) $y_N^*(0) = q$..
(ii) $\dot{y}_N^*(t) = \widehat{y}_{k+1}^* \in -A^*\left(q + \tau\left(\widehat{y}_1^* + \cdots + \widehat{y}_k^*\right) + 2\widehat{v}_k\right), t \in [k\tau, (k+1)\tau]$..

If we denote $z_N^*(t) = 2\widehat{v}_k - (t - k\tau)\widehat{y}_{k+1}^*, \quad t \in [k\tau, (k+1)\tau]$, then

$$\dot{y}_N^*(t) \in -A^*\left(y_N^*(t) + z_N^*(t)\right). \tag{3.3.11}$$

By the definition of $z_N^*(t)$, it is easy to obtain that $\sup\left|z_N^*(t)\right| \leq 2\max\|\widehat{v}_k\| + \tau\left\|\widehat{y}_{k+1}^*\right\|$. A^* is a closed and bounded operator, by Relation (3.3.10), $\widehat{y}_{k+1}^*$ is bounded. Using Inequality (3.3.7) and the definition of $\tau = T/N$, we conclude that $\sup\left|z_N^*(t)\right| \to 0$ $(N \to \infty)$.

$$\begin{aligned} y_N^*(k\tau) &= q + \tau\left(\widehat{y}_1^* + \cdots + \widehat{y}_k^*\right) = q + \tau\left(\widehat{y}_1^* + \cdots + \widehat{y}_{k-1}^*\right) + \tau\widehat{y}_k^* \\ &= y_N^*((k-1)\tau) + \tau\widehat{y}_k^*. \end{aligned}$$

By Eq. (3.3.10), we have

$$\begin{aligned} \left\|y_N^*(k\tau)\right\| &\leq \left\|y_N^*((k-1)\tau)\right\| + \tau\left\|\widehat{y}_k^*\right\| \\ &\leq \left\|y_N^*((k-1)\tau)\right\| + \tau\|A\|\left\|q + \tau\left(y_1^* + \cdots + y_{k-1}^*\right) + 2\widehat{v}_k\right\| \\ &\leq \left\|y_N^*((k-1)\tau)\right\| + \tau\|A\|\left\|y_N^*((k-1)\tau)\right\| + 2\tau\|A\|\|\widehat{v}_k\| \\ &\leq (1 + \tau\|A\|)\left\|y_N^*((k-1)\tau)\right\| + \frac{2\tau cT\|A\|}{\sqrt{N}}. \end{aligned}$$

Recursively, we have

$$\begin{aligned}\left\|y_N^*(k\tau)\right\| &\leq (1+\tau\|A\|)\left\|y_N^*((k-1)\tau)\right\| + \frac{2\tau cT\|A\|}{\sqrt{N}} \\ &\leq (1+\tau\|A\|)^2\left\|y_N^*((k-2)\tau)\right\| + ((1+\tau\|A\|)+1)\frac{2\tau cT\|A\|}{\sqrt{N}} \\ &\quad \cdots\cdots \\ &\leq (1+\tau\|A\|)^k\|q\| + \left((1+\tau\|A\|)^{k-1} + \cdots + (1+\tau\|A\|)+1\right)\frac{2\tau cT\|A\|}{\sqrt{N}} \\ &\leq (1+\tau\|A\|)^N\|q\| + \left((1+\tau\|A\|)^N - 1\right)\frac{2cT}{\sqrt{N}}.\end{aligned}$$

$\tau = T/N$, hence $(1+\tau\|A\|)^N = (1+(T\|A\|/N))^N \to \mathrm{e}^{\frac{1}{T\|A\|}}$ $(N\to\infty)$. There is a constant $\tilde{c}$ such that for every $t \in [0,T]$ and every integer N $\left\|y_N^*(t)\right\| \leq \tilde{c}$, and $\left\|\dot{y}_N^*(t)\right\| \leq \|A\|\,\tilde{c}$ by Inc. (3.3.11). Using Arzela-Ascoli Theorem, there is a subseries of $\left\{y_N^*(t), N=1,2,\ldots\right\}$ and $\left\{\dot{y}_N^*(t), N=1,2,\ldots\right\}$, which is uniformly convergent. Without loss of generality, we can say $y_N^*(t) \to y^*(t)$, $\dot{y}_N^*(t) \to \dot{y}^*(t)$ uniformly. Finally, using Inc. (3.3.11), we obtain $y^*(t) \in S_{[0,T]}(-A^*, q)$. The theorem is verified. □

Remark Transforming the optimization problem to the (3.3.4) is a meaningful step. The optimization introduces $\Lambda^* K^*$ so that we can transfer the problem to the adjoint process. The transformation can be treated as an extension of Lagrange multiplication. □

3.3.2 Controllability

This subsection extends the discussion P_T to the more general case.

Denote $P = \underset{T>0}{\cup} P_T$. Because $P_{T_1} \subset P_{T_2}$ $(T_1 \leq T_2)$, we have $P = \lim\limits_{T\to\infty} P_T$. Dually, if we define $Q = \underset{T>0}{\cap} Q_T$, similarly, we have $Q_{T_1} \supset Q_{T_2}$ $(T_1 \leq T_2)$. Thus, $Q = \lim\limits_{T\to\infty} Q_T$. By Theorem 3.3.1, it is direct that $P^* = -Q$.

Definition 3.3.2 The set P is called by the controllability set of Inc. (3.3.1). If $P = \mathbb{R}^n$ then the Inc. (3.3.1) is said to be controllable, or the convex process A is controllable or A holds controllability. □

Theorem 3.3.2 If Inc. (3.3.1) is controllable, then there is a $T < \infty$, such that $P_T = \mathbb{R}^n$.

Proof: We denote $\overline{B}$ for the closed ball in $\mathbb{R}^n$. There exist $n+1$ vectors $x_1, x_2, \ldots, x_{n+1} \in \mathbb{R}^n$ such that $\operatorname{conx}(x_1, x_2, \ldots, x_{n+1}) \supset \overline{B}$ where $\operatorname{conx}(x_1, x_2, \ldots, x_{n+1})$ is the simplex generated by $x_1, x_2, \ldots, x_{n+1}$.

Because Inc. (3.3.1) is controllable, for every x_i, $i \in \{1, 2, \ldots, n+1\}$, there is a $x_i(t) \in S_{[0,T_i]}(A, x_i)$ such that $x_i(T_i) = 0$. Let $T = \max\{T_i, i = 1, 2, \ldots, n+1\}$. Then by $P_T \supset P_{T_i}$, we have an $\tilde{x}_i(t) \in S_{[0,T]}(A, x_i)$, $\tilde{x}_i(T) = 0$ for every $i \in \{1, 2, \ldots, n+1\}$. Because $\operatorname{conx}(x_1, x_2, \ldots, x_{n+1}) \supset \overline{B}$, for every $x_0 \in \overline{B}$, there exist $\lambda_i \in [0, 1]$, $i \in \{1, 2, \ldots, n+1\}$, $\sum_{i=1}^{n+1} \lambda_i = 1$ such that $x_0 = \sum_{i=1}^{n+1} \lambda_i x_i$. By Lemma 3.3.1, we conclude $x(t) = \sum_{i=1}^{n+1} \lambda_i \tilde{x}_i(t) \in S_{[0,T]}(A, x_0)$ and $x(T) = 0$.

Now, let $x_0 \in \mathbb{R}^n$, there is a $\alpha_0 \in \mathbb{R}^+$ such that $\alpha x_0 \in \overline{B}$. Using Lemma 3.3.1 again, $\alpha^{-1}x(t) \in S_{[0,T]}(A, x_0)$. $x(T) = 0$, consequently, $\alpha^{-1}x(T) = 0$. □

The following corollary is a natural result of Theorem 3.3.2.

Corollary 3.3.1 If Inc. (3.3.1) is controllable, then there is a $T < \infty$ such that $Q_T = \{0\}$. □

The main result of this subsection is the following theorem.

Theorem 3.3.3 Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be strictly closed and convex process. Then A is controllable if and only if A^* has no an invariant subspace and also no nontrivial eigenvector.

Proof Necessity. If A^* holds a nontrivial invariant subspace S, then by the Corollary 3.1.1 and illustration given after the corollary, the restriction of A^* on S, $A^*|_S$, is a linear mapping. Using S_1 to denote the least invariant subspace of $A^*|_S$, we know the dimension of S_1 is one or two by the conclusion of linear algebra (or the explanation given before Lemma 3.2.2 in this book)

If $\dim S_1 = 1$, then S_1 is a real eigenspace of $A^*|_S$. It implies that there is for every $x_0 \in S_1$, $x_0 \neq 0$ we have $A^*x = \lambda x$. Furthermore, $\mathrm{e}^{-\lambda t}x_0 \in S_1 \cap S_{[0,\infty)}(-A^*, x_0)$. The fact implies $S_1 \subset Q$, hence, $Q \neq \{0\}$, i.e. $P \neq \mathbb{R}^n$.

If $\dim S_1 = 2$, then there exist $\alpha, \beta \in R$ and $x, y \in S_1$, $\|x\| = \|y\| = 1$ such that $A^*x = \alpha x - \beta y$ and $A^*y = \beta x + \alpha y$. By the conclusions established in the theory of differential equations, for $ax + by \in \operatorname{span}(x, y)$ where a and b are arbitrarily real numbers, we have $\in S_1 \cap S_{[0,\infty)}(-A^*, x_0)$ $\mathrm{e}^{-\alpha t}(a\cos\beta t - b\sin\beta t)x + \mathrm{e}^{-\alpha t}(b\cos\beta t + a\sin\beta t)y$. The fact also implies $S_1 \subset Q$, hence, $Q \neq \{0\}$, i.e. $P \neq \mathbb{R}^n$.

Sufficiency. It is sufficient to show $Q = \{0\}$ which is verified by contradiction. The proof contains two steps.

(1) We assume Q is not a proper set, i.e., Q contains a subspace of $\mathbb{R}^n$. Let S be the maximal subspace contained in Q, i.e., $S = Q \cap (-Q)$. Let $y_0^* \in S$. Then by the definition of Q, there exists a $y^*(t) \in S_{[0,\infty)}\left(-A^*, y_0^*\right)$. On the other hand, because $-y_0^* \in S$, there is also a $\overline{y}^*(t) \in S_{[0,\infty)}\left(-A^*, -y_0^*\right)$. Thus, by Lemmas 3.3.1 and 3.3.2, we have

$$y^*(t) + \overline{y}^*(t) \in S_{[0,\infty)}\left(-A^*, y_0^* - y_0^*\right) = S_{[0,\infty)}\left(-A^*, 0\right) = \{0\}.$$

The above relation implies the following facts:

(1) The set $S_{[0,\infty)}\left(-A^*, y_0^*\right)$ has only one element;
(2) $y^*(t) = -\overline{y}^*(t)$;
(3) For every $t \in [0,\infty)$, $y^*(t) \in Q$, then by the above fact ii, $y^*(t) \in S$.

Because for $t, t_0 \in [0,\infty)$, $y^*(t) \in S$ and $y^*\,(t_0) \in S$,

$$\frac{y^*(t) - y^*\,(t_0)}{t - t_0} \in S.$$

Since S is a linear subspace. S is closed; hence, when $t \to t_0$, we obtain $\dot{y}^*(t) \in S$. Especially, $\dot{y}^*(0) = -A^*\,(y^*(t))|_{t=0} \in S$, i.e., $A^*\left(y_0^*\right) \in S$. It leads to that S is an invariant subspace of A^*. We have gotten a contradiction. Therefore, Q is proper.

(2) We show that if $Q \neq \{0\}$, then Q has an eigenvector. Let k be an positive integer. We construct a convex process $A_k : \mathbb{R}^n \to \mathbb{R}^n$ as follows:

$$A_k\left(x^*\right) = R_{[0,\frac{1}{k}]}\left(-A^*, x^*\right),$$

where $R_{[0,\frac{1}{k}]}\,(-A^*, x^*)$ has been defined in Definition 2.4.1.

It is obvious $Q \subset \operatorname{dom} A_k$ and $A_k\,(x^*) \cap Q \neq \varnothing$ for every $x^* \in Q$. Because we have verified that Q is proper, by Lemma 3.2.1, there exists an eigenvalue $\lambda_k \geq 0$ and eigenvector $x_k \in Q$, $\|x_k\| = 1$ such that $\lambda_k x_k^* \in A_k\left(x_k^*\right) = R_{[0,\frac{1}{k}]}\left(-A^*, x_k^*\right)$. Suppose $x_k^*(t) \in S_{[0,1]}\left(-A^*, x_k^*\right)$ and $x_k^*\,(1/k) = \lambda_k x_k^*$, i.e.,

$$\lambda_k x_k^* = x_k^*\left(\frac{1}{k}\right) = x_k^* + \int_0^{1/k} \dot{x}_k^*(t)dt. \tag{3.3.12}$$

Because A is a bounded and closed process, by Theorem 3.2.3 (2) and Gronwall inequality (Lemma 2.3.1), we can conclude that both $\left\{x_k^*(t)\right\}$ and $\left\{\dot{x}_k^*(t)\right\}$ are bounded sets, it implies $\left\{x_k^*(t)\right\}$ is uniformly continuous. Therefore, $\left\{x_k^*(t)\right\}$ holds a convergent subseries. Without loss of generality, we assume $\left\{x_k^*(t)\right\}$ is convergent to $x^*(t)$ uniformly. Denote $x_k^* \to x_0$, then $\|x_0\| = 1$. By Theorem 2.4.4, we conclude $x^*(t) \in S_{[0,1]}\,(-A^*, x_0)$.

Given an $\varepsilon > 0$, there is a K when $k > K$, we have

$$-A^*\left(x_k^*(t)\right) \subset -A^*\left(x_0^*\right) + \varepsilon B, \quad \text{for} \quad t \in \left[0, \frac{1}{k}\right].$$

By Equation (3.3.12) and Lemma 2.4.1, we obtain

$$\lambda_k x_k^* = x_k^* + \int_0^{1/k} \dot{x}_k^*(t)dt \in x_k^* + \frac{1}{k}\left(-A^*\left(x_0^*\right) + \varepsilon B\right),$$

or it can be written as

$$k(\lambda_k - 1)x_k^* \in -A^*\left(x_0^*\right) + \varepsilon B. \tag{3.3.13}$$

The right side of Relation (3.3.13) is a bounded set, and $\left\|x_k^*\right\| = 1$; hence, the set $\{k(\lambda_k - 1)\}$ holds a convergent subseries. Without loss of generality, we can assume $k(\lambda_k - 1) \to \lambda_0$. Relation (3.3.13) then leads to $\lambda_0 x_0^* \in -A^*\left(x_0^*\right) + \varepsilon B$. Since ε is selected arbitrarily, we conclude $\lambda_0 x_0^* \in -A^*\left(x_0^*\right)$, i.e., A^* has an eigenvector which belongs to Q. This is a contradiction to the condition of theorem. The sufficiency is verified. □

Theorem 3.3.2 implies a conclusion that if $y^* \in -A^*(y^*)$ holds a solution whose domain can be extended to $[0, \infty)$, then dom A^* contains a subspace or A^* has an nontrivial eigenvector.

Problems

1. Prove that $\dot{y}_N^*(t)$ defined in the proof of Theorem 3.3.1 is bounded.
2. Prove that A_k defined in the proof of Theorem 3.3.3 is a closed convex process.
3. Suppose the set-valued mapping $A(x)$ is defined by $A(x) = Cx + K$, where $C \in \mathbb{R}^{n\times n}$ and $K \subset \mathbb{R}^n$ is closed and convex cone. Prove the following statements:

   (1) $A(x)$ is a closed and convex process;
   (2) $A^*(y^*) = \begin{cases} C^*y^*, & y^* \in K^*, \\ \varnothing, & y^* \notin K^*; \end{cases}$
   (3) The convex process $\dot{x} \in A(x)$ is controllable if and only if $C^* \in \mathbb{R}^{n\times n}$ has no eigenvector in K^*;
   (4) The convex process $\dot{x} \in A(x)$ is controllable if and only if there exists an integer $m \geq 1$ such that

$$K + CK + \cdots + C^m K = K - CK + \cdots + (-1)^m C^m K = \mathbb{R}^n.$$

(Note: If K is a subspace of $\mathbb{R}^n$, then there is a matrix D such that $K = D\mathbb{R}^n$. Moreover, $K + CK + \cdots + C^m K = \mathbb{R}^n$ is equivalent to rank $\left[C\ CD \cdots C^{n-1}D\right] = n$; the later is known as the controllability criterion of linear system $\dot{x} = Cx + Du$.)

3.4 Stability of Convex Process Differential Inclusions

This section deals with the stability of differential inclusions of convex processes. We mainly consider the weak stability. The necessary and sufficient conditions of the weak stability will be presented, and their Lyapunov functions will be constructed. Because the convex process is an extension of the linear mapping in theory of set-valued mapping, the concept of exponential stability can be extended to the convex process differential inclusions. Readers will find that the exponential stability plays a very important role in the stability theory of convex systems.

Let us start to recall the definition of stability of differential inclusions. Consider the Cauchy problem of differential inclusion described as follows:

$$\dot{x}(t) \in F\left(x(t)\right), \quad x(0) = x_0. \tag{3.4.1}$$

$S_{[0,T]}(F, x_0)$ is the set of solutions of Inc. (3.4.1). $[0, T]$ is the interval of time where the solutions exist. If $T = \infty$, the set $S_{[0,\infty)}\left(F, x_0\right)$ is simplified as $S(F, x_0)$ for convenience.

If $0 \in F(0)$, then the origin is an equilibrium of Inc. (3.4.1). From now on, we always assume $0 \in F(0)$, and the stability always considers for the equilibrium.

Inc. (3.4.1) is said to be (strong) stable if there is a function α which belongs to Class **K** such that for every $x(t) \in S\left(F, x_0\right)$, we have

$$\|x(t)\| \leq \alpha\left(\|x_0\|\right). \tag{3.4.2}$$

Inc. (3.4.1) is said to be (strong) asymptotically stable if there is a function β in Class **KL** such that for every $x(t) \in S\left(F, x_0\right)$, we have

$$\|x(t)\| \leq \beta\left(\|x_0\|, t\right). \tag{3.4.3}$$

Inc. (3.4.1) is said to be weakly stable if there is a function α in Class **K** and there exists at least one $x(t) \in S\left(F, x_0\right)$ such that

$$\|x(t)\| \leq \alpha\left(\|x_0\|\right).$$

It is said to be weakly asymptotically stable if there is a function β in Class **KL** and there exists at least one $x(t) \in S\left(F, x_0\right)$ such that

$$\|x(t)\| \leq \beta\left(\|x_0\|, t\right).$$

Inc. (3.4.1) is said to be unstable if it is not weakly stable.

It has seen that for differential inclusions there are “strongly stable”, “weakly stable” and “unstable” different definitions. The difference between “strong” and “weak” is “all” and “one”, and the “unstable” is the negative of “weakly stable”.

3.4.1 Stability of Convex Processes

Consider the following differential inclusion

$$\dot{x}(t) \in A\left(x(t)\right), \tag{3.4.4}$$

where A: $\mathbb{R}^n \to \mathbb{R}^n$ is a strict, closed and convex process. For the sake of convenience, Inc. (3.4.4) is called as convex system.

The dual system of Inc. (3.4.4) is defined as

$$\dot{y}^*(t) \in -A^*\left(y^*(t)\right), \tag{3.4.5}$$

where $A^*\colon \mathbb{R}^n \to \mathbb{R}^n$ is the adjoint process of A. Inc. (3.4.5) is also called as adjoint system of Inc. (3.4.4), sometimes.

Because $0 \in A(0)$ and $0 \in -A^*(0)$, the origin is always the equilibrium of both Incs. (3.4.4) and (3.4.5).

From Theorem 3.3.2, we can find that if Inc. (3.4.4) is controllable then it is weakly asymptotically stable. Therefore, the weakly asymptotic stability is a weaker concept than the controllability. The controllability is somewhat similar to the finite time stability in the theory of differential equations.

Let S be the maximal A^*-invariant subspace contained in $\operatorname{dom} A^* \cap (-\operatorname{dom} A^*)$, and $T = S^\perp$. By Lemma 3.2.2, T is the minimal A-invariant subspace. Because $\mathbb{R}^n = S \oplus T$, for every $x \in \mathbb{R}^n$ there exist uniquely $x_S \in S,\ x_T \in T$ such that $x = x_S + x_T$. Consequently, sometimes we denote $x^T = \left[x_S^T\ x_T^T\right]$ by treating $[x_S^T\ \ 0]^T$ as x_S and $[0\ \ x_T^T]^T$ as x_T. We have mentioned that two norms in a finite space are always equivalent (Theorem 1.1.1), hence the norm of x takes that $\|x\| = \max\left\{\|x_S\|, \|x_T\|\right\}$ in this section, and both $\|x_S\|$ and $\|x_T\|$ are their Euclidean norms, i.e., $\|x_S\| = \sqrt{x_S^T x_S}$ and $\|x_T\| = \sqrt{x_T^T x_T}$. By Theorem 3.2.6 and the illustration after the theorem, $\bar{A}$, the induced mapping of A on $\mathbb{R}^n/T$, is a simple-valued linear mapping, and $\left(\overline{A}\right)^*$ can be treated to be identical to the restriction of A^* on S. When A is restricted on T, i.e., $A|_T$, has a Jordan-like construction. Denote $L_k(\lambda) = (A - \lambda I)^{-k}(0)$. When $S = \{0\}$ and $\lambda > \lambda_M(A^*)$, where $\lambda_M(A^*)$ is the maximal eigenvalue of A^* (we have stipulated $\lambda_M(A^*) = -\infty$ when A^* has no eigenvalue), $L_k(\lambda) = \mathbb{R}^n$ for some integer k. If $S \neq \{0\}$, the conclusion is still valid for the restriction of A on T, i.e., $T = \left(A|_T - \lambda I|_T\right)^{-k}(0)$ for some constants k and λ.

Theorem 3.4.1 If $S = \{0\}$ and $0 > \lambda_M(A^*)$, the convex system Inc. (3.4.4) is weakly asymptotically stable.

Proof Because $0 > \lambda_M(A^*)$, we can select an λ such that $0 > \lambda > \lambda_M(A^*)$. By the statement above, we conclude that $\mathbb{R}^n = (A - \lambda I)^{-k}(0)$ for this λ. It follows that for every $x_0 \in \mathbb{R}^n$, there exist $y_k, y_{k-1}, \ldots, y_1 \in \mathbb{R}^n$ such that

$$x_0 = y_k \in (A - \lambda I)^{-1}(y_{k-1}), y_{k-1} \in (A - \lambda I)^{-1}(y_{k-2}), \ldots, y_1 \in (A - \lambda I)^{-1}(0).$$

We now construct a function $x(t)$ as follows

$$x(t) = \mathrm{e}^{\lambda t}\left(\frac{t^{k-1}}{(k-1)!}y_1 + \frac{t^{k-2}}{(k-2)!}y_2 + \cdots + y_k\right). \tag{3.4.6}$$

It is obvious that $x(t)$ can define on $[0, \infty)$. Taking the derivate of $x(t)$, we obtain

$$\begin{aligned}
\dot{x}(t) &= \lambda e^{\lambda t}\left(\frac{t^{k-1}}{(k-1)!}y_1 + \frac{t^{k-2}}{(k-2)!}y_2 + \cdots + y_k\right) \\
&\quad + \mathrm{e}^{\lambda t}\left(\frac{t^{k-2}}{(k-2)!}y_1 + \frac{t^{k-3}}{(k-3)!}y_2 + \cdots + y_{k-1}\right) \\
&= e^{\lambda t}\left(\frac{t^{k-1}}{(k-1)!}\lambda y_1 + \frac{t^{k-2}}{(k-2)!}\left(y_1 + \lambda y_2\right) + \cdots + \left(y_{k-1} + \lambda y_k\right)\right) \\
&\in e^{\lambda t}\left(\frac{t^{k-1}}{(k-1)!}A\left(y_1\right) + \frac{t^{k-2}}{(k-2)!}A\left(y_2\right) + \cdots + A\left(y_k\right)\right) \\
&\quad + \subset A\left(e^{\lambda t}\left(\frac{t^{k-1}}{(k-1)!}y_1 + \frac{t^{k-2}}{(k-2)!}y_2 + \cdots + y_k\right)\right) \\
&= A\left(x(t)\right).
\end{aligned}$$

And $x(0) = y_k = x_0$. Consequently, $x(t) \in S_{[0,\infty)}\left(A, x_0\right)$ and $x(t) \to 0\ (t \to \infty)$. □

If $S \neq \{0\}$, Theorem 3.4.1 should be revised as follows.

Corollary 3.4.1 If $0 > \lambda > \lambda_M\left(A^*\right)$ and $x_0 \in L_k(\lambda)$, then there exists an $x(t) \in S\left(A, x_0\right)$, such that $\lim\limits_{t\to\infty} x(t) = 0$. □

The proof of Corollary 3.4.1 is exactly the same as that of Theorem 3.4.1 and omitted.

Theorem 3.4.1 and Corollary 3.4.1 provide conclusions which are very similar to those for differential equations.

In the theory of differential equations, the exponential convergence is a faster convergence than asymptotic convergence. We now extend this concept to differential inclusions.

Definition 3.4.1 Inc. (3.4.1) is exponentially stable, if for every x_0 and every $x(t) \in S\left(F, x_0\right)$, there exist constants $\alpha > 0$ and $\lambda > 0$ such that $\|x(t)\| \leq \alpha \left\|x_0\right\| e^{-\lambda t}$.

Inc. (3.4.1) is weakly exponentially stable, if for every x_0 there exist $\alpha > 0$ and $\lambda > 0$, and $x(t) \in S\left(F, x_0\right)$ such that $\|x(t)\| \leq \alpha \left\|x_0\right\| e^{-\lambda t}$. □

The following lemma extends the conclusion of Theorem 3.4.1.

Lemma 3.4.1 If Inc. (3.4.4) is weakly asymptotically stable, then it is weakly exponentially stable.

Proof We use bd$B$ to denote the shell of closed unit ball of $\mathbb{R}^n$. Because the unit ball can be contained in a simplex, there exist $y_1, y_2, \ldots, y_{n+1} \in \mathbb{R}^n$ such that bd $B \subset$ conx$\left(y_1, y_2, \ldots, y_{n+1}\right)$. Consequently, for every $y \in$ bd B, there exist $\lambda_1, \lambda_2, \ldots, \lambda_{n+1} \in [0, 1]$ and $1 = \sum\limits_{i=1}^{n+1} \lambda_i$ such that $y = \sum\limits_{i=1}^{n+1} \lambda_i y_i$.

Inc. (3.4.4) is weakly asymptotically stable, hence, for every i, $i \in \{1, 2, \ldots, n+1\}$, there exists an $x_i(t) \in S\left(A, y_i\right)$ such that $\lim\limits_{t\to\infty} x_i(t) = 0$. Further-

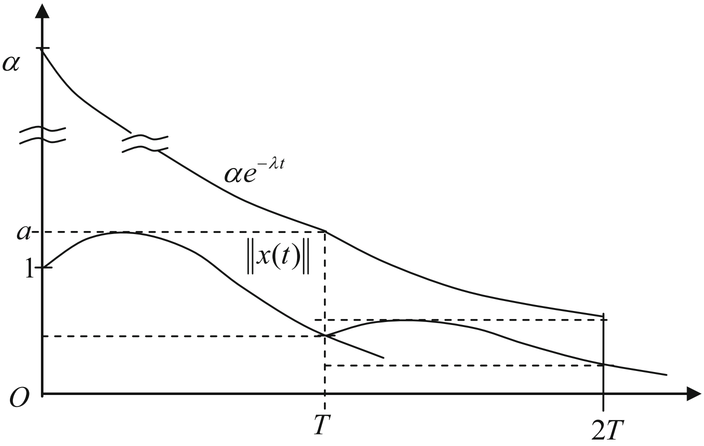


Fig. 3.2 The feature of $\|x(t)\|$

more, there is a $T > 0$ such that $\|x_i(T)\| \le a\mathrm{e}^{-1}$ for every i, $\quad i \in \{1, 2, \dots, n+1\}$ where the constant a is defined as $a = \max\{\|x_i(t)\|, t \in [0, \infty), i = 1, 2, \dots, n+1\}$.

Define $x(t) = \sum_{i=1}^{n+1} \lambda_i x_i(t)$, then by Lemma 3.3.1, we have $x(t) \in S(A, y)$. Moreover, the $x(t)$ satisfies that (1) $\lim_{t\to\infty} x(t) = 0$, (2) $\|x(T)\| \le a\mathrm{e}^{-1}$, (3) $\|x(t)\| \le a,\ t \in [0, T]$.

Let us denote $\lambda = T^{-1}$ and $\alpha = ae$, where the α is independent of $x(t)$ and T. Then $\|x(t)\| \le \alpha e^{-\lambda t}$ as $t \in [0, T]$. Furthermore, if we define

$$y(t) = x(mT)x(t - mT), \quad t \in [mT, (m+1)T) \quad m = 0, 1, 2, \dots,$$

then $y(t) \in S(A, y)$ and $\|y(t)\| \le \alpha\mathrm{e}^{-\lambda t}$ (Fig. 3.2).

For arbitrary $x_0 \in \mathbb{R}^n$, $y = x_0\|x_0\|^{-1} \in \operatorname{bd} B$. There is a $y(t) \in S(A, y)$ such that $\|y(t)\| \le \alpha e^{-\lambda t}$. By Lemma 3.3.1, $x(t) = \|x_0\|\, y(t) \in S(A, x_0)$, then $\|x(t)\| \le \alpha\,\|x_0\|\, e^{-\lambda t}$. It completes the proof. □

Lemma 3.4.1 reveals the uniformness of weakly asymptotic stability of convex system. We have obtained α and λ which are independent of the initial condition. In the proof of Lemma 3.4.1, we have applied the features of finite dimension linear space and convex process sufficiently. The lemma will play an important role in the following discussion.

Theorem 3.4.2 Suppose A is a strict, closed and convex process. The following statements are equivalent.

(1) Inc. (3.4.4) is weakly asymptotically stable.
(2) The dual Inc. (3.4.5) is nowhere stable, i.e., for every $y_0^* \in \mathbb{R}^n$, $y_0^* \ne 0$ then every solution $y^*(t) \in S\left(-A^*, y_0^*\right)$, $y^*(t)$ does not hold a bounded subsequence $\{y^*(t_k)\}$.

(3) $\lambda_M(A^*) < 0$, moreover, if $S \neq \{0\}$, then the restriction of A^* on S, i.e. $A^*|_S$ is a Hurwitz matrix.

Proof (1) ⇒ (2). Suppose Inc. (3.4.4) is weakly asymptotically stable, i.e., for every $x_0 \in \mathbb{R}^n$, there is an $x(t) \in S(A, x_0)$ such that $x(t) \to 0$ $(t \to \infty)$. If there are a y_0^* and a solution $y^*(t) \in S(-A^*, y_0^*)$ which has a bounded subsequence $\{y^*(t_k)\}$. Consider now

$$\begin{aligned}\langle y^*(t), x(t)\rangle &= \langle y_0^*, x_0\rangle + \int_0^t d\langle y^*(s), x(s)\rangle \\ &= \langle y_0^*, x_0\rangle + \int_0^t \left(\langle \dot{y}^*(s), x(s)\rangle + \langle x^*(s), \dot{x}(s)\rangle\right) ds.\end{aligned}$$

By the definition of adjoint process, we have

$$\langle \dot{y}^*(t), x(t)\rangle + \langle y^*(t), \dot{x}(t)\rangle \geq 0.$$

It follows that $\langle y^*(t), x(t)\rangle \geq \langle y_0^*, x_0\rangle$. We now restrict the inequality on the sequence $\{y^*(t_k)\}$. Taking the limitation as $t \to \infty$, the inequality yields $0 \geq \langle y_0^*, x_0\rangle$. The x_0 can be selected arbitrary, hence the inequality $0 \geq \langle y_0^*, x_0\rangle$ implies $y_0^* = 0$. The result illustrate that only $y^*(t) \in S(-A^*, 0)$ has a bounded sequence.

Furthermore, because $\dot{y}^*(t) \in -A^*(y^*(t))$, from Theorem 3.2.3 (2), we obtain $\|\dot{y}^*(t)\| \leq \|A\| \, \|y^*(t)\|$. Using the Gronwall inequality, it leads to

$$\|y^*(t)\| \leq e^{\|A\|t} \|y^*(0)\| = 0.$$

The fact illustrates that a bounded solution of adjoint process $\dot{y}^*(t) \in -A^*(y^*(t))$ is only the zero solution. Hence, the solution of $\dot{y}^*(t) \in -A^*(y^*(t))$ is nowhere stable, i.e., when $y_0^* \neq 0$ every solution $y^*(t) \in S(-A^*, y_0^*)$ is divergent.

(2) ⇒ (3). We still use method of contradiction.

If A^* holds an nonnegative eigenvalue $\lambda \geq 0$, and the responding eigenvector is $y_0^* \in \mathbb{R}^n$, $y_0^* \neq 0$, then $y(t) = e^{-\lambda t} y_0^* \in S_{[0,\infty)}(-A^*, y_0^*)$. $e^{-\lambda t} y_0^*$ is bounded. We have a contradiction.

If $\dim S = \upsilon \neq 0$, then the restriction of A^* on S, $A^*|_S$, can expressed as a $\upsilon \times \upsilon$ matrix A_S^*. If $A^*|_S$ is not asymptotically stable. Let λ be the eigenvalue of A_S^*. Then Re $\lambda \geq 0$. The above discussion shows that the λ cannot be real. Hence, it is a complex number. Let $\lambda = \alpha + \mathrm{j}\beta$ with $\alpha \geq 0$, then $\lambda = \alpha - \mathrm{j}\beta$ is also an eigenvalue of A_S^*. Their responding eigenvectors are $u_0^* \pm \mathrm{j} v_0^*$, $u_0^*, v_0^* \in S \subset \mathbb{R}^n$. By the theory of differential equations, the following facts are true:

$$y^*(t) = e^{-\alpha t}\left(u_0^* \cos \beta t - v_0^* \sin \beta t\right) \in S_{[0,\infty)}\left(-A^*, u_0\right),$$

Table 3.1 The sequence of yielding $\{x_{kT}(t)\}$ and $\{x_k(t)\}$

$\mathbb{R}^n$	x_0	$x_0(t) \rightarrow\rightarrow$	$x_0(\tau) = x_1$	$x_1(t) \rightarrow\rightarrow$	$x_1(\tau) = x_2$
	$\downarrow$ Projection	$\uparrow$ Th. 2.3.3	$\downarrow$ Projection	$\uparrow$ Th. 2.3.3	$\downarrow$ Projection
T	$x_{0T} \rightarrow\rightarrow$	$x_{0T}(t)$	$x_{1T} \rightarrow\rightarrow$	$x_{1T}(t)$	$x_{2T} \rightarrow\rightarrow x_{2T}(t)$

and

$$y^*(t) = e^{-\alpha t}(v_0 \cos \beta t + u_0 \sin \beta t) \in S_{[0,\infty)}\left(-A^*, v_0\right).$$

They are both bounded even for the case of $\alpha = 0 \quad (\text{Note}: \quad \beta \neq 0)$. We also obtain a contradiction.

(3) $\Rightarrow$ (1). The proof of this part is constructive and consists of two steps.

(i) Constructing an $x(t) \in S_{[0,\infty)}(A, x_0)$.

Because $\lambda_M(A^*) < 0$, by Corollary 3.4.1 and Lemma 3.4.1, for every $x_0 \in T$,[5] there is an $x(t) \in S_{[0,\infty)}(A, x_0)$ such that $\|x(t)\| \leq \alpha \|x_0\| e^{-\lambda t}$, where α and λ are positive numbers and independent of x_0 and t. We can find a positive number τ such that $\alpha e^{-\lambda\tau} < 1$.

For every $x_0 \in \mathbb{R}^n$, $x_0 = x_{0S} + x_{0T}$ is an orthogonal decomposition of x_0 where $x_{0S} \in S$ and $x_{0T} \in T$. For this x_{0T}, we have an weakly exponentially stable solution $x_{0T}(t) \in S_{[0,\infty)}(A, x_{0T})$ such that $\|x_{0T}(t)\| \leq \alpha \|x_{0T}\| e^{-\lambda t}$. Of course, $x_{0T}(t) \in S_{[0,\tau]}(A, x_{0T})$.

Let $M = \{x_{0T}(t)\}$ and $r_0(x) \equiv x_0$. Then by Theorem 2.3.3, there is an $x_0(t) \in S_{[0,\tau]}(A, x_0)$ which satisfies that $\|x_0(t) - x_{0T}(t)\| \leq \|x_0 - x_{0T}\| e^{\|A\|t}$, where we have applied $\|A\|$ as the Lipschitz constant of A.

Let $x_1 = x_0(\tau)$. Then by replacing x_0 by x_1 and repeating the above procedure, we can obtain an $x_{1T}(t) \in S_{[0,\tau]}(A, x_{1T})$ and $x_1(t) \in S_{[0,\tau]}(A, x_1)$. Fixing $x_2 = x_1(\tau)$ and repeating the above procedure, we can obtain $x_{2T}(t) \in S_{[0,\tau]}(A, x_{2T})$ and $x_2(t) \in S_{[0,\tau]}(A, x_2)$, ... Thus, we can obtain two series $\{x_{kT}(t)\}$ and $\{x_k(t)\}$. The Table 3.1 presents the procedure of construction of the two series.

The $\{x_k(t)\}$ and $\{x_{kT}(t)\}$ satisfy the following relations:

(i) $x_k(\tau) = x_{k+1} = x_{k+1}(0); x_k = x_{kS} + x_{kT}$;

(ii) $x_{kT}(t) \in S_{[0,\tau]}(A, x_{kT})$ and $x_k(t) \in S_{[0,\tau]}(A, x_k)$,
where $x_{kT}(t) \in S_{[0,\tau]}(A, x_{kT})$ is a weakly exponentially stable solution for initial condition x_{kT}, it satisfies $\|x_{kT}(t)\| \leq \alpha \|x_{kT}\| e^{-\lambda t}$;

(iii) $\|x_{kT}(t) - x_k(t)\| \leq e^{\|A\|t} \|x_{kT}(0) - x_k(0)\| = e^{\|A\|t} \|x_{kT} - x_k\| = e^{\|A\|t} \|x_{kS}\|$, especially,

$$\|x_{kT}(\tau) - x_k(\tau)\| \leq e^{\|A\|\tau} \|x_{kT} - x_k\| = e^{\|A\|\tau} \|x_{kS}\|. \tag{3.4.7}$$

[5] The T used is a subspace and equal to $S^{\perp}$.

Let

$$x(t) = x_k\,(t - k\tau)\,,\quad t \in [k\tau, (k+1)\,\tau)\,.$$

By the property (i), $x(t)$ is continuous, and moreover, $x(t) \in S_{[0,\infty)}\,(A, x_0)$.

(ii) Verifying $x(t) \to 0\,(t \to \infty)$.

Let P_T be the projection from $\mathbb{R}^n$ to T,[6] and P_S the projection from $\mathbb{R}^n$ to S. Since $\mathbb{R}^n/T$ is isomorphic to S, in this section, P_S is also used as the projection form $\mathbb{R}^n$ to $\mathbb{R}^n/T$. Denote $P_S x(t) = x_S(t)$ and $P_T x(t) = x_T(t)$, then $x(t) = x_T(t) + x_S(t)$. By the norm defined at the beginning of this section, we can have $\|x(t)\| = \max\{\|x_S(t)\|, \|x_T(t)\|\}$. Hence, if $x_S(t) \to 0\;\;(t \to \infty)$ and $x_T(t) \to 0\;\;(t \to \infty)$, then $x(t) \to 0\;\;(t \to \infty)$.

T is an A -invariant subspace, then, A yields an induced mapping $\bar{A}$ on $\mathbb{R}^n/T$. The induced mapping satisfies $P_S A = \overline{A} P_S$. It follows that

$$\dot{x}_S(t) = P_S \dot{x}(t) \in P_S A\,(x(t)) = \overline{A} P_S\,(x(t)) = \overline{A} x_S(t).$$

By Lemma 3.2.5, $\bar{A}$ is a single-valued linear mapping, its eigenvalues are equal to those of $\left(\bar{A}\right)^*$ or $A^*|_S$. By the condition of (3), $\bar{A}$ is asymptotically stable. There exist positive numbers β and μ, which are independent of $x_S(0)$ and t, such that

$$\|x_S(t)\| \le \beta\,\|x_S(0)\|\,e^{-\mu t}, \tag{3.4.8}$$

for every $x_S(0) \in S$.

Hence, it is sufficient to show $x_T(t) \to 0\;\;(t \to \infty)$.

By the definition of $x(t)$, $x\,((k-1)\,\tau) = x_k$, hence, $x_T\,((k-1)\,\tau) = x_{kT}$. By the definition of norm, we have $\left\|x_{kT} - x_{(k-1)T}\,(\tau)\right\| \le \left\|x_k - x_{(k-1)T}\,(\tau)\right\|$. From the definition of x_k, we obtain $\left\|x_k - x_{(k-1)T}\,(\tau)\right\| = \left\|x_{k-1}\,(\tau) - x_{(k-1)T}\,(\tau)\right\|$. Using Inequality (3.4.7),

$$\left\|x_k - x_{(k-1)T}\,(\tau)\right\| = \left\|x_{k-1}\,(\tau) - x_{(k-1)T}\,(\tau)\right\| \le e^{\|A\|\tau}\,\left\|x_{(k-1)S}\right\|.$$

Hence, x_{kT} can be estimated by

$$\begin{aligned} \|x_{kT}\| &\le \left\|x_{kT} - x_{(k-1)T}\,(\tau)\right\| + \left\|x_{(k-1)T}\,(\tau)\right\| \\ &\le \left\|x_k - x_{(k-1)T}\,(\tau)\right\| + \left\|x_{(k-1)T}\,(\tau)\right\| \\ &\le \left\|x_{(k-1)S}\right\|\,e^{\|A\|\tau} + \alpha e^{-\lambda\tau}\,\left\|x_{(k-1)T}\right\|. \end{aligned} \tag{3.4.9}$$

[6] Note, the meaning of P_T is different from the controllability cone defined at the last section.

The second term comes from the step (i), i.e., $x_{(k-1)T}(t)$ is a weakly stable solution. For the sake of convenience, we denote $f = \alpha e^{-\lambda\tau}$ and $g = e^{(\|A\|+\mu)\tau}$, then $f < 1$ by the selection of τ. From Inequalities (3.4.8) and (3.4.9), we have

$$\|x_{kT}\| \leq \|x_{0S}\| \, e^{\|A\|\tau-\mu(k-1)\tau} + f \left\|x_{(k-1)T}\right\| = g \, \|x_{0S}\| \, e^{-\mu k\tau} + f \left\|x_{(k-1)T}\right\| .$$

Thus,

$$\begin{aligned}
\|x_{kT}\| &\leq g \, \|x_{0S}\| \, e^{-\mu k\tau} + f \left\|x_{(k-1)T}\right\| \\
&\leq g \, \|x_{0S}\| \, e^{-\mu k\tau} + f \left(g \, \|x_{0S}\| \, e^{-\mu(k-1)\tau} + f \left\|x_{(k-2)T}\right\|\right) \\
&\leq g \, \|x_{0S}\| \, e^{-\mu k\tau} + fg \, \|x_{0S}\| \, e^{-\mu(k-1)\tau} + f^2 \left(g \, \|x_{0S}\| \, e^{-\mu(k-2)\tau} + f \left\|x_{(k-3)T}\right\|\right) \\
&\quad \cdots\cdots \\
&\leq g \, \|x_{0S}\| \, e^{-\mu k\tau} + fg \, \|x_{0S}\| \, e^{-\mu(k-1)\tau} \cdots + f^{k-1} g \, \|x_{0S}\| \, e^{-\mu\tau} + f^k \, \|x_{0T}\| \\
&= g \, \|x_{0S}\| \frac{f^k - \left(e^{-\mu\tau}\right)^k}{f - e^{-\mu\tau}} + f^k \left(\|x_{0T}\| - g \, \|x_{0S}\|\right) .
\end{aligned}$$

Because $f < 1, \;\; e^{-\mu\tau} < 1, x_{kT} \to 0 \;\; (k \to \infty)$.

When $t \in [k\tau, (k+1)\,\tau)$, we have

$$\begin{aligned}
\|x_T(t)\| &\leq \|x_T(t) - x_{kT}\,(t-k\tau)\| + \|x_{kT}\,(t-k\tau)\| \\
&\leq \|x(t) - x_{kT}\,(t-k\tau)\| + \|x_{kT}\,(t-k\tau)\| \\
&\leq \|x_{kS}\,(t-k\tau)\| + \|x_{kT}\,(t-k\tau)\| \\
&\leq \|x_{kS}\| \, e^{\|A\|\tau} + f \, \|x_{kT}\| .
\end{aligned}$$

Therefore, $\bar{x}_T(t) \to 0 \, (k \to \infty)$. □

From Theorem 3.4.2, we can obtain the following corollary for the unstability. The proof is left to readers as an exercise.

Theorem 3.4.3 Suppose A is a strict, closed and convex process. If the restriction of A^* on S is an unstable linear single-valued mapping, i.e. $A^*|_S$ is not a Hurwitz matrix, or $\lambda_M\,(A^*) > 0$, then A is unstable.

3.4.2 Construction of Lyapunov Functions

It is difficult for us to find an analytical solution of a differential equation. Hence, researchers presented the qualitative theory of differential equations. In the theory, Lyapunov direct method is very powerful and widely used. An evident advantage is that it can discriminate the stability without solving the equation. Theorem 2.5.5, given at the last chapter, presents a criterion for the weakly stable of differential inclusions. It illustrates that if there exist a positive definite function $V : \mathbb{R}^n \to \mathbb{R}$, a semi-negative definite function $W : \mathbb{R}^n \to \mathbb{R}$, and a vector $v \in \mathbb{R}^n$ such that $D^- V(x)(v) \leq W(x)$ then the differential inclusion is weakly asymptotically stable. By Lemma 3.4.1, for a convex process, the solution is exponential stable. It is regretful that we still lack a usable method to construct Lyapunov functions. However, for convex system, we have a method to construct a Lyapunov function.

We now consider convex system Inc. (3.4.4). If there is a positive definite and convex function $V : \mathbb{R}^n \to \mathbb{R}$, $v \in A(x)$, and a positive real number $\delta \in (0,\ 1)$ such that inequality

$$V(x + \tau v) \leq \delta V(x) \tag{3.4.10}$$

holds for a positive real number τ, then $V(x)$ is a Lyapunov function of Inc. (3.4.4). The fact is verified as follows. Because $V(x)$ is a convex function, By Theorem 1.3.4, the following function decreases

$$\frac{V(x + \tau v) - V(x)}{\tau}$$

as τ decreases. Hence, we obtain

$$DV(x)(v) \leq \frac{V(x + \tau v) - V(x)}{\tau} \leq \frac{\delta - 1}{\tau} V(x). \tag{3.4.11}$$

Let $\theta = \frac{1-\delta}{\tau}$. Then $\theta > 0$ and $DV(x)(v) \leq -\theta V(x)$.

The next theorem gives the existence of Lyapunnov function for the convex system Inc. (3.4.4). We give there exists a positive function $V : \mathbb{R}^n \to \mathbb{R}$ to satisfy Inequality (3.4.10).

Theorem 3.4.4 If $A^*|_S$ is asymptotically stable and $\lambda_M(A^*) < 0$, then there exists a positive function $V : \mathbb{R}^n \to \mathbb{R}$, positive number $\delta \in (0, 1)$ and vector $v \in A(x)$ such that $V(x + \tau v) \leq \delta V(x)$.

Proof The proof consists of three steps.

(1) Construction of positive definite function $V : \mathbb{R}^n \to \mathbb{R}$.

Let $\bar{A}$ be the induced mapping of A on the quotient space $\mathbb{R}^n/T$. Then the condition of theorem guarantee $\bar{A}$ is an asymptotically stable matrix, i.e., it can be expressed as a Hurwitz matrix. Hence the Lyapunov equation $\overline{A}^T P + P\overline{A} = -I$ has a positive definite solution P such that $V_S(x_S) = x_S^T P x_S$ is a Lyapunov function of differential equation $\dot{x}_S = \overline{A} x_S$. Denote

$$M_S = \{x_S; x_S \in S,\ V_S(x_S) \leq 1\}.$$

Denote B for the open unit ball in $\mathbb{R}^n$, then $B_T = B \cap T$ is the unit ball of T. If $\dim T = m$, then there exists a simplex $\operatorname{co}\{y_1, y_2, \ldots, y_{m+1}\}$ such that $B_T \subset \operatorname{conx}\{y_1, y_2, \ldots, y_{m+1}\} \subset T$.

Let us fix a real number λ such that $\lambda_M(A^*) < \lambda < 0$. By Theorem 3.2.5, there is a positive integer k_0 such that $T = (A - \lambda I)^{-k_0}(0)$. Denote $L_k = (A - \lambda I)^{-k}(0)$ for every $1 \le k \le k_0$. By the property of L_k, for every y_i, $1 \le i \le m+1$, there exists a k_i such that $y_i \in L_{k_i}$, $y_i \notin L_{k_i - 1}$. By Theorem 3.2.5, for every y_i, there exist $y_i = y_i^{k_i}, y_i^{k_i - 1}, \ldots, y_i^1 \in T$ such that

$$\begin{aligned} &\lambda y_i^1 \in A\left(y_i^1\right), \\ &y_i^1 + \lambda y_i^2 \in A\left(y_i^2\right), \\ &\cdots\cdots \\ &y_i^{k_i - 1} + \lambda y_i^{k_i} \in A\left(y_i^{k_i}\right). \end{aligned} \tag{3.4.12}$$

A constant α is selected such that $\alpha > \max\{1, |\lambda|\}$. Define

$$z_i^j = \left(\frac{\alpha}{|\lambda|}\right)^{k_i - j} y_i^j, \tag{3.4.13}$$

and

$$M_T = \text{clco}\left\{z_i^j; j = 1, 2, \ldots, k_i, i = 1, 2, \ldots, m+1\right\}.$$

Then $M_T \supset \text{co}\,(y_1, y_2, \ldots, y_{m+1}) \supset B_T$. Let ω be a positive constant, and denote

$$M_\omega = M_T \oplus \omega M_S \subset T \oplus S = \mathbb{R}^n.$$

M_ω is bounded, hence, it is compact no matter the selection of ω. We now define $V(x)$ to be the Minkovski function of M_ω.[7]

(2) Properties of M_S and M_T.

Let $x_S \in \text{bd}M_S \subset S$. We now prove there exist positive numbers τ_S and δ_S with $0 < \delta_S < 1$ and $x_S + \tau_S \overline{A} x_S \in \delta_S M_S$. The two constants are independent of x_S.

Consider

$$\begin{aligned} V_S\left(x_S + \tau \overline{A} x_S\right) &= \left(x_S + \tau \overline{A} x_S\right)^T P \left(x_S + \tau \overline{A} x_S\right) \\ &= x_S^T P x_S - \tau x_S^T x_S + \tau^2 x_S^T \overline{A}^T P \overline{A} x_S \\ &= 1 - \tau x_S^T x_S + \tau^2 x_S^T \overline{A}^T P \overline{A} x_S \ , \end{aligned}$$

Since $\overline{A}$ is asymptotically stable, $\overline{A}$ is nonsingular. $\overline{A}^T P \overline{A}$ is then positive definite. If $\tau = \frac{x_S^T x_S}{x_S^T \overline{A}^T P \overline{A} x_S}$, then $V\left(x_S + \tau \overline{A} x_S\right) = 1$, and if $0 < \tau <$

[7] Minkovski function $\mu(x, A)$ has been defined in Sect. 1.3. In that section $\mu(x, A)$ has been proved to be positive definite and homogenous.

$\frac{x_S^T x_S}{x_S^T \overline{A}^T P \overline{A} x_S}$, then $V\left(x_S + \tau \overline{A} x_S\right) < 1$. Hence, $\tau_S = \min\limits_{x_S \in \mathrm{bd} M_S} \frac{x_S^T x_S}{2 x_S^T \overline{A}^T P \overline{A} x_S}$ and $\delta_S = \max\limits_{x_S \in \mathrm{db} M_S} V\left(x_S + \tau_S \overline{A} x_S\right) < 1$ meet with the requirement.

If $x_S \in \mathrm{int} M_S$ then there is $a > 1$ such that $a x_S \in \mathrm{bd}\, M_S$. The above proof implies $a x_S + \tau_S \overline{A} a x_S \in \delta_S M_S$. Thus, $a x_S + \tau_S \overline{A} a x_S \in \delta_S M_S$ and $x_S + \tau_S \overline{A} x_S \in a^{-1} \delta_S M_S \subset \delta_S M_S$. Therefore, we conclude that for every $x_S \in M_S$, $x_S + \tau_S \overline{A} x_S \in \delta_S M_S$.

Now let $x_T \in M_T \subset T$. It will be verified that there is a $v \in A\left(\left(0 \;\; x_T^T\right)^T\right)$ and positive numbers τ_T and δ_T where $0 < \delta_T < 1$, such that $x_T + \tau_T v \in \delta_T M_T$. The two constants τ_T and δ_T are also independent of x_T.

For $i = 1, 2, \ldots, m + 1$, let us define

$$\begin{aligned} v_i^1 &= \lambda z_i^1, \\ v_i^j &= \tfrac{|\lambda|}{\alpha} z_i^{j-1} + \lambda z_i^j, \quad j = 2, 3, \cdots k_i. \end{aligned}$$

From Relation (3.4.12) and Eq. (3.4.13), we can find $v_i^j \in A\left(z_i^j\right)$ and

$$\begin{aligned} |\lambda|^{-1} v_i^1 + z_i^1 &= 0 \in \mathrm{int}\, M_T, \\ |\lambda|^{-1} v_i^j + z_i^j &= \tfrac{1}{\alpha} z_i^{j-1} \in \mathrm{int}\, M_T, \quad j = 2, 3, \ldots k_i. \end{aligned}$$

If $x_T \in M_T$, then there exist $\lambda_i^j \in [0, 1]$, $1 \le i \le m + 1$, $1 \le j \le k_i$ such that $x_T = \sum\limits_{i,j} \lambda_i^j z_i^j$. Correspondingly, we obtain $v_T = \sum\limits_{i,j} \lambda_i^j v_i^j$. It follows that $v_T \in A\left(x_T\right)$ and

$$x_T + |\lambda|^{-1} v_T = \alpha^{-1} \sum_{i,j} \lambda_i^j z_i^{j-1} \in \alpha^{-1} M_T,$$

where $z_i^{-1} = 0$. Taking $\tau_T = |\lambda|^{-1}$ and $\delta_T = \alpha^{-1}$, the conclusion is followed.

(3) $V(x)$ can satisfy Inequality (3.4.10) by selecting ω.

Because the Minkovski function is positive and homogeneous, it is sufficient for points on boundary of M_ω to check Inequality (3.4.10).

Suppose $\left[x_T^T \;\; x_S^T\right]^T \in \mathrm{bd} M_\omega \subset M_T \oplus \omega M_S$, where $x_S \in \omega M_S$, $x_T \in M_T$.

If we repeat the discussion for M_S, then we can conclude that there exist τ_S and δ_S which are independent of x_S such that for every $x_S \in \omega M_S$, $x_S + \tau_S \overline{A} x_S \in \delta_S \omega M_S$, where $0 < \delta_S < 1$.[8]

[8] The τ_S, δ_S obtained here may be different from those obtain before, in the following if we mention τ_S, δ_S, then they take the values obtained for ωM_S.

Set now $\tau = \min(\tau_S, \tau_T)$, $\delta = \max(\delta_S, \delta_T)$, then for $x_S \in \omega M_S$ and $x_T \in M_T$, equations

$$x_S + \tau \overline{A} x_S \in \delta \omega M_S, \quad x_T + \tau v_T \in \delta M_T$$

hold simultaneously. Let $b = \max\{\|x_S\|, x_S \in M_S\}$, $\omega_0 = (1-\delta)/4\tau\,\|A\|\,b$ and $v \in A(x) = A(x_S + x_T)$. Then $\|v - v_T\| \le \|A\|\,\|x_S\| \le \|A\|\,\omega b$ for every $\omega \in (0, \omega_0)$. Thus, we have

$$\begin{aligned}
\left[0 \;\, x_T^T\right]^T + \tau P_T v &= \left[0 x_T^T\right]^T + \tau v_T + \tau P_T v - \tau v_T \\
&\in \delta M_T + \tau\,||P_T v - v_T||\,B_T \\
&\subset \delta M_T + \tau\,||v - v_T||\,B_T \\
&\subset \delta M_T + \tfrac{1-\delta}{4} B_T \\
&\subset \delta M_T + \tfrac{1-\delta}{4} M_T \\
&= \tfrac{1+3\delta}{4} M_T \\
&\subset \tfrac{1+\delta}{2} M_T.
\end{aligned}$$

Because $P_S v \in P_S A(x_S \oplus x_T) = \overline{A} P_S (x_S \oplus x_T) = \overline{A} x_S$, $x_S + \tau \overline{A} x_S \in \delta \omega M_S$. It is equivalent to

$$x_S + \tau P_S v \in \delta \omega M_S \subset \frac{1+\delta}{2} \omega M_S.$$

Therefore,

$$\begin{bmatrix} x_S \\ x_T \end{bmatrix} + \tau v \in \frac{1+\delta}{2} \begin{bmatrix} \omega M_S \\ M_T \end{bmatrix},$$

where $v \in A(x)$. Inequality (3.4.10) holds if we replace $(1+\delta)/2$ by δ and $(1+\delta)/2 < 1$. □

Theorem can be treated the Lyapunov converse theorem. Combining Theorems 3.4.2 and 3.4.3, we get that convex system (3.4.4) is weakly asymptotically stable, if and only if there is a $v \in \operatorname{Im}(A)$, two positive definite functions $V(x)$ and $W(x)$ such that

$$D^- V(x)(v) \le -W(x).$$

Problems

1. Consider the convex system (3.4.4), if Inequality (3.4.10) holds and $v \in A(x)$, then $D^- V(x)(v) \le -\theta V(x)$ with $\theta > 0$.

2. Suppose $A(x) = Cx+K$ where $C \in \mathbb{R}^{n\times n}$ and $K \subset \mathbb{R}^n$ is a close and convex cone. Find the condition under which convex system $\dot{x} \in A(x)$ is weakly asymptotically stable.
3. Consider the convex system (3.4.4) where $A(x) = Cx + K$, $C = \begin{bmatrix} 0 & 1 \\ -3 & -2 \end{bmatrix}$ and K is the first quadrant.

   (1) Is the convex system weakly asymptotically stable? If yes, give a stable solution for the convex system.

   (2) Is the convex system asymptotically stable? If no, find the restriction of K, under which the convex system is asymptotically stable.

Reference

Aubin J-P, Cellina A (1984) Differential Inclusions – set-valued maps and viability theory [M]. Springer, Berlin

Chapter 4
Linear Polytope Control Systems

A class of differential inclusion systems – the linear polytope systems – is discussed in this chapter. This kind of system can be viewed as another extension of the linear control systems to the set-valued mappings. This chapter contains four sections. In the first section, we present the definition of the linear polytope system and the motivation of investigation. Section 4.2 deals with the convex hull Lyapunov function which is the main tool in this chapter. Then Sect. 4.3 considers the control of the linear polytope system. We apply the conclusions of Sect. 4.3 to deal with saturated control at the last section of this chapter.

In the theory of differential inclusion control systems, the linear polytope system is relatively simple. Moreover, only the linear polytope control system and the Luré differential inclusion system discussed in next chapter are investigated comparatively deeply by so far.

4.1 Polytope Systems

In order to conveniently narrate later, we introduce some results of linear systems and the matrix inequalities at the beginning of this section. They are needed for the following presentation, especially, the matrix inequality has become an important tool for the control system design in recent years. The polytope system is defined and the motivation for the investigation is introduced in this section.

4.1.1 Linear Control Systems and Matrix Inequalities

1. Linear Control Systems

We start with a brief review of linear control systems.

Z. Han et al., *Theory of Control Systems Described by Differential Inclusions*,
Springer Tracts in Mechanical Engineering, DOI 10.1007/978-3-662-49245-1_4

A typical linear control system is described as follows:

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t),\\ y(t) &= Cx(t),\end{aligned} \tag{4.1.1}$$

where for each given $t \in [0, \infty)$, $x(t) \in \mathbb{R}^n$ is the state of the system, $u(t) \in U \subset \mathbb{R}^m$ is the input or control, U is the set of permissible controls, and $y(t) \in \mathbb{R}^r$ is the output. The first one in Eq. (4.1.1) is called the state equation, it is a differential equation; and the second equation is the output equation, it is an algebraic equation. The dimension n of the state $x(t)$ is called the order of system (4.1.1), A, B, C are real coefficient matrices with appropriate dimensions. When not including time variable t, system (4.1.1) is called time invariant, otherwise, it is called time variant.

Since this book mainly discusses time invariant differential inclusion, only the linear time invariant control system is introduced here. A is an $n \times n$ matrix in System (4.1.1) and is called the dynamic matrix of the system. The eigenvalues of A are known as the poles of the system. The stability of the system is completely determined by the eigenvalues of A. The $(n + r) \times (n + m)$ matrix

$$\begin{bmatrix} A & B \\ C & 0 \end{bmatrix}$$

is known as the system matrix of System (4.1.1). It is also a completely mathematical description for System (4.1.1). In this context, we often use (A, B) to replace the first equation in System (4.1.1) and (C, A, B) to express System (4.1.1).

By the Laplace transformation and eliminating the state variable, System (4.1.1) yields

$$Y(s) = C(sI - A)^{-1}BU(s).$$

$C(sI - A)^{-1}B$ is known as the transfer function matrix of System (4.1.1), the degree of its denominator is larger than that of every element in numerator matrix, so it is a strictly proper real rational fraction matrix.

System (4.1.1) is often described by using block diagram as follows.

In Fig. 4.1, the part in the dashed box is called the interior of the system and is called exterior of the system outside the dashed box. So $x(t)$ is called the interior variable, $u(t)$ and $y(t)$ are called exterior variables. Generally, it is believed that the internal state variable x is not directly measured. Thus, signals can be obtained only the input u and the output y.

In this context, the transfer function matrix is known as the external description of system (4.1.1). $u = F(x)$ is the state feedback of system (4.1.1); accordingly, $u = K(y)$ is the output feedback. When $F(x)$ and $K(y)$ are linear mapping, these closed loop systems are also linear. Otherwise, they are regarded as nonlinear systems. The following mapping

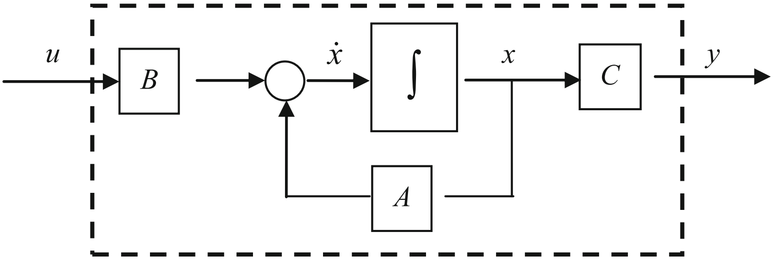


Fig. 4.1 The linear control system

$$\begin{aligned} \dot{\xi}(t) &= M\xi(t) + Ny(t), \\ u(t) &= H\xi(t) + Ky(t), \end{aligned} \tag{4.1.2}$$

is the dynamic linear output feedback of System (4.1.1), where $\xi(t) \in \mathbb{R}^p$ is the state of dynamic compensator and p is the order of the dynamic compensator. State feedback, output feedback and dynamic output feedback are the main control method in the control design. The common control objectives are pole configuration, stabilization, tracking, regulation, optimal control and so on.

Because in Fig. 4.1, the interior of the dashed box is a black box, that is, the state $x(t)$ is not directly available, the observation is an important task in control system design. Σ_{O} is called as an asymptotic observer of System (4.1.1), if the output $\hat{x}$ of the Σ_{O} satisfies that $\lim\limits_{t\to\infty}(\hat{x}(t) - x(t)) = 0$. At present, the most common observer is called Lunberger observer and is constructed as follows:

$$\dot{\hat{x}} = (A - KC)\,\hat{x} + Ky + Bu \tag{4.1.3}$$

when (A, C) is observable, it always can be found a feedback gain K such that $\lim\limits_{t\to\infty}(\hat{x}(t) - x(t)) = 0$.

The observer Eq. (4.1.3) is called full order observer, since its dimension is equal to that of (4.1.1). In fact, the output y directly includes some information about x, we can design an observer with order $n - r$, this observer is called reduced observer.

2. Matrix Inequalities

The matrix inequality is an effective tool for the control system design at present. Since the interior point matrix has been put forward, solving the linear matrix inequality becomes very convenient. Moreover, a lot of effective computer aided design software have been developed and widely used; thus the application of the matrix inequality has been greatly extended, and many design conditions are given by these inequalities.

Assume that $P \in \mathbb{R}^{n \times n}$ is a symmetric matrix. $P > 0$ implies that P is a positive definite matrix; $P \geq 0$ means that P is a positive semi-definite matrix. Similarly, it can be defined that $P < 0$ and $P \leq 0$. The application of inequality signs can be extended to two matrices. If $A, B \in \mathbb{R}^{n \times n}$ are two symmetric matrices, then $A > B$ means $A - B > 0$. Similarly, it can be defined that $A \geq B$, $A < B$ and $A \leq B$.

The following lemma, called by Schur complement lemma, plays an important role in the later discussion.

Lemma 4.1.1 Given a symmetric matrix

$$A = \begin{bmatrix} P & R \\ R^T & Q \end{bmatrix} \in \mathbb{R}^{n \times n},$$

with $P \in \mathbb{R}^{m \times m}$, $Q \in \mathbb{R}^{r \times r}$, $m + r = n$, then the following three conditions are equivalent:

1. $A < 0$;
2. $P < 0$, $Q - R^T P^{-1} R < 0$;
3. $Q < 0$, $P - R Q^{-1} R^T < 0$. □

The proof of Lemma 4.1.1 can be found in a general textbook for matrix theory, here omitted. For Lemma 4.1.1, there exists the corresponding conclusion for the positive definite matrix; readers are suggested to state and the proof. In order to facilitate the description, we call Conclusion 2 to be the application of Schur lemma about the matrix P and Conclusion 3 about the matrix Q. There exists a version of Schur lemma for negative semi-definite matrices which is given in Lemma 4.2.1.

Lemma 4.1.2 Given a symmetric matrix

$$A = \begin{bmatrix} P & R \\ R^T & Q \end{bmatrix} \in \mathbb{R}^{n \times n},$$

with $P \in \mathbb{R}^{m \times m}$, $Q \in \mathbb{R}^{r \times r}$, $m + r = n$, then the following conditions are equivalent:

(1) $A \leq 0$;
(2) $P \leq 0$, $Q - R^T P^{-1} R \leq 0$, $R^T \left(I - P P^{-1}\right) = 0$;
(3) $Q \leq 0$, $P - R Q^{-1} R^T \leq 0$, $R \left(I - Q Q^{-1}\right) = 0$;

where P^{-1} and Q^{-1} are Moore-Penrose inverse[1] of P and Q, respectively. The third equality holds in Conclusion 2 or 3 if P or Q is the invertible matrix.

[1] Recall the definition of Moore-Penrose inverse, assume that A is an $n \times n$ symmetric matrix, $n \times n$ matrix G *is* a Moore-Penrose inverse of A, such that $AGA = G$, $GAG = A$, $(AG)^T = AG$, $(GA)^T = GA$ hold.

First, the meaning of the two lemmas is that it transfers the problem of verifying negative definite of a matrix with order n into two low dimensional matrices being negative definite, thereby it reduces the complexity of checking. This application is referred to as the positive application. Secondly, it transforms a quadratic matrix inequality into a linear matrix inequality with high dimension; this application is called reverse application of Schur lemma. Opportunities of the reverse application will be more than the positive application in the control system design.

For example, we want to solve the matrix inequality $X^T X < I$ where $X \in \mathbb{R}^{p \times q}$; this is a quadratic matrix inequality by Schur lemma, and the inequality is equivalent to $\begin{bmatrix} -I & X \\ X^T & -I \end{bmatrix} < 0$. The latter can be solved since it is a linear one.

For instant, the sufficient and necessary condition of matrix A being a Hurwitx matrix is for any positive definite matrix Q, there exists a positive definite solution P for Lyapunov equation $PA + A^T P = -Q$. This is equivalent to solve a linear matrix inequality

$$\begin{bmatrix} -P & 0 \\ 0 & PA + A^T P \end{bmatrix} < 0.$$

One more example is about Riccati inequality. $KA + A^T K + KBR^{-1} B^T K + Q < 0$ is quadratic for the unknown matrix K, it can be transformed into

$$\begin{bmatrix} A^T K + KA + Q & KB \\ B^T K & -R \end{bmatrix} < 0.$$

This is a linear matrix inequality of K, which can be conveniently obtained using the software.

The following two inequalities are basic, and they can be proved based on Schwarz inequality, and the detailed proofs are omitted.

Assume that U and V are two matrices with appropriate dimensions, then for any positive real number α, we have

$$U^T V + V^T U \leq \alpha U^T U + \alpha^{-1} V^T V. \tag{4.1.4}$$

Let x and y be two n-dimensional real vectors, Q is an $n \times n$ positive definite matrix, then it holds

$$2x^T y \leq x^T Q x + y^T Q^{-1} y. \tag{4.1.5}$$

Afterward it will mention the Hurwitz matrix, recall its definition: a matrix is called Hurwitz matrix, if its eigenvalues have negative real parts. The necessary and sufficient condition for the linear system $\dot{x} = Ax$ asymptotically stable is that A is a Hurwitz matrix.

4.1.2 Linear Polytope Systems

1. The Description of Linear Polytope Systems

If $A_i \in \mathbb{R}^{n\times n}$, $i = 1, 2, \ldots, N$, are N real matrices, then the set

$$\mathrm{co}\,(A_i,\ i = 1, 2, \ldots, N) = \left\{ A = \sum_{i=1}^{N} \gamma_i A_i;\ \left[\gamma_1\ \gamma_2\ \ldots\ \gamma_N \right]^T \in \Gamma \right\}$$

is a convex hull composed by A_i, where

$$\Gamma = \left\{ \boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_N)\,; 0 \le \gamma_i \le 1,\ i = 1, 2, \ldots, N, \sum_{i=1}^{N} \gamma_i = 1 \right\}.$$

Γ is a closed convex set of the first octant on $\mathbb{R}^N$, for example, it is a closed line segment connecting $(1, 0)$ and $(0, 1)$ on $\mathbb{R}^2$; it is a closed triangle with vertices $(1, 0, 0), (0, 1, 0),\ (0, 0, 1)$ in $\mathbb{R}^3$.

Because $\mathrm{co}\,(A_i,\ i = 1, 2, \ldots, N)$ is a convex hull obtained by composing finite elements, it is naturally a closed set. Thus, $\overline{\mathrm{co}}\,(A_i,\ i = 1, 2, \ldots, N) = \mathrm{co}\,(A_i,\ i = 1, 2, \ldots, N)$. For simplicity, the convex hull is still denoted as $\mathrm{co}(A_i)$.

When N is a finite integer, $\mathrm{co}(A_i)$ is called a polytope composed of $A_i,\ i = 1, 2, \ldots, N$. By the definition, we can see that the polytope is a convex set generated by finite linear elements (matrices, vectors).

Assume that $[0, T)$ is a time interval (where T can be ∞), for an any given $t \in [0, T)$, $x(t) \in \mathbb{R}^n$, we define the set $\mathrm{co}\{A_i x(t)\}=\{Ax(t),\ A \in \mathrm{co}\,\{A_i, i=1, 2, \ldots, N\}\}$, the differential inclusion

$$\dot{x}(t) \in \mathrm{co}\,(A_i\, x(t))\,, \tag{4.1.6}$$

is called polytope differential inclusion. The Cauchy problem of polytope differential inclusion Eq. (4.1.6) is

$$\dot{x}(t) \in \mathrm{co}\,(A_i\, x(t))\,, x(0) = x_0.$$

In the control theory, the system considered should hold control input. Therefore, let us consider N linear control systems

$$\begin{bmatrix} A_i & B_i \\ C_i & 0 \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+r)},\ \ i = 1, 2, \cdots N,$$

then

$$\mathrm{co}\left(\begin{bmatrix} A_i & B_i \\ C_i & 0 \end{bmatrix}\right) = \left\{ \sum_{i=1}^{N} \gamma_i \begin{bmatrix} A_i & B_i \\ C_i & 0 \end{bmatrix},\ \gamma \in \Gamma \right\}$$

is a convex hull of these N linear control systems. According to the above discussion, it is a closed set, the polytope system about $\mathrm{co}\left(\begin{bmatrix} A_i & B_i \\ C_i & 0 \end{bmatrix}\right)$ is denoted by

$$\begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix} \in \mathrm{co}\left(\begin{bmatrix} A_i & B_i \\ C_i & 0 \end{bmatrix}\begin{bmatrix} x(t) \\ u(t) \end{bmatrix}\right). \tag{4.1.7}$$

Inclusion (4.1.7) is called a linear polytope system. Note that the second relation in Inc. (4.1.7) is an algebraic relation.

2. The Background of Polytope Systems

Linear polytope system control is the earliest formally presented by Hu et al. (2006), when they studied the saturated control for linear systems (to see the fourth section of this chapter). In fact, many problems can be transformed into the linear polytope problem. For example, the discontinuous Eq. (2.3.5) in Sect. 2.3 of this book can be transformed into the differential inclusion, it is really a polytope system.

This section will continue to give some examples to illustrate applications of the polytope system in practical problems.

Example 4.1.1 Consider the following differential equation

$$\dot{x} = \sqrt[3]{x}\sin x^2.$$

This differential equation cannot be solved by the elementary function. But we know $\sin x^2 \in [-1, 1]$, then the differential equation can be extended into a differential inclusion

$$\dot{x} \in \mathrm{co}\left(-\sqrt[3]{x}, \sqrt[3]{x}\right). \tag{4.1.8}$$

Inclusion (4.1.8) is a polytope system. Solving the differential equation $\dot{x} = \beta\sqrt[3]{x}$, with $\beta \in [-1, 1]$, yields

$$x(t) = \left[\frac{2}{3}\beta t + x^{\frac{2}{3}}(0)\right]^{\frac{3}{2}}. \tag{4.1.9}$$

If the time domain is divided into some small intervals $[t_k,\ \ t_{k+1})$, in each interval, the solution $x(t) = \left[\frac{2}{3}\beta t + x^{\frac{2}{3}}\left(t_k\right)\right]^{\frac{3}{2}}$ given by Eq. (4.1.9) is approximate to the origin equation, the parameter β can be selected that $\beta = \sin\left(x\left(t_k\right)\right)^2$.

Example 4.1.2 Figure 4.2 is a typhoon forecast issued by China National Meteorological Center in 2011. The fold line in the lower right of figure depicts the historical track of typhoon center motion. At the end of this path is typhoon center location at 8 a.m. in August 5, 2011, its upper left coated with the conical area shadow is 48 h ahead that typhoon center may reach area. Although it shows “path probability forecast map” in the next 48 h, a curve is scaled out in the shaded part, that is, the center of typhoon track which is forecasted, two black spots on it are the estimations of the typhoon center after 24 and 48 h in the future, respectively. But it cannot provide any probability of typhoon center transfer, also did not have any predictive model of typhoon motion. These trajectories and the center position which were estimated look like a center line.

If we use $x(t)$ and $y(t)$ to express the longitude and latitude of the typhoon center at t moment, then the vector $[x(t) \;\; y(t)]^T$ is the coordinate of the typhoon center, and $\left[\dot{x}(t) \;\; \dot{y}(t)\right]^T$ is the velocity vector of the typhoon center which gives the moving direction as well as speed.

Let $[a_1(t) \;\; b_1(t)]^T$ and $[a_2(t) \;\; b_2(t)]^T$ indicate the upper and lower curves of shadow areas in Fig. 4.2. We depict them again in Fig. 4.3, then

$$\dot{a}_2(t) \le \dot{x}(t) \le \dot{a}_1(t) \quad \text{and} \quad \dot{b}_2(t) \le \dot{y}(t) \le \dot{b}_1(t),$$

or

$$\begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} \in \text{co}\left(\begin{bmatrix} a_2(t) \\ b_2(t) \end{bmatrix}, \begin{bmatrix} a_1(t) \\ b_1(t) \end{bmatrix} \right), \quad \begin{bmatrix} x(t_0) \\ y(t_0) \end{bmatrix} = \begin{bmatrix} x(0) \\ y(0) \end{bmatrix}.$$

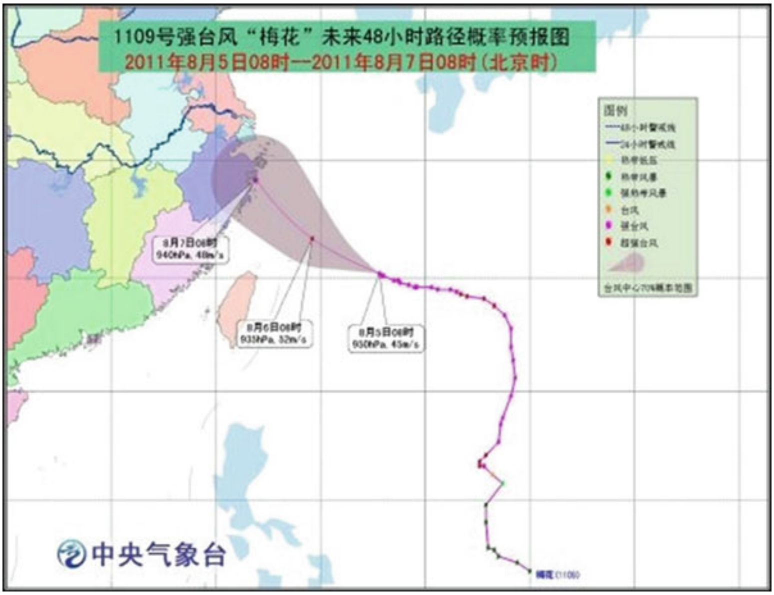


Fig. 4.2 A typhoon forecast in 2011

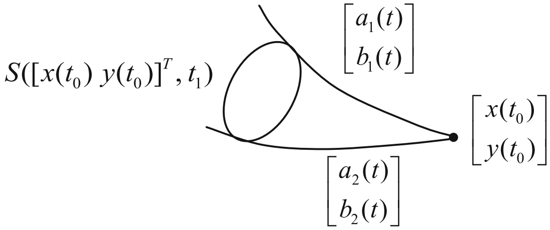

$S([x(t_0)\ y(t_0)]^T, t_1)$
$\begin{bmatrix} a_1(t) \\ b_1(t) \end{bmatrix}$
$\begin{bmatrix} x(t_0) \\ y(t_0) \end{bmatrix}$
$\begin{bmatrix} a_2(t) \\ b_2(t) \end{bmatrix}$

Fig. 4.3 Description of typhoon center movement by differential inclusion

This is a Cauchy problem of polytope differential inclusion. Let $S([x(t_0)\ y(t_0)]^T, t)$ be the solution set of this differential inclusion. Then $S([x(t_0)\ y(t_0)]^T, t_1)$ is the set of all of possible situations of typhoon center at t_1 moment. Moreover, if we have experience distribution $F(x, y)$ for this set, then $[Ex\ \ Ey]^T$ gives the optimal position of typhoon center at time t_1. If this distribution is uniform, then the desired position is the center of gravity of $S([x(t_0)\ y(t_0)]^T, t_1)$; if the distribution is not uniform, so the most likely trajectory of moving is not a center line as given in Fig. 4.2. □

Example 4.1.3 Many robust control problems can be described by polytope systems.

Consider linear system

$$\dot{x} = (A + \Delta A)\, x + Bu, \tag{4.1.10}$$

where $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ are the state and the input of the system, respectively; $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are the two known matrices, $\Delta A \in \mathbb{R}^{n\times n}$ stands for uncertainty, it is unknown, may even be a matrix variable. In robust control, we often assume that it is a bounded matrix. Let $\Delta A = \left(\alpha_{ij}\right)_{n\times n}, \underline{\alpha} \leq \alpha_{ij} \leq \overline{\alpha}$, then Eq.(4.1.10) is equivalent to a differential inclusion system described as follows

$$\dot{x} \in \operatorname{co}\left((A + \underline{\alpha} I)\, x + Bu, (A + \overline{\alpha} I)\, x + Bu\right). \tag{4.1.11}$$

For Inc. (4.1.10), under the feedback $u = Fx$, the corresponding closed-loop system is a differential inclusion

$$\dot{x} \in \operatorname{co}\left((A + BF + \underline{\alpha} I)\, x, (A + BF + \overline{\alpha} I)\, x\right). \tag{4.1.12}$$

The robust stability of the closed-loop system is equivalent to strong stability of the differential inclusion system Inc. (4.1.12). □

Problems

1. Assume that $F_0, F_1, \ldots, F_n$ are symmetric matrices with the same number of rows and columns, $x_1, \ldots, x_n$ are n unknown real number, solving the linear matrix inequality is to compute the vector $x = [x_1 \ldots x_n]^T$, such that

$$F_0 + x_1 F_1 + \cdots + x_n F_n < 0.$$

   Give a sufficient condition for the solution set being non-empty, and prove that the solution set is a convex set.
2. Prove Lemma 4.1.2.
3. (1) Extend Inequality (4.1.4) to show that $U^T V + V^T U \le U^T Q U + V^T Q^{-1} V$ where Q is a positive definite matrix.
   (2) Prove that Inequality (4.1.5) becomes equation if and only if $y = Qx$.
4. Prove that Γ is a closed hull set in the first octant on $\mathbb{R}^N$.
5. There are two polytope systems

$$\begin{bmatrix} A_i & B_i \\ C_i & 0 \end{bmatrix} \in R^{(n+m)\times(n+m)},\ i{=}1,2,\cdots N \begin{bmatrix} A_j & B_j \\ C_j & 0 \end{bmatrix} \in R^{(n+m)\times(n+m)},\ i{=}1,2,\ldots M.$$

   Write the expressions of series system and parallel system, respectively.

4.2 Convex Hull Lyapunov Functions

This section discusses a kind of special Lyapunov function, called the convex hull Lyapunov function. In fact, *perhaps* it may be *better to* call it "convex hull quadratic Lyapunov function", since it takes a form of quadratic function. Because it is defined by a convex combination of some quadratic functions, it has more advantages than a single quadratic function. The convex hull quadratic Lyapunov function is proposed in the early twenty-first century. It is an important tool for the analysis and design of polytope systems and is used to get the invariable set and controlled invariable set, and also becomes a useful tool in the saturated control for linear systems (Hu and Lin 2004; Hu et al. 2006). In this section, we discuss the definition and fundamental properties of convex hull Lyapunov functions, and it will provide a foundation for further design.

4.2.1 Convex Hull Quadratic Lyapunov Functions

Let $P_i \in \mathbb{R}^{n\times n}$, $i = 1, 2, \ldots, N$, be N positive definite matrices, and $Q_i = P_i^{-1}$. Then Q_i is also positive definite for every $i \in \{1, 2, \ldots, N\}$. Define

$$\Gamma = \left\{ \gamma = [\gamma_1,\ \gamma_2, \ldots, \gamma_N]^T \in \mathbb{R}^N;\ \ \gamma_i \geq 0, i = 1, 2, \ldots, N, \sum_{i=1}^{N} \gamma_i = 1 \right\},$$

$\mathit{\Gamma}$ is a simplex on $\mathbb{R}^N$ with its vertices $e_i = \begin{bmatrix} 0 & \cdots & \overbrace{1}^{i} & \cdots & 0 \end{bmatrix}^T$. Γ is a bounded, closed and convex set. For every $\gamma \in \Gamma$, let $Q(\gamma)$ be the convex combination of Q_i, i.e.,

$$Q\left(\gamma\right) = \sum_{i=1}^{N} \gamma_i Q_i, \tag{4.2.1}$$

$Q(\gamma)$ is always positive definite, and a linear mapping on $\mathbb{R}^n$, it is also a linear mapping on Γ. Define $P\left(\gamma\right) = Q^{-1}\left(\gamma\right)$, $P(\gamma)$ is also positive definite. Further, because $Q\left(\gamma\right) \neq 0$ (zero matrix), $P(\gamma)$ is also a continuous mapping on the simplex Γ. Especially, when $N = 1, P\left(\gamma\right) = P$, and when $N > 1$, for $i = 1, 2, \ldots, N$, $P_i \in \left\{P\left(\gamma\right),\ \gamma \in \Gamma\right\}$, it is the reason why $P(\gamma)$ can be expected to have better properties than a single P_i.

Definition 4.2.1 The following positive definite function

$$V_c(x) = \min_{\gamma \in \Gamma} x^T P\left(\gamma\right) x = \min_{\gamma \in \Gamma} x^T \left(\sum_{i=1}^{N} \gamma_i Q_i \right)^{-1} x$$

is the convex hull quadratic Lyapunov function of $P_i,\ i = 1, 2, \ldots, N$. □

Although $V_c(x)$ is called the convex hull quadratic Lyapunov function of P_i, $P(\gamma)$ is not a convex combination of P_i, where the "convex combination" comes from that $Q(\gamma)$ is a convex combination of Q_i.

Since Γ is a compact set, for each x, there exists a $\gamma^* = \gamma^*(x) \in \Gamma$, such that

$$V_c(x) = \min_{\gamma \in \Gamma} x^T P\left(\gamma\right) x = x^T P\left(\gamma^*\right) x.$$

Note that the notation $\gamma^* = \gamma^*(x)$ only means that γ^* depends on x, but $\gamma^*(x)$ may not be a function of x, it is usually a set-valued mapping. At the end of this section, we will give the condition under which $\gamma^*(x)$ is a function of x.

It can be verified directly that $\gamma^*(x) = \gamma^*(ax)$ for every $a \in \mathbb{R}$. We now give an alternative definition for $V_c(x)$.

Theorem 4.2.1 The value of $V_c(x)$ can be obtained by the following optimization problem:

$$V_c(x) = \min \left\{\alpha; \text{there exists a } \ \gamma \in \Gamma, \text{such that } \ \alpha \geq x^T P\left(\gamma\right) x\right\}$$

Proof Denote

$$A(x) = \left\{\alpha; \text{there exists a } \ \gamma \in \Gamma, \text{such that } \ \alpha \geq x^T P\left(\gamma\right) x\right\}$$

and $\alpha_0(x) = \inf A(x)$. Because there exists a $\gamma^* \in \Gamma$ such that $V_c(x) = x^T P\left(\gamma^*\right) x$, $V_c(x) \in A(x)$. By the definition of $\alpha_0(x)$, we have $V_c(x) \geq \alpha_0(x)$.

On the other hand, for any $\varepsilon > 0$, there exists $\alpha \in A(x)$, such that $\alpha_0(x) + \varepsilon > \alpha$, i.e., there exists $\gamma \in \Gamma$, such that

$$\alpha_0(x) + \varepsilon > \alpha \geq x^T P\left(\gamma\right) x \geq V_c(x).$$

Since ε can be selected arbitrarily, we obtain $\alpha_0(x) = V_c(x)$. □

Theorem 4.2.1 shows that the set $A\left(x\right)$ exists the minimum value for every x, and the minimum value is exactly equal to $V_c(x)$.

Owing to $P^{-1}\left(\gamma\right) = Q\left(\gamma\right) > 0$, so $\alpha \geq x^T P\left(\gamma\right) x$ is equivalent to

$$\begin{bmatrix} \alpha & x^T \\ x & Q\left(\gamma\right) \end{bmatrix} \geq 0.$$

From Theorem 4.2.1, the another description of $V_c(x)$ is

$$\begin{aligned} V_c(x) &= \min \ \ \alpha, \\ \text{s.t.} & \begin{bmatrix} \alpha & x^T \\ x & Q\left(\gamma\right) \end{bmatrix} \geq 0, \ \ \exists \gamma \in \Gamma \ . \end{aligned} \tag{4.2.2}$$

The optimization problem with constraint (4.2.2) is convenient to be solved by the existing software.

We have verified in Sect. 1.3 that a positive quadratic function $x^T P x$ is a convex function; however, $V_c(x) = x^T P\left(\gamma^*\right) x$ is not a convex function. The fact is illustrated as follows.

Consider

$$\begin{bmatrix} \alpha & \lambda x_1^T + (1-\lambda)\, x_2^T \\ \lambda x_1 + (1-\lambda)\, x_2 & Q\left(\gamma\right) \end{bmatrix} = \lambda \begin{bmatrix} \alpha & x_1^T \\ x_1 & Q\left(\gamma\right) \end{bmatrix} + (1-\lambda) \begin{bmatrix} \alpha & x_2^T \\ x_2 & Q\left(\gamma\right) \end{bmatrix}. \tag{4.2.3}$$

We still adopt the notations used in the proof of Theorem 4.2.1. If $\alpha_0\left(x_1\right) \neq \alpha_0\left(x_2\right)$, without the loss of generality, we assume $\alpha_0\left(x_1\right) > \alpha_0\left(x_2\right)$, then for any $\gamma \in \Gamma$, it holds $\begin{bmatrix} \alpha_0\left(x_2\right) & x_1^T \\ x_1 & Q\left(\gamma\right) \end{bmatrix} < 0$. When $\lambda \to 1^-$, by Eq. (4.2.3), for any $\gamma \in \Gamma$, $\begin{bmatrix} \alpha_0\left(x_2\right) & \lambda x_1^T + (1-\lambda)\, x_2^T \\ \lambda x_1 + (1-\lambda)\, x_2 & Q\left(\gamma\right) \end{bmatrix} < 0$. By the definition of $\alpha_0(x)$, when $\lambda \in [1-\varepsilon, 1]$, it yields $\alpha_0\left(\lambda x_1 + (1-\lambda)\, x_2\right) > \alpha_0\left(x_2\right)$. It follows that the inequality

$$\alpha_0 (\lambda x_1 + (1-\lambda) x_2) \leq \lambda \alpha_0 (x_1) + (1-\lambda) \alpha_0 (x_2)$$

does not hold for all λ in $0 \leq \lambda \leq 1$.

Since $V_c(x)$ is not a convex function, it causes many difficulty for study. In the research of convex hull quadratic functions, we will consider the layer set and prove that the layer set is always convex.

Recall the layer set defined in Sect. 1.2. $\mathrm{lev}\,(f \leq \alpha)$ is defined as the set $\{x;\ x \in \mathrm{dom}\, f, f(x) \leq \alpha\}$. If it does not illustrate especially, this chapter only considers the layer set given by "$\leq$".

In order to simplify the notation, let $L_c(\alpha)$ denotes the set $\mathrm{lev}\,(V_c(x) \leq \alpha)$ and $L_{P_i}(\alpha)$ denotes the $\mathrm{lev}\left(x^T P_i x \leq \alpha\right)$. Similar notations will be employed for other functions, and we do not point out later. Especially, when $\alpha = 1$, we denote $L_c = L_c(1)$ for simplicity. For a quadratic function $x^T \Lambda x$, it is easy to see $L_\Lambda(\alpha) = \sqrt{\alpha} L_\Lambda$ for any $\alpha \geq 0$, so it is sufficient to discuss L_Λ only for a quadratic function $x^T \Lambda x$. Since $\gamma^*(x) = \gamma^*(ax)$ for $V_c(x)$, it can be proved $L_c(\alpha) = \sqrt{\alpha} L_c$.

If $\gamma^* = \gamma^*(x)$ is a continuous function of x, all lever sets mentioned above are compact.

4.2.2 Layer Sets for the Convex Hull Quadratic Function

This section gives some basic properties of the layer set. We first prove a useful lemma.

Lemma 4.2.1 For a given vector $f \in \mathbb{R}^n$ and a matrix $P \in \mathbb{R}^{n\times n} > 0$, $L_P \subset L_f$ if and only if $f^T P^{-1} f \leq 1$, where $L_f = \{x; |\langle f, x\rangle| \leq 1\}$.

Proof Sufficiency. If $f^T P^{-1} f \leq 1$, that is, $1 - f^T P^{-1} f \geq 0$, from Lemma 4.1.1, it holds

$$\Phi = \begin{bmatrix} 1 & f^T P^{-1} \\ P^{-1} f & P^{-1} \end{bmatrix} \geq 0\ . \tag{4.2.4}$$

Using Lemma 4.1.1, we have $P^{-1} - P^{-1} f f^T P^{-1} \geq 0$, and multiplying from left and from right with P, it yields $P \geq ff^T$. If $x \in L_P$, then $1 \geq x^T P x$, we have $1 \geq x^T P x \geq x^T f f^T x$, that is, $x \in L_f$.

Necessity. Suppose that $P \geq ff^T$ does not hold, then there exists an $x_0 \in \mathbb{R}^n$, such that $x_0^T P x_0 < x_0^T f f^T x_0$. Denote $l = x_0^T P x_0$, then $l > 0$ and $\left(\sqrt{l}\right)^{-1} x_0 \in L_P$, but $\left(\left(\sqrt{l}\right)^{-1} x_0\right)^T ff^T \left(\left(\sqrt{l}\right)^{-1} x_0\right) = l^{-1} x_0^T f f^T x_0 > l^{-1} x_0^T P x_0 = 1$, $\left(\sqrt{l}\right)^{-1} x_0 \notin L_f$. This is a contradiction to $L_P \subset L_f$, so $P \geq ff^T$, that is, $P^{-1} - P^{-1} f f^T P^{-1} \geq 0$, it implies (4.2.4), by Lemma 4.1.1, we get $f^T P^{-1} f \leq 1$. □

There are three remarks for Lemma 4.2.1.

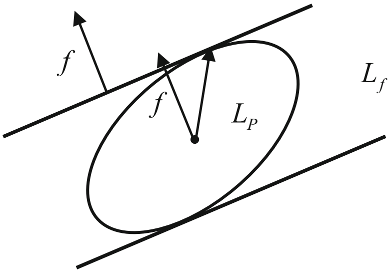


Fig. 4.4 Illustration for $L_P \subset L_f$

Remark 1 From the view of geometry, $x^T Px = 1$ is an ellipsoid on $\mathbb{R}^n$, $f^T x = 1$ is a $n-1$ dimensional hyperplane in the $\mathbb{R}^n$, and f is the normal vector of the hyperplane. $\left|f^T x\right| \leq 1$ is the area between the two hyper planes $f^T x = \pm 1$.

$$x_0 = P^{-1} f$$

Thus, $L_P \subset L_f$ shows that the ellipsoid L_P is between the two hyperplanes as shown in Fig. 4.4. $1 = x^T Px = x^T ff^T x$ implies that the ellipsoid is tangent to the hyperplane $f^T x = 1$. The point of tangency is $x_0 = P^{-1} f$. For every $x \in L_P \backslash \{x_0, -x_0\}$, it holds $\left|f^T x\right| < f^T x_0 = 1$.

If $L_P \subset L_f$, and $\mathrm{db} L_P \cap \left\{x; \left|f^T x\right| = 1\right\} = \varnothing$, then the ellipsoid $x^T Px = 1$ is between the two hyperplanes $f^T x = \pm 1$ without touching them, consequently, $f^T P^{-1} f < 1$.

If f_1 is a linearly dependent vector with f, i.e. $f_1 = af$, then L_{f_1} and L_f are two parallel zonal areas, and L_{f_1} will be contained by L_f if $|a| \geq 1$ □

Remark 2 If f is replaced by $H \in \mathbb{R}^{m \times n}$ where

$$H = \begin{bmatrix} h_1^T \\ \vdots \\ h_m^T \end{bmatrix},$$

then it can be proved by a similar method that $L_P \subset L_H$ if and only if $h_j^T P^{-1} h_j \leq 1$, for all $j = 1, 2, \ldots, m$, where $L_H = \underset{i}{\cap} L_{h_i^T}$, i.e., $L_H = \{x;\ \|Hx\|_\infty \leq 1\}$, $\|Hx\|_\infty$ is the infinity norm of Hx. If $\left\{f_j,\ j = 1, 2, \ldots\right\}$ is the set of intersection points of hyperplanes $h_i^T x = \pm 1$, then L_H is the simplex with vertices $\left\{f_j,\ j = 1, 2, \ldots\right\}$. Then $L_P \subset L_H$ implies that the ellipsoid L_P is contained in the simplex. The reason why L_P is an ellipsoid and L_H is a simplex is that their norms calculated by the two different ways. □

Remark 3 If $f^T P^{-1} f = 1$, then Φ defined in Eq. (4.2.4) is only a semi-definite matrix. Therefore, there must be x such that $1 = x^T Px = x^T ff^T x$, particularly, it can be required $f^T x = x^T f = 1$. The following will prove that such x is unique.

Since the rank of matrix Φ defined by Eq. (4.2.4) is larger than $n-1$, and there exists a nonzero vector $\left[1 \ \ (-f)^T\right]^T$ such that

$$\Phi\begin{bmatrix} 1 \\ -f \end{bmatrix} = \begin{bmatrix} 1 & f^T P^{-1} \\ P^{-1}f & P^{-1} \end{bmatrix}\begin{bmatrix} 1 \\ -f \end{bmatrix} = 0,$$

we obtain that rank $\Phi = n - 1$. It follows that there is unique x, such that $\left[1 \ \ -x^T P\right]\Phi\left[1 \ \ -x^T P\right]^T = 0$. The x can be solved $x = P^{-1}f$ and satisfies that $1 = x^T P x = x^T f f^T x$. □

The following four theorems present basic properties of the layer set L_c of the quadratic function $V_c(x)$.

For the sake of simplicity, we denote $L_P = \text{co}\left\{L_{P_i}, i = 1, 2, \ldots, N\right\}$.

Theorem 4.2.2 $L_c = L_P = \bigcup\limits_{\gamma\in\Gamma} L_{P(\gamma)}$.

Proof (1) At first, we prove $L_P \subset L_c$. For every $x \in L_P$, there exists $x_i \in L_{P_i}$, such that $x = \sum\limits_{i=1}^{N} \gamma_i x_i$. $x_i \in L_{P_i}$ implies $1 \geq x_i^T Q_i^{-1} x_i$, then we have

$$\begin{bmatrix} 1 & x_i^T \\ x_i & Q_i \end{bmatrix} \geq 0, \begin{bmatrix} 1 & x^T \\ x & Q(\gamma) \end{bmatrix} = \sum_{i=1}^{N} \gamma_i \begin{bmatrix} 1 & x_i^T \\ x_i & Q_i \end{bmatrix} \geq 0,$$

i.e., $1 \geq x^T P(\gamma) x \geq V_c(x), x \in L_c$.

(2) Secondly, we prove $\bigcup\limits_{\gamma\in\Gamma} L_{P(\gamma)} \subset L_P$. It is sufficient, for any $\gamma \in \Gamma$, to prove $L_{P(\gamma)} \subset L_P$.

The conclusion is proved by contradiction. If there exists a $\gamma^* \in \Gamma$ such that $x_0 \in L_{P(\gamma^*)}, \ x_0 \notin L_P$. Since L_P is a convex set, from Lemma 1.2.2 and Remark 3 behind the lemma, there exists a vector n, such that $\langle n, \ x_0\rangle > \langle n, \ x\rangle$, for all $x \in L_P$.

Consider an optimization problem

$$y = \max_{x\in L_{P(\gamma^*)}} \langle n, \ x\rangle .$$

Since $L_{P(\gamma^*)}$ is a closed ellipsoid, there exists unique maximum value. Denote $y_{\max} = \langle n, \ x^*\rangle$, where $x^* \in L_{P(\gamma^*)}$. Let $n^* = \frac{n}{\langle n, \ x^*\rangle}$. Then $1 \geq \langle n^*, \ x\rangle$ for all $x \in L_{P(\gamma^*)}$. Further, by the definition of L_{n^*}, we have $L_{P(\gamma^*)} \subset L_{n^*}$. The hyperplane $(n^*)^T x = 1$ is tangent to $L_{P(\gamma^*)}$. From Remark 2 given after Lemma 4.2.1, we have

$$\left(n^*\right)^T Q\left(\gamma^*\right) n^* = 1. \tag{4.2.5}$$

But, for all $x \in L_P$, it holds that $\langle n, \ x_0\rangle > \langle n, \ x\rangle$. Replacing n by n^*, it yields

$$\left(n^*\right)^T x < \left(n^*\right)^T x_0 \leq \left(n^*\right)^T x^* = 1.$$

Using Remark 2 given after Lemma 4.2.1, for any P_i, $i = 1, 2, \ldots, N$, L_{P_i} is between hyper planes $(n^*)^T x = \pm 1$, without touching them. Therefore, for all $i = 1, 2, \ldots, N$, we have

$$\begin{bmatrix} 1 & (n^*)^T Q_i \\ Q_i n^* & Q_i \end{bmatrix} > 0. \tag{4.2.6}$$

By Inequality (4.2.6), if $\gamma^* = \left[\gamma_1^* \ \gamma_2^* \ \ldots \gamma_N^*\right]^T$, then

$$\begin{bmatrix} 1 & (n^*)^T Q(\gamma^*) \\ Q(\gamma^*) n^* & Q(\gamma^*) \end{bmatrix} = \sum_{i=1}^{N} \gamma_i^* \begin{bmatrix} 1 & (n^*)^T Q_i \\ Q_i n^* & Q_i \end{bmatrix} > 0.$$

By Remark 2 for Lemma 4.2.1, it holds $(n^*)^T Q(\gamma^*) n^* < 1$. This contradicts to Eq. (4.2.5).

(3) Finally, we prove $L_c \subset \underset{\gamma \in \Gamma}{\cup} L_{P(\gamma)}$. Since $x \in L_c$, by the illustration given after Definition 4.2.1, there exists a $\gamma^* \in \Gamma$, such that $1 = V_c(x) = \min_{\gamma \in \Gamma} x^T P(\gamma) x = x^T P(\gamma^*) x$, for $x \in L_{P(\gamma^*)} \subset \underset{\gamma \in \Gamma}{\cup} L_{P(\gamma)}$.

In conclusion, we have proved $L_c = L_P = \underset{\gamma \in \Gamma}{\cup} L_{P(\gamma)}$. □

Theorem 4.2.2 presents the geometric characteristic of the layer L_c of the convex hull quadratic function $V_c(x)$. It illustrates that L_c is a closed convex hull of those L_{P_i}. The fact somewhat explains the reason why we call $V_c(x) = \min_{\gamma \in \Gamma} x^T P(\gamma) x$ the quadratically convex hull Lyapunov function.

Using $L_c(\rho) = \sqrt{\rho} L_c$ and the other equalities, it is direct to obtain

$$L_c(\rho) = \operatorname{co}\{L_{P_i}(\rho), i = 1, 2, \ldots, N\} = \underset{\gamma \in \Gamma}{\cup} L_{P(\gamma)}(\rho).$$

In order to give a geometrical illustration of Theorem 4.2.2, we need to introduce a concept of convex analysis. The concept is also useful in the next theorem which deals with algebraic property of the boundary L_c.

Let A be a closed convex set. $x \in A$ is called an extreme point of A if x cannot be represented as a convex combination of any other points in A. It is obvious that the set of extreme points is contained by the boundary of A (Theorem 1.2.2), i.e., bd$A$. But, usually, it is a proper subset of bd$A$. For instance, in a simplex, its vertices are extreme points, but other points in edges are not extreme points. The closed convex set A is said to be strictly convex, if every point in bd$A$ is an extreme point. For example, every point of the circumference of a closed circle in $\mathbb{R}^2$ is extreme point. The level set L_P is a strictly convex set provided that $P \in \mathbb{R}^{n \times n}$ is a positive definite matrix.

The following lemma gives a characteristic of extreme points.

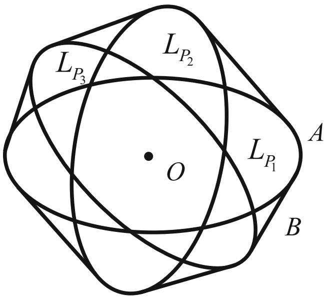


Fig. 4.5 The construction of L_c

Lemma 4.2.2 Let A be a closed convex set, and $x_0 \in A$ be an extreme point of A. Let $f(x)$ be a convex function defined on A. Then $f(x)$ reaches its minimal value at x_0 if and only if $\frac{\partial}{\partial x_i} f(x_0) \leq 0$. □

The lemma can be proved by the knowledge of calculus and is omitted.

Theorem 4.2.2 shows L_c is a closed convex hull of L_{P_i}'s. The boundary of L_c consists of two kinds points (to see Fig. 4.5 for a schematic figure). Point A is located at the boundary of one ellipsoid, it is an extreme point; Point B is in a flat surface (hyperplane) and is not an extreme point. The hyperplane is the common tangent plane of several ellipsoids. Hence, bd L_c is a continuous differentiable curved surface with order one.

In the theory of convex analysis, many properties of a convex set can be described by its extreme points. Hence, it is advanced to get properties of a convex set from studying on its extreme points.

For $i = 1, 2, \ldots, N$, we denote

$$E_i = \text{bd}\, L_c \cap \text{bd}\, L_{P_i} = \left\{x;\ \ V_c(x) = x^T P_i x = 1\right\}.$$

The above analysis shows that E_i is a subset of extreme points of L_c. The further characteristic of E_i is depicted by the following Theorem 4.2.3.

Theorem 4.2.3 For $i = 1, 2, \ldots, N$,

$$E_i = \left\{x \in \text{bd}\, L_c : x^T Q_i^{-1} \left(Q_j - Q_i\right) Q_i^{-1} x \leq 0,\ \ j = 1, 2, \ldots, N\right\}.$$

Proof Without loss of generality, let $i = 1$. Then it needs us to prove only

$$E_1 = \left\{x \in \text{bd}\, L_c : x^T Q_1^{-1} \left(Q_j - Q_1\right) Q_1^{-1} x \leq 0,\ \ j = 1, 2, \ldots, N\right\}. \tag{4.2.7}$$

Firstly, when $j = 1$, it always holds $x^T Q_1^{-1} \left(Q_j - Q_1\right) Q_1^{-1} x \leq 0$, hence, we only consider the case $j \neq 1$. $V_c(x)$ is rewritten as follows

$$V_c(x) = \min \left\{ x^T \left[Q_1 + \sum_{j=2}^{N} \gamma_j \left(Q_j - Q_1\right) \right]^{-1} x,\ \ \sum_{j=2}^{N} \gamma_j \leq 1,\ \ \gamma_j \geq 0 \right\}.$$

We now denote $x^T P(\gamma)x$ in an alternative form. Denote

$$\varphi_x(\gamma_2, \ldots, \gamma_N) = x^T \left[Q_1 + \sum_{j=2}^{N} \gamma_j \left(Q_j - Q_1 \right) \right]^{-1} x = x^T P(\gamma) x, \tag{4.2.8}$$

where x is treated as a parameter. Define a set of $\mathbb{R}^{n-1}$ as follows

$$\Gamma_2 = \left\{ (\gamma_2, \ldots, \gamma_N) : \sum_{j=2}^{N} \gamma_j \leq 1, \ \ \gamma_j \geq 0 \right\},$$

Γ_2 can be treated as the projection of Γ on span $\{e_i, \ i = 2, 3, \ldots, n\}$, hence, is bounded and closed. An elements of Γ_2 is denoted by $\hat{\gamma}$. By the definition of $V_c(x)$, we have $V_c(x) = \min_{\hat{\gamma} \in \Gamma_2} \varphi_x(\hat{\gamma})$, or $\varphi_x(\hat{\gamma}) \geq V_c(x), \hat{\gamma} \in \Gamma_2$.

If $x \in E_1$, then $V_c(x) = x^T Q_1^{-1} x = 1$. It implies that the minimum value of function φ_x is obtained when $(\gamma_2, \ldots, \gamma_N) = (0, \ldots, 0)$, then by Lemma 4.2.2

$$\frac{\partial \varphi_x}{\partial \gamma_j}\bigg|_{(\gamma_2, \ldots \gamma_N) = (0, \ldots, 0)} \leq 0, \ \ j = 2, \ldots, N. \tag{4.2.9}$$

It is known that for an differentiable function matrix $A(t)$, $\frac{d}{dt} A^{-1}(t) = A^{-1}(t) \frac{dA(t)}{dt} A^{-1}(t)$. Using the rule, Inequality (4.2.8) yields

$$x^T Q_1^{-1} \left(Q_j - Q_1 \right) Q_1^{-1} x \leq 0, \quad j = 2, 3, \ldots, N. \tag{4.2.10}$$

It means that for $x \in E_1$, Inequality (4.2.10) holds.

Now, we prove that if $x \in$ db L_c and Inequality (4.2.10) holds, then $x \in$ db L_{P_1}, hence, $x \in E_1$.

Because

$$\frac{\partial \varphi_x}{\partial \gamma_j}\bigg|_{(\gamma_2, \ldots \gamma_N) = (0, \ldots, 0)} = x^T Q_1^{-1} \left(Q_j - Q_1 \right) Q_1^{-1} x, \quad j = 2, \ldots, N,$$

by Lemma 4.2.2, Inequality (4.2.9) implies that $\varphi_x(\gamma_2, \gamma_3, \ldots, \gamma_N)$ reaches its minimal value. Thus,

$$x^T P_1 x = \varphi_x(0, 0, \ldots, 0) = \min_{\hat{\gamma} \in \Gamma_2} \varphi_x(\hat{\gamma}) = V_c(x).$$

$x \in$ db L_c, consequently, $x^T P_1 x = 1, x \in$ db L_{P_1}.

In conclusion, we have

$$E_1 = \left\{ x \in \partial L_c : x^T Q_1^{-1} \left(Q_j - Q_1 \right) Q_1^{-1} x \leq 0, \ \ j = 2, \ldots, N \right\}.$$

The theorem is verified □

Theorem 4.2.4 Let $x_0 \in$ bd L_c. Suppose that $\gamma_k^* > 0$ for $k = 1, 2, \ldots, N_0$; and $\gamma_k^* = 0$ for $k = N_0 + 1, \ldots, N$, i.e., $Q(\gamma^*) = \sum_{k=1}^{N_0} \gamma_k^* Q_k$. Denote $x_k = Q_k Q^{-1}(\gamma^*) x_0$, $k = 1, 2, \ldots, N_0$, then

(1) $x_0 = \sum_{k=1}^{N_0} \gamma_k^* x_k$;
(2) $x_k \in$ bd L_{P_k};
(3) $\frac{\partial V_c(x_0)}{\partial x} = \frac{\partial V_c(x_k)}{\partial x_k} = 2\left(Q_k^{-1} x_k\right)^T = 2\left(Q(\gamma^*)^{-1} x_0\right)^T$[2];
(4) $V_c(x_k) = V_c(x_0) = 1, k = 1, 2, \ldots, N_0$.

Proof (1) By the definition, $x_k = Q_k Q(\gamma^*)^{-1} x_0$, $k = 1, 2, \ldots, N_0$, hence,

$$\sum_{k=1}^{N_0} \gamma_k^* x_k = \sum_{k=1}^{N_0} \gamma_k^* Q_k Q(\gamma^*)^{-1} x_0 = Q(\gamma^*) Q(\gamma^*)^{-1} x_0 = x_0.$$

(2) Recall the definition of the convex hull quadratic function V_c, it is equivalent to calculate:

$$\begin{aligned} \min \ & x_0^T \Big(\sum_{k=1}^{N} \gamma_k Q_k\Big)^{-1} x_0, \\ \text{s.t. } & \gamma = [\gamma_1, \gamma_2, \ldots, \gamma_N] \in \Gamma. \end{aligned} \tag{4.2.11}$$

We use the Lagrange multiplier method to solve the optimization problem. By introducing Lagrange multipliers α, β_k, and r_k, $k = 1, 2, \ldots, N$, the Lagrange function for Problem (4.2.11) is obtained as follows

$$L(\gamma, r, \alpha, \beta) = x_0^T \left(\sum_{k=1}^{N} \gamma_k Q_k\right)^{-1} x_0 + \alpha \left(\sum_{k=1}^{N} \gamma_k - 1\right) + \sum_{k=1}^{N} \beta_k \left(\gamma_k - r_k^2\right), \tag{4.2.12}$$

where $\beta = [\beta_1\ \beta_2 \ldots \beta_N]^T$ and $r = [r_1\ r_2 \ldots r_N]^T$. The last two terms in Eq. (4.2.12) are used to realize $\gamma \in \Gamma$. Thus the optimal solution $(\gamma^*, r^*, \alpha^*, \beta^*)$ of Problem (4.2.12) must satisfy

$$\begin{cases} \partial L / \partial \gamma_j |_{(\gamma^*, r^*, \alpha^*, \beta^*)} = 0, \\ \partial L / \partial r_j |_{(\gamma^*, r^*, \alpha^*, \beta^*)} = 0, \\ \partial L / \partial \alpha |_{(\gamma^*, r^*, \alpha^*, \beta^*)} = 0, \\ \partial L / \partial \beta_j |_{(\gamma^*, r^*, \alpha^*, \beta^*)} = 0, \end{cases} \quad j = 1, 2, \ldots, N. \tag{4.2.13}$$

[2] In accordance with the general notation, $\frac{\partial V_c(x_0)}{\partial x}$ represents $\left.\frac{\partial V_c(x)}{\partial x}\right|_{x=x_0}$.

By computing (4.2.13), it yields

$$x_0^T\left(\sum_{k=1}^{N}\gamma_k^* Q_k\right)^{-1} Q_j\left(\sum_{k=1}^{N}\gamma_k^* Q_k\right)^{-1} x_0-\alpha^*+\beta_j^*=0,\ \ j=1,2,\ldots,N, \tag{4.2.14a}$$

$$\beta_j^* r_j=0, j=1,2,\ldots,N, \tag{4.2.14b}$$

$$\sum_{k=1}^{N}\gamma_k^*=1, \tag{4.2.14c}$$

$$\gamma_j^*=\left(r_j^*\right)^2, j=1,2,\ldots,N. \tag{4.2.14d}$$

Since $\gamma_k^*>0$, $k=1,2,\ldots,N_0$, and $\gamma_k^*=0$, $k=N_0+1,\ldots,N$, we can require that $r_k>0$ for $k=1,2,\ldots,N_0$. By Eq. (4.2.14b), it yields

$$\beta_k^*=0, k=1,2,\ldots,N_0. \tag{4.2.15}$$

Substituting Eq. (4.2.15) into Eq. (4.2.14a), we have

$$x_0^T Q^{-1}\left(\gamma^*\right) Q_k Q^{-1}\left(\gamma^*\right) x_0=\alpha^*, k=1,2,\ldots,N_0, \tag{4.2.16}$$

hence,

$$x_0^T Q^{-1}\left(\gamma^*\right)\Big(\sum_{k=1}^{N_0}\gamma_k^* Q_k\Big) Q^{-1}\left(\gamma^*\right) x_0=\alpha^*,$$

that is

$$V_c\left(x_0\right)=x_0^T Q^{-1}\left(\gamma^*\right) x_0=\alpha^*. \tag{4.2.17}$$

Because $x_0\in\operatorname{bd} L_c$, $\alpha^*=1$.

It is easy to obtain that

$$x_k^T Q_k^{-1} x_k=\left(Q_k Q^{-1}\left(\gamma^*\right) x_0\right)^T Q_k^{-1}\left(Q_k Q^{-1}\left(\gamma^*\right) x_0\right)=x_0^T Q^{-1}\left(\gamma^*\right) Q_k Q^{-1}\left(\gamma^*\right) x_0\ .$$

Therefore, we have $x_k^T Q_k^{-1} x_k=1$, and $x_k\in\operatorname{bd} L_{P_k}$, $k=1,2,\ldots,N_0$.

(3) Since $x_0\in\operatorname{bd} L_c$, and L_c is a closed convex set, by Lemma 1.2.2, there exists an $h_0\in\mathbb{R}^n$, such that $1=\langle h_0,x_0\rangle\geq\langle h_0,x\rangle$ for all $x\in L_c$. For this x_0, there exists a γ^*, and then $P\left(\gamma^*\left(x_0\right)\right)$, such that $V_c\left(x_0\right)=x_0^T P\left(\gamma^*\left(x_0\right)\right) x_0$. Denote $P_0=P\left(\gamma^*\left(x_0\right)\right)$ for simplicity, note that P_0 depends on x_0.

From Lemma 4.2.1, $h_0^T x = 1$ is tangent to the ellipsoid $x^T P_0 x = 1$ at x_0, hence,

$$L_{h_0} \supset L_c \supset L_{P_0}, \tag{4.2.18}$$

in other words

$$x_0 = P_0^{-1} h_0, \quad h_0^T P_0^{-1} h_0 = 1. \tag{4.2.19}$$

Define a function $\phi(x) = \sqrt{x^T P_0 x}$, then $\phi(x)$ is positively homogeneous, and for every $x \in \text{bd}\, L_c$, we have

$$h_0^T x \le \sqrt{V_c(x)} \le \phi(x), \tag{4.2.20}$$

where we apply $h_0^T x \le 1 = \sqrt{V_c(x)}$. Since the three functions in Inequality (4.2.20) are all positively homogeneous, Inequality (4.2.20) holds for all $k \ge 0$ and $x \in \text{bd}\, L_c(k)$. It means the inequality holds for all $x \in \mathbb{R}^n$.

Since $\dfrac{\partial \phi(x)}{\partial x} = \dfrac{2x^T P_0}{2\sqrt{x^T P_0 x}}$, we have $\left.\dfrac{\partial \phi(x)}{\partial x}\right|_{x\,=\,x_0} = x_0^T P_0 = h_0^T$, for $x = x_0 \in \text{bd}\, L_c$. By the continuity of the quadratic function, we have

$$\phi\left(x_0 + \Delta x\right) = \phi\left(x_0\right) + h_0^T \Delta x + o\left(\|\Delta x\|\right),$$

where $o(\|\Delta x\|)$ is the higher order infinitesimal of $\|\Delta x\|$, and $\phi\left(x_0\right) = 1$. On the other hand, $h_0^T\left(x_0 + \Delta x\right) = 1 + h_0^T \Delta x$, by substituting $x = x_0 + \Delta x$ into Inequality (4.2.20), we obtain

$$1 + h_0^T \Delta x \le \sqrt{V_c\left(x_0 + \Delta x\right)} \le 1 + h_0^T \Delta x + o\left(\|\Delta x\|\right).$$

Hence,

$$h_0^T \Delta x \le \sqrt{V_c\left(x_0 + \Delta x\right)} - \sqrt{V_c\left(x_0\right)} \le h_0^T \Delta x + o\left(\|\Delta x\|\right).$$

Multiplying by $(\|\Delta x\|)^{-1}$ for every item, and letting $\|\Delta x\| \to 0$, it leads to

$$\left.\frac{\partial \sqrt{V_c(x)}}{\partial x}\right|_{x=x_0} = h_0^T = x_0^T P_0. \tag{4.2.21}$$

Thus,

$$\frac{\partial V_c(x)}{\partial x} = \frac{\partial\left(\sqrt{V_c(x)}\right)^2}{\partial x} = 2\frac{\partial \sqrt{V_c(x)}}{\partial x}.$$

Let $x = x_0$, we have $\left.\frac{\partial V_c(x)}{\partial x}\right|_{x=x_0} = 2\left.\frac{\partial \sqrt{V_c(x)}}{\partial x}\right|_{x=x_0} = 2(P_0 x_0)^T$, or $\left.\frac{\partial V_c(x)}{\partial x}\right|_{x=x_0} = 2(P(\gamma^*)\,x_0)^T$. By using $x_k = Q_k Q(\gamma^*)^{-1} x_0$, we at last obtain $\left.\frac{\partial V_c(x)}{\partial x}\right|_{x=x_0} = 2(P_k x_k)^T$.

(4) Since $x = \sum_{k=1}^{N_0} \gamma_k^* x_k$, the x_k and x_0 are on one hyperplane, and the hyperplane must be the tangent plane of $L_{P(\gamma^*)}$ at x_0. Therefore, $x_k \in$ db L_c and $V_c\,(x_0) = V_c\,(x_k) = x_k^T Q_k^{-1} x_k = 1$, for $k = 1, 2, \ldots, N_0$. □

Using $L_c\,(\rho) = \sqrt{\rho} L_c$, Theorem 4.2.4 has the following corollary.

Corollary 4.2.1 If $x_0^T P\,(\gamma^*)\,x_0 = \rho$, i.e., $x_0 \in$ bd $L_c\,(\rho)$, then the following conclusions are valid.

(1) $x_0 = \sum_{k=1}^{N_0} \gamma_k^* x_k$, with $x_k \in$ bd $L_{P_k}\,(\rho)$;
(2) $\frac{\partial V_c\,(x_0)}{\partial x} = \frac{\partial V_c\,(x_k)}{\partial x_k} = 2\big(Q_k^{-1} x_k\big)^T = 2\Big(Q\big(\gamma^*\big)^{-1} x_0\Big)^T$;
(3) $V_c\,(x_k) = V_c\,(x_0) = \rho, \quad k = 1, 2, \ldots, N_0$. □

Remark The equation $\frac{\partial V_c(x)}{\partial x} = 2P\left(\gamma^*\right) x$ cannot be simply obtained by differentiating the function $V_c(x) = x^T P\,(\gamma^*)\,x$. Since $\gamma^* = \gamma^*(x)$ is also a function of $x, P\,(\gamma^*)$ is not a constant matrix. □

From the proof of the last theorem, we can state another corollary.

Corollary 4.2.2 If $x_0 \in$ db L_c, and its responding γ^* satisfies that $\gamma_k^* \neq 0,\ \ k = 1, 2, \ldots, N_0, \gamma_k^* = 0,\ \ k = N_0 + 1, \ldots, N$. If there exist $x_k \in L_{P_k},\ \ k = 1, 2, \ldots, N_0$, such that $x_0 = \sum_{k=1}^{N} \gamma_k^* x_k$. Then

(1) $x_k \in$ db L_{P_k};
(2) If $k \neq j$, then $P_k \neq P_j$;
(3) $N_0 \leq n$;
(4) $x_0, x_k, k = 1, 2, \ldots, N_0$ on an identical hyperplane $f^T x = 1$, where f is a normal vector of the hyperplane. □

Corollary 4.2.2 has a version for $V_c\,(x_0) = \rho$, the readers are suggested to write the conclusions.

Some properties of the function $\gamma^*(x)$ are given in the following theorem.

Theorem 4.2.5 (1) For any $x \in$ dbL_{V_c}, if x can be expressed as $x = \sum_{k=1}^{N} \gamma_k x_k$ uniquely, with $x_j \in L_{P_j}$, $\gamma \in \Gamma$, then γ^* is a function of x.
(2) If γ^* is a function of x, then $\gamma^*(x)$ is a continuous function.

Proof (1) The conclusion is proved by contradiction. If $\gamma^*(x)$ is not a function of x, i.e., there is an x and more than one γ^* corresponding to it, e.g. $\gamma^{*\,(1)}$ and $\gamma^{*\,(2)}$. From Theorem 4.2.3, we have $x = \sum_{k=1}^{N} \gamma_k^{*(1)} x_k^{(1)} = \sum_{k=1}^{N} \gamma_k^{*(2)} x_k^{(2)}$. It contradicts with that x can be represented uniquely as $x = \sum_{k=1}^{N} \gamma_k x_k$. Thus γ^* is a function of x.

(2) It is also proved by contradiction. If $\gamma^*(x)$ is not continuous at x_0, then there exists a sequence $\{x^{(n)}\}$ such that $\lim\limits_{n\to\infty} x^{(n)} = x_0$, but $\gamma^*(x^{(n)})$ does not converge at $\gamma^*(x_0)$. Since for any n, $\gamma^*\left(x^{(n)}\right) \in \Gamma$, and Γ is a closed convex set, there exists a subsequence $\{x^{(1n)}\}$ of $\{x^{(n)}\}$ such that $\lim\limits_{n\to\infty} x^{(1n)} = x_0$, and $\lim\limits_{n\to\infty} \gamma^*\left(x^{(1n)}\right) = \gamma^{*(1)} \neq \gamma^*\left(x_0\right)$. We denote $\gamma^*\left(x^{(1n)}\right) = \gamma^{*(1n)}$. By Theorem 4.2.3, there exist $x_k^{(1n)} \in L_{P_k}$, $k = 1, 2, \ldots, N$, such that

$$x^{(1n)} = \sum_{k=1}^{J} \gamma_k^{*(1n)} x_k^{(1n)}. \tag{4.2.22}$$

Since Γ, L_c and L_{P_j} are all compact, an integer sequence $\{k_i;\ i = 1, 2, \ldots\}$ can be found such that the subsequence $\left\{x_{k_i}^{(1n)}\right\}$ of $\left\{x_{k_i}^{(1n)}\right\}$ satisfies $\left\{x_{k_i}^{(1n)}\right\} \subset \left\{x_k^{(1n)}\right\}$ and $\lim\limits_{i\to\infty} x_{k_i}^{(1n)} = x_k^{(2)}$ with $x_k^{(2)} \in L_{P_k}$. Since $\{x^{(1n)}\}$ and $\left\{\gamma^{*(1n)}\right\}$ are all convergent sequences, so $\lim\limits_{i\to\infty} \gamma_{k_i}^{*(1n)} = \gamma^{*(1)} \neq \gamma^*\left(x_0\right)$ and $\lim\limits_{i\to\infty} x^{(1k_i)} = x_0$. Taking limit on the both sides of (4.2.22) yields

$$\lim_{i\to\infty} x^{(1k_i)} = \lim_{i\to\infty} \sum_{k=1}^{N} \gamma_{k_i}^{*(1n)} x_{k_i}^{(1n)},$$

so

$$x_0 = \sum_{k=1}^{J} \gamma_k^{*(1)} x_k^{(2)}.$$

Owing to $\gamma^{*(1)} \neq \gamma^*\left(x_0\right)$,

$$x_0 = \sum_{k=1}^{N} \gamma_k^*\left(x_0\right) x_k = \sum_{k=1}^{N} \gamma_k^{*(1)} x_k^{(2)},$$

note $\gamma^{*(1)} \neq \gamma^*\left(x_0\right)$, we get two different expressions about x_0, it contradicts with the prerequisite, that is, the expression of x_0 is unique, so the assumption does not hold. That is, $\gamma^*(x)$ is a continuous function of x. □

Problems

1. Prove $\gamma^*(x) = \gamma^*(\alpha x)$, for all $\alpha \in \mathbb{R}$.
2. Prove $L_c(a) = \sqrt{a} L_c(1)$, for all $\alpha \in \mathbb{R}_+$.
3. Prove that $V_c(x)$ is a continuous function on $\mathbb{R}^n$.
4. Let $P_1 = \begin{bmatrix} 5 & 2 \\ 2 & 1 \end{bmatrix}$, $P_2 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$. Calculate $V_c(x)$.
5. Prove Corollary 4.2.1 and Corollary 4.2.2.
6. If $P \in \mathbb{R}^{n \times n}$ is a positive definite matrix, then the level set L_P is a strictly convex set.
7. Give an example that $x_0 \in L_c$, but x_0 can be expressed as two different equations of $x = \sum_{k=1}^{N} \gamma_k^{*(1)} x_k^{(1)} = \sum_{k=1}^{N} \gamma_k^{*(2)} x_k^{(2)}$.

4.3 Control of Linear Polytope Systems

This section considers control problems for the linear polytope system with control input described as follows

$$\dot{x}(t) \in \text{co}\,\{A_i x(t) + B_i u(t), i = 1, 2, \ldots, N\}, \tag{4.3.1}$$

where $x(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$ are the state and the input of the system, respectively, and $A_i \in \mathbb{R}^{n \times n}$, $B_i \in \mathbb{R}^{n \times m}$, $i = 1, 2, \ldots, N$. The main design objective is to establish a state feedback law $u(t) = K(x(t))$, such that the closed-loop polytope system

$$\dot{x}(t) \in \text{co}\,\{A_i x(t) + B_i K(x(t)), i = 1, 2, \ldots, N\} \tag{4.3.2}$$

is strong stable or is strong exponentially stable. From Inclusion (4.3.2), it can be seen that the design objective is equivalent to find the $u(t) = K(x(t))$ such that N linear systems $\dot{x}(t) = A_i x(t) + B_i K(x(t))$, $i = 1, 2, \ldots, N$ are stable, or exponentially stable, simultaneously.

The feedback law $K(x(t))$ given in this section is completed by using the convex hull quadratic function discussed in the last section, so it is not linear in general.

In order to facilitate the description, when it is no confusion, we will omit the argument (t) in $x(t)$ and $u(t)$, and omit $i = 1, 2, \ldots, N$ in Inclusions (4.3.1) and (4.3.2).

In order to provide convenience to readers, a conclusion for the linear system is given here.

Consider the linear system described by

$$\dot{x} = Ax + Bu,$$

there exists the linear feedback control $u = Kx$, such that the closed-loop system asymptotically stable if and only if there exist solutions X and Y the following linear matrix inequality

$$AX + XA^{\mathrm{T}} + BY + Y^{\mathrm{T}}B^{\mathrm{T}} < 0, \tag{4.3.3}$$

holds with X being positive definite. If Inequality (4.3.3) holds, then we have $K = YX^{-1}$.

The proof is direct, and it is omitted. It is emphasized that there exists a solution for Inequality (4.3.3) if and only if (A, B) is stabilizable. A further conclusion is: If (A, B) is not be stabilized, then for any feedback $u = K(x)$, even though $K(x)$ is nonlinear, it is impossible to make $\dot{x} = Ax + BK(x)$ stable.

4.3.1 Feedback Stabilizability for Linear Polytope Systems

Consider the linear polytope system descripted by Inclusion (4.3.1), and given positive definite matrices $P_j,\ j = 1, 2, \ldots, J$, denote $Q_j = P_j^{-1}$, constructing the convex hull quadratic function $V_c(x)$ by Definition 4.2.1. The meaning of layers $L_c(\alpha)$, $L_{P_j}(\alpha)$ are given in Sect. 4.2. Suppose that $\gamma^*(x)$ is a function throughout of this section.

Consider the set $\mathrm{bd}\, L_c(\alpha)$. If for any $x \in \mathrm{bd}\ L_c(\alpha)$, it holds $\frac{\partial V_c(x)}{\partial x} f(x) \le 0$ with arbitrary selection $f(t) \in \mathrm{co}\ \left(A_i x(t) + B_i K(x(t))\right)$, then by Lasalle invariable principle, any trajectory will tend to an invariable set in $L_c(\alpha)$. Since $L_c(\alpha) = \sqrt{\alpha} L_c$, if for $x \in \mathrm{bd}\ L_c$, it always holds $\frac{\partial V_c(x)}{\partial x} f(x) < 0$, then the polytope system must be strong stable. In what follows, this argument is main issue for further investigation.

In order to simplify the proofs afterward, first, given a direct result and its detailed proof, the purpose is to provide some ideas for similar proofs.

Theorem 4.3.1 If there exist matrices $P_j \in \mathbb{R}^{n\times n}$ and $F_j \in \mathbb{R}^{m\times n}$, with $P_j > 0$, $j = 1, 2, \ldots, J$, such that the inequalities

$$A_i Q_j + Q_j A_i^{\mathrm{T}} + B_i F_j + F_j^{\mathrm{T}} B_i^{\mathrm{T}} < 0, \tag{4.3.4}$$

hold for $i = 1, 2, \ldots, N$ simultaneously, where $Q_j = P_j^{-1}$, then there exists a state feedback $u(t) = K(x(t))$ such that the closed Inclusion (4.3.2) is strong stable.

Proof If there exist positive definite matrices $P_j \in \mathbb{R}^{n\times n}$ and matrices $F_j \in \mathbb{R}^{m\times n}$ $j = 1, 2, \ldots, J$, such that Inequalities (4.3.4) hold for every $i = 1, 2, \cdots, N$, then we construct $V_c(x)$ by Definition 4.2.1, and obtain $\gamma^* = \gamma^*(x)$.

Denote $F(\gamma^*) = \sum\limits_{j=1}^{J} \gamma_j^* F_j$ and $Q(\gamma^*) = \sum\limits_{j=1}^{J} \gamma_j^* Q_j$, the feedback law is constructed as follows

$$u = F\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x. \tag{4.3.5}$$

Since it has been assumed that $\gamma^*(x)$ is a function in the section, by Theorem 4.2.5, the feedback given in Eq. (4.3.5) is continuous.

Multiplying Inequality (4.3.4) from left and right by Q_j^{-1}, respectively, it yields

$$Q_j^{-1}A_i + A_i^{\mathrm{T}}Q_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}F_j^{\mathrm{T}}B_i^{\mathrm{T}}Q_j^{-1} < 0. \tag{4.3.6}$$

The proof of strong stability of the closed-loop system combined by Inclusion (4.3.2) and Eq. (4.3.5) is proceeded with two steps.

(1) If $x \in E_j = \mathrm{bd}\ L_c \cap \mathrm{bd}\ L_{P_j}$, then $V_c(x) = V_c\left(x_j\right) = x_j^{\mathrm{T}}P_jx_j = 1$. Because $\gamma^* = \gamma^*(x)$ is a function of x, $\gamma_j^* = 1$ and $\gamma_i^* = 0,\ \ i \neq j$. Thus,

$$\begin{aligned}\dot{V}_c(x) = \dot{V}_c\left(x_j\right) &\in \mathrm{co}\left\{\frac{\partial}{\partial x}V_c(x)\left[A_ix + B_iF\left(\gamma^*\right)Q^{-1}\left(\gamma^*\right)x\right]\right\}\Bigg|_{x=x_j}\\ &= \mathrm{co}\left\{2x_j^{\mathrm{T}}Q_j^{-1}\left(A_i + B_iF_jQ_j^{-1}\right)x_j\right\}\\ &= \mathrm{co}\left\{x_j^{\mathrm{T}}\left(Q_j^{-1}A_i + A_i^{\mathrm{T}}Q_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}F_j^{\mathrm{T}}B_i^{\mathrm{T}}Q_j^{-1}\right)x_j\right\}.\end{aligned}$$

By Inequality (4.3.6), the quadratic form $x_j^{\mathrm{T}}\left(Q_j^{-1}A_i + A_i^{\mathrm{T}}Q_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}F_j^{\mathrm{T}}B_i^{\mathrm{T}}Q_j^{-1}\right)x_j$ is negative definite, and E_j is a compact set, hence,

$$\max_{x_j \in E_j} x_j^{\mathrm{T}}\left(Q_j^{-1}A_i + A_i^{\mathrm{T}}Q_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}F_j^{\mathrm{T}}B_i^{\mathrm{T}}Q_j^{-1}\right)x_j < 0.$$

Furthermore,

$$\max_i \max_{x_j \in E_j} x_j^{\mathrm{T}}\left(Q_j^{-1}A_i + A_i^{\mathrm{T}}Q_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}F_j^{\mathrm{T}}B_i^{\mathrm{T}}Q_j^{-1}\right)x_j < 0.$$

It leads to $\dot{V}_c\left(x_j\right) < 0$.

(2) If $x \in \mathrm{bd}\ L_c$, by Conclusion (1) of Theorem 4.2.3, there exist $x_j \in E_j$, $j = 1, 2, \ldots, J$ such that $x = \sum_{j=1}^{J}\gamma_j^*x_j$. For simplicity, we assume that $\gamma_j^* > 0, j = 1, 2, \ldots, J_0$; $\gamma_j^* = 0, j = J_0 + 1, \ldots, J$. By Conclusion (3) of Theorem 4.2.3, we have $Q_j^{-1}x_j = Q(\gamma^*)^{-1}x$ and $V_c(x) = V_c\left(x_j\right) = 1$. By Conclusion (3) of Theorem 4.2.3 again, we have

$$\begin{aligned}\dot{V}_c(x) &\in \mathrm{co}\left\{\frac{\partial}{\partial x}V_c(x)\left[A_ix + B_iF\left(\gamma^*\right)Q\left(\gamma^*\right)^{-1}x\right]\right\}\\ &= \mathrm{co}\left\{2x^{\mathrm{T}}Q\left(\lambda^*\right)^{-1}\left[A_i + B_iF\left(\gamma^*\right)Q\left(\gamma^*\right)^{-1}x\right]\right\}.\end{aligned}$$

Because

$$
\begin{aligned}
&2x^{\mathrm{T}}Q(\lambda^*)^{-1}\left[A_i x + B_i F\left(\gamma^*\right)Q(\gamma^*)^{-1}x\right] \\
&\quad= \sum_{j=1}^{J_0} 2x^{\mathrm{T}}Q(\lambda^*)^{-1}\left[\gamma_j^* A_i x_j + \gamma_j^* B_i F_j Q(\gamma^*)^{-1}x\right] \\
&\quad= \sum_{j=1}^{J_0} 2x_j^{\mathrm{T}}Q_j^{-1}\left[\gamma_j^* A_i x_j + \gamma_j^* B_i F_j Q_j^{-1}x_j\right] \\
&\quad= \sum_{j=1}^{J_0} \gamma_j^* x_j^{\mathrm{T}}\left[Q_j^{-1}A_i + A_i^{\mathrm{T}}Q_j^{-1} + Q_j^{-1}B_i F_j Q_j^{-1} + Q_j^{-1}F_j^{\mathrm{T}}B_i^{\mathrm{T}}Q_j^{-1}\right]x_j \\
&\quad< 0,
\end{aligned}
$$

where the first equality is obtained by applying $x = \sum_{j=1}^{J_0}\gamma_j^* x_j$ and $F\left(\gamma^*\right) = \sum_{j=1}^{J_0}\gamma_j^* F_j$, the second equation is because $Q(\gamma^*)^{-1}x = Q_j^{-1}x_j$. Therefore, using the arguments made in Step (1), we have $V_c(x) < 0$ for all $x \in \mathrm{bd}\ L_c$. □

There are the following remarks about the above proof.

Remark 1 Readers can find that a series of simplifications have been made in the above proof. First, the discussion for any $L_c(\alpha)$ is simplified as that for the fixed L_c, and the scope of the discussion is narrowed; then the discussion for $x \in \mathrm{bd}\ L_c$ is restricted to that for $x_j \in E_j$. By Theorem 4.2.4, there are some useful properties of E_j, it is easy to carry on the discussion by using these properties. From the proof of Theorem 4.3.1, it can be seen as long as the conclusion holds for $x_j \in E_j$, it is convenient to extend it to $\mathrm{bd}L_c$. Hence, the key is to prove the conclusion for $x_j \in E_j$. Later on, we mainly prove the conclusion for $x_j \in E_j$, and that for $x \in \mathrm{bd}\ L_c$ is simplified or omitted. □

Remark 2 From the proof Step (1) of Theorem 4.3.1, the condition $Q_j^{-1}A_i + A_i^{\mathrm{T}}Q_j^{-1} + Q_j^{-1}B_i F_j Q_j^{-1} + Q_j^{-1}F_j^{\mathrm{T}}B_i^{\mathrm{T}}Q_j^{-1} < 0$ can be relaxed. In fact, it is sufficient for us to prove that $x_j^{\mathrm{T}}\left(Q_j^{-1}A_i + A_i^{\mathrm{T}}Q_j^{-1} + Q_j^{-1}B_i F_j Q_j^{-1} + Q_j^{-1}F_j^{\mathrm{T}}B_i^{\mathrm{T}}Q_j^{-1}\right)x_j < 0$ for x_j in E_j. □

Remark 3 When $a{>}0$, $\gamma^*(ax) = \gamma^*(x)$, the feedback $u = F\left(\gamma^*(x)\right)Q^{-1}\left(\gamma^*(x)\right)x$ is positively homogeneous for x. □

Remark 4 Solving inequalities $A_iQ_j{+}Q_jA_i^{\mathrm{T}}{+}B_iF_j{+}F_j^{\mathrm{T}}B_i^{\mathrm{T}} < 0$, for $i = 1, 2, \ldots, N$, is equivalent to find a common solution (X, Y) for the set of inequalities as follows

$$\begin{cases} A_1X + XA_1^{\mathrm{T}} + B_1Y + Y^{\mathrm{T}}B_1^{\mathrm{T}} < 0, \\ \qquad \cdots\cdots \\ A_NX + XA_N^{\mathrm{T}} + B_NY + Y^{\mathrm{T}}B_N^{\mathrm{T}} < 0, \end{cases}$$

where X is required to be positive definite. A sufficient and necessary condition is that (A_i, B_i), $i = 1, 2, \ldots, N$, can be stabilized, simultaneously, and a necessary condition is that (A_i, B_i), $i = 1, 2, \ldots, N$, can be stabilized. Therefore, the condition of Theorem 4.3.1 is also necessary for the linear feedback. □

Corollary 4.3.1 Under the conditions of Theorem 4.3.1, there exists a feedback such that Inclusion (4.3.2) under feedback $u(t) = K(x(t))$ is strong exponentially stable.

Proof Under the condition of Theorem 4.3.1, we will prove that there exists a positive real number β, such that for all $j = 1, 2, \ldots, J$,

$$Q_jA_i^{\mathrm{T}} + A_iQ_j + F_j^{\mathrm{T}}B_i^{\mathrm{T}} + B_iF_j \le -\beta Q_j. \tag{4.3.7}$$

with $Q_i = P_i^{-1}$. If inequality (4.3.7) is verified, then it will be obtained $\dot{V}_c(x) \le -\beta V_c(x)$, for all $x \in \operatorname{bd} L_c$ by the proof which is almost the same as that of Theorem 4.3.1. Hence, the closed-loop system is strongly exponentially stable.

Denote $Q_jA_i^{\mathrm{T}} + A_iQ_j + F_j^{\mathrm{T}}B_i^{\mathrm{T}} + B_iF_j = -W_{ij}$, then W_{ij} is a positive definite matrix for every $i \in \{1, 2, \ldots, N\}$ and every $j \in \{1, 2, \ldots, J\}$. It can be proved that there exists a matrix C such that $C^{\mathrm{T}}Q_jC$ and $C^{\mathrm{T}}W_{ij}C$ are simultaneously diagonalizable; consequently, there exists a $\beta_{ij} > 0$ such that $C^{\mathrm{T}}W_{ij}C \ge \beta_{ij}C^{\mathrm{T}}Q_jC$, that is,

$$Q_jA_i^{\mathrm{T}} + A_iQ_j + F_j^{\mathrm{T}}B_i^{\mathrm{T}} + B_iF_j \le -\beta_{ij}Q_j.$$

Let $\beta = \min\left\{\beta_{ij}, i = 1, 2, \ldots, N; j = 1, 2, \ldots, J\right\}$, then Inequality (4.3.7) holds. □

From Theorem 4.3.1, nonlinear feedback Eq. (4.3.5) does not have any advantage. This is the main reason to give Theorem 4.3.2, which can relax slightly the conditions of the above theorem.

Theorem 4.3.2 Consider Inc. (4.3.1), if there exist matrices $P_j \in \mathbb{R}^{n\times n}$ and $F_j \in \mathbb{R}^{m\times n}$, with $P_j > 0$, $j = 1, 2, \ldots, J$, and constant $\beta > 0$ and $\lambda_{ijk} \ge 0$, $i = 1, 2, \ldots, N; j, k = 1, 2, \ldots, J$ such that

$$Q_jA_i^{\mathrm{T}} + A_iQ_j + F_j^{\mathrm{T}}B_i^{\mathrm{T}} + B_iF_j \le \sum_{k=1}^{J} \lambda_{ijk}\left(Q_k - Q_j\right) - \beta Q_j \tag{4.3.8}$$

where $Q_j = P_j^{-1}$, then there exists a feedback $u(t) = K(x(t))$, such that the closed-loop system is strongly exponentially stable.

Proof Constructing the function $V_c(x)$ defined by the Definition 4.2.1. Based on the above remark, $\gamma^* = \gamma^*(x)$ is a function. We can then construct $F(\gamma^*) = \sum_{j=1}^{J} \gamma_j^* F_j$, $Q(\gamma^*) = \sum_{j=1}^{J} \gamma_j^* Q_j$ and design a continuous feedback control law

$$u = F\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x$$

which is the same as Eq. (4.3.5).

Multiplying Inequality (4.3.8) from left and from right sides by Q_j^{-1}, it yields

$$\begin{aligned} & A_i^{\mathrm{T}} Q_j^{-1} + Q_j^{-1} A_i + Q_j^{-1} F_j^{\mathrm{T}} B_i^{\mathrm{T}} Q_j^{-1} + Q_j^{-1} B_i F_j Q_j^{-1} \\ & \leq \sum_{k=1}^{J} \lambda_{ijk} Q_j^{-1}\left(Q_k - Q_j\right) Q_j^{-1} - \beta Q_j^{-1} . \end{aligned}$$

We give the proof with two steps as those of Theorem 4.3.1.

(1) If $x \in E_j = \mathrm{bd}\, L_c \cap \mathrm{bd}\, L_{P_j}$, by the definition of E_j, $V_c(x) = x^T Q_j^{-1} x = 1$, the jth component of γ^* is 1, the others are all 0. Hence, $F(\gamma^*)\, Q^{-1}(\gamma^*) = F_j Q_j^{-1}$ and $x = x_j$, and $\frac{\partial}{\partial x} V_c(x) = \frac{\partial}{\partial x_j} V_c\left(x_j\right) = \left(2 Q_j^{-1} x_j\right)^T$. The derivation along with the trajectory of the closed-loop system obtained from Inclusion (4.3.1) and Eq. (4.3.5) satisfies

$$\begin{aligned} \dot{V}_c(x) & \in \mathrm{co}\left\{\frac{\partial}{\partial x} V_c(x)\left[A_i x + B_i F\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x\right]\right\} \\ & = \mathrm{co}\left\{\frac{\partial}{\partial x} V_c(x)\left(A_i x_j + B_i F_j Q_j{}^{-1} x_j\right)\right\} \\ & = \mathrm{co}\left\{2 x_j^T Q_j^{-1}\left(A_i + B_i F_j Q_j{}^{-1}\right) x_j\right\} \\ & = \mathrm{co}\left\{x_j^T\left(Q_j^{-1} A_i + A_i Q_j^{-1} + Q_j^{-1} B_i F_j Q_j{}^{-1} + Q_j^{-1} F_j^T B_j^T Q_j{}^{-1}\right) x_j\right\} . \end{aligned}$$

By Theorem 4.2.2, we get

$$\begin{aligned} \dot{V}_c(x) & \leq \max_{x_j \in E_j}\left\{x_j^T\left(Q_j^{-1} A_i + A_i Q_j^{-1} + Q_j^{-1} B_i F_j Q_j{}^{-1} + Q_j^{-1} F_j^T B_j^T Q_j{}^{-1}\right) x_j\right\} \\ & \leq \max_{x_j \in E_j}\left\{x_j^T\left(\sum_{k=1}^{N} \lambda_{ijk} Q_j^{-1}\left(Q_k - Q_j\right) Q_j^{-1} - \beta Q_j^{-1}\right) x_j\right\} \\ & \leq \max_{x_j \in E_j}\left\{-\beta x_j^T Q_j^{-1} x_j\right\} \\ & \leq -\beta V_c(x) . \end{aligned}$$

(2) For $x \in \text{bd } L_{V_c}$, using the arguments as done in Step (2) of Theorem 4.3.1, we have

$$\dot{V}_c(x) \leq -\beta V_c(x),$$

the detailed proof is omitted.

According to the above two steps, under the feedback $u = F(\gamma^*) \, Q(\gamma^*)^{-1} x$, the closed-loop system is strongly exponentially stable. □

For Theorem 4.3.2, we have the following two remarks.

Remark 1 By the proof of Corollary 4.3.1, the Inequality (4.3.8) is equivalent to

$$Q_j A_i^T + A_i Q_j + F_j^T B_i^T + B_i F_j < \sum_{k=1}^{J} \lambda_{ijk} \left(Q_k - Q_j \right). \tag{4.3.9}$$

□

Remark 2 Inequality (4.3.9) is easier to be satisfied than Inequality (4.3.4), since it does not need that $\sum_{k=1}^{J} \lambda_{ijk} \left(Q_k - Q_j \right)$ is negative definite. The reason that the condition can be relaxed in Theorem 4.3.2 is that we do not need Inequality (4.3.4) holding for all $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, J$, but we just need

$$x_j^T \left(Q_j A_i^T + A_i Q_j + F_j^T B_i^T + B_i F_j \right) x_j < 0 \tag{4.3.10}$$

for $x_j \in E_j = \text{bd } L_c \cap \text{bd } L_{P_j}$. By Theorem 4.2.2

$$x_j^T Q_j^{-1} \left(Q_k - Q_j \right) Q_j^{-1} x_j \leq 0. \tag{4.3.11}$$

Inequality (4.3.11) does not imply $Q_j^{-1} \left(Q_k - Q_j \right) Q_j^{-1} \leq 0$. It is difficult to verify (4.3.10), but Inequality (4.3.9) is a linear matrix inequality, it can be solved by existing software. □

The following corollary also illustrates that the condition of Theorem 4.3.2 is lower than that of Theorem 4.3.1.

Corollary 4.3.2 If the condition of Theorem 4.3.1 is satisfied, there exists a nonzero λ_{ijk} such that Inequality (4.3.8) holds.

Proof Let $\lambda = [\lambda_{111} \; \lambda_{112} \ldots \lambda_{11J} \; \lambda_{121} \ldots \lambda_{12J} \ldots \lambda_{NJJ}]^T$ and define a function of λ as follows

$$f(\lambda) = A_i^T Q_j^{-1} + Q_j^{-1} A_i + Q_j^{-1} F_j^T B_i^T Q_j^{-1} + Q_j^{-1} B_i F_j Q_j^{-1} \\ - \sum_{k=1}^{J} \lambda_{ijk} Q_j^{-1} \left(Q_k - Q_j \right) Q_j^{-1} + \beta Q_j^{-1}.$$

Then it is a linear mapping for λ, so it is continuous. Since $f(0) < 0$, there exists a $\delta > 0$, such that $f(\lambda) < 0$ when $\|\lambda\| < \delta$. □

4.3.2 Feedback Stabilization for Linear Polytope Systems with Time-delay

A real process has to take time, hence, a real control always time-delay during the operation. It is natural to consider the time-delay for a control system design. Sometimes, we do not consider time-delay because we have no effective method to handle it rather than it does not exist. This subsection deals with the linear polytope systems with time-delay. Readers have found in Sect. 2.3 that it becomes easy for differential inclusions since we only require the solution exists almost every time.

Consider the system described by Inclusion (4.3.12). Comparing with Inc. (4.3.1), it has an extra delayed state $x(t-\tau)$, where $\tau > 0$ is called time-delay.

$$\begin{cases} \dot{x}(t) \in \mathrm{co}\left\{A_i x(t) + A_{di} x(t-\tau) + B_i u(t)\right\}, \\ x(t) = \phi(t), \quad t \in [-\tau, 0] \ . \end{cases} \tag{4.3.12}$$

In System (4.3.12), $i = 1, 2, \ldots, N$, $x(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$ are the state and the input, respectively, for $t \in [-\tau, 0]$, $\phi(t)$ is a continuous vector-valued function, and gives the initial state of the system, and $A_i \in \mathbb{R}^{n\times n}$, $A_{di} \in \mathbb{R}^{n\times n}$, $B_i \in \mathbb{R}^{n\times m}$. The design objective is still to find a state feedback law $u(t) = K(x(t))$, such that the closed-loop system

$$\dot{x}(t) \in \mathrm{co}\ \left\{A_i x(t) + A_{di} x(t-\tau) + B_i K(x(t))\right\}, i = 1, 2, \ldots, N \tag{4.3.13}$$

is strongly stable.

To start with, we give a Lemma

Lemma 4.3.1 Let $f(t) \in \mathbb{R}^+$ be a continuous function. If there exist two constants, $\alpha,\ \beta$ such that $\beta > \alpha > 0$ and

$$\dot{f}(t) \leq -\beta f(t) + \alpha \sup_{t-\tau\leq s\leq t} f(s)$$

for $t \geq t_0$. Let ν be the unique positive solution of $\nu - \beta + \alpha e^{\nu\tau} = 0$, then

$$f(t) \leq \left[\sup_{t_0-\tau\leq s\leq t_0} f(s)\right] e^{-\nu(t-t_0)}.$$

Lemma 4.3.1 is called Halany inequality, it can be direct verified, or, referred to (Filippov 1988).

Theorem 4.3.3 If there exist positive definite matrices $P_j \in \mathbb{R}^{n\times n}$ and matrices $F_j \in \mathbb{R}^{m\times n}, j = 1, 2, \ldots, J$, such that for $i = 1, 2, \ldots, N$

$$A_iQ_j + Q_jA_i^T + B_iF_j + F_j^TB_i^T + A_{di}Q_jA_{di}^T \le -\beta Q_j, \tag{4.3.14}$$

where $\beta > 1$, and $Q_j = P_j^{-1}$. Then there exists a feedback $u(t) = K(x(t))$ such that System (4.3.13) is strongly exponentially stable.

Proof We continue to apply the notations used in Theorem 4.3.1, i.e., $F(\gamma^*) = \sum_{j=1}^{J} \gamma_j^* F_j$ and $Q(\gamma^*) = \sum_{j=1}^{J} \gamma_j^* Q_j$, the feedback law is

$$u = F\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x.$$

We now prove exponential stability for the closed-loop system (4.3.13). Multiplying Inequality (4.3.14) from left and from right by Q_j^{-1} yields

$$A_i^TQ_j^{-1} + Q_j^{-1}A_i + Q_j^{-1}F_j^TB_i^TQ_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}A_{di}Q_jA_{di}^TQ_j^{-1} \le -\beta Q_j^{-1}.$$

First, we assume $x \in E_j$, it can be deduced that $V_c(x) = V_c\left(x_j\right) = x_j^TP_jx_j = 1$ and $\dfrac{\partial V_c(x)}{\partial x} = \dfrac{\partial V_c\left(x_j\right)}{\partial x_j} = 2Q_j^{-1}x_j$. Therefore, for $j \in \{1, 2, \ldots, J\}$, the derivation of $V_c(x)$ along the trajectory of the closed-loop system (4.3.13) is

$$\dot{V}_c(x) \in \text{co}\left\{\frac{\partial}{\partial x}V_c(x)^T\left[A_ix + A_{id}x(t-\tau) + B_iF\left(\gamma^*\right)Q\left(\gamma^*\right)^{-1}x\right],\ i{=}1, 2, \ldots, N\right\}. \tag{4.3.15}$$

For the sake of simplicity, we omit (t) in $x(t)$, only maintain $(t - \tau)$ in the term with time-delay. Since $x \in E_j$, that is, $\gamma_j^* = 1,\ \ \gamma_k^* = 0,\ \ k \neq j$, then we can obtain

$$\begin{aligned}
&\frac{\partial}{\partial x_j}V_c\left(x_j\right)\left[A_ix_j + A_{id}x_j(t-\tau) + B_iF_jQ_j^{-1}x_j\right] \\
&\quad = 2x_j^TQ_j^{-1}\left[A_ix_j + A_{id}x_j(t-\tau) + B_iF_jQ_j^{-1}x_j\right] \\
&\quad = 2x_j^TQ_j^{-1}\left[A_ix_j + B_iF_jQ_j^{-1}x_j\right] + 2x_j^TQ_j^{-1}A_{id}x_j(t-\tau).
\end{aligned}$$

By Relation (4.1.5), we obtain

$$2x_j^TQ_j^{-1}A_{di}x_j(t-\tau) \le x_j^TQ_j^{-1}A_{di}Q_jA_{di}^TQ_j^{-1}x_j + x_j^T(t-\tau)Q_j^{-1}x_j(t-\tau).$$

Hence,

$$\begin{aligned}
&\frac{\partial}{\partial x_j} V_c\left(x_j\right)\left[A_i x_j + A_{di} x_j\left(t-\tau\right) + B_i F_j Q_j^{-1} x_j\right] \\
&\leq x_j^T\left[Q_j^{-1} A_i + A_i^T Q_j^{-1} + Q_j^{-1} B_i F_j Q_j^{-1} + Q_j^{-1} F_j^T B_j^T Q_j^{-1} + Q_j^{-1} A_{di} Q_j A_{di} Q_j^{-1}\right] x_j \\
&\quad + x_j^T\left(t-\tau\right) Q_j^{-1} x_j\left(t-\tau\right).
\end{aligned}$$

By the conditions of the theorem, for every $j = 1, 2, \ldots, J$, it holds

$$A_i^T Q_j^{-1} + Q_j^{-1} A_i + Q_j^{-1} F_j^T B_i^T Q_j^{-1} + Q_j^{-1} B_i F_j Q_j^{-1} + Q_j^{-1} A_{id} Q_j A_{id}^T Q_j^{-1} \leq \beta Q_j^{-1},$$

thus

$$\begin{aligned}
\dot{V}_c\left(x_j\right) &\leq -\beta x_j^T Q_j^{-1} x_j + x_j^T\left(t-\tau\right) Q_j^{-1} x_j\left(t-\tau\right) \\
&= -\beta V_c\left(x_j\right) + V_c\left(x_j\left(t-\tau\right)\right).
\end{aligned}$$

Using the arguments as done in Theorem 4.3.1, it can be deduced that for all $x \in \operatorname{bd} L_{V_c}$,

$$\dot{V}_c(x) \leq -\beta V_c(x) + V_c\left(x\left(t-\tau\right)\right).$$

Since $\beta > 1$, by Lemma 4.3.1, we have $V\left(x(t)\right) \leq \sup_{s\in[-\tau,0]} \left|\phi(s)\right| e^{-\nu\tau}$, where ν is the unique solution of $\nu - \beta + e^{\nu\tau} = 0$. The closed-loop system is strongly exponentially stable. □

Similar conclusion can be established by using the arguments as that completed in Theorem 4.3.2. The proof is left to readers. We just list the conclusions below.

Corollary 4.3.3 Consider system (4.3.12). If there exist matrices $P_j \in \mathbb{R}^{n\times n}$ and $F_j \in \mathbb{R}^{m\times n}$, with $P_j > 0, j = 1, 2, \ldots, J$, and constants $\lambda_{ijk} \geq 0$ and $\beta > 1$, such that

$$Q_j A_i^T + A_i Q_j + F_j^T B_i^T + B_i F_j + A_{di} Q_j A_{di}^T \leq \sum_{k=1}^{J} \lambda_{ijk}\left(Q_k - Q_j\right) - \beta Q_j$$

where $Q_j = P_j^{-1}$ for all $i = 1, 2, \ldots, N$, then there exists a feedback law $u(t) = K\left(x(t)\right) = F\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x(t)$, such that the closed-loop system is strongly exponentially stable. □

In this subsection, we apply Halany inequality to deal with the stability of linear polytope systems with time-delay. In Lemma 4.3.1, the requirement $\beta > 1$ is to assure the solution ν of equation $\nu - \beta + e^{\nu\tau} = 0$ satisfies $\nu > 0$. In fact, the Halany inequality can be replaced by different inequalities, then we can obtain different conditions for the existence of feedback law to stabilize the inclusion. However, the process of the proof is the same, we always prove the case of $x \in E_j$ firstly and then extend the result to $\operatorname{bd} L_c$.

4.3.3 Disturbance Rejection for Linear Polytope Systems

This subsection considers another designing problem for control systems described by differential inclusions, we now deal with disturbance rejection. Let us consider the following linear differential inclusion systems with disturbance:

$$\begin{bmatrix} \dot{x} \\ y \end{bmatrix} \in \operatorname{co}\left\{ \begin{bmatrix} A_i x + B_i u + T_i \omega \\ C_i x \end{bmatrix},\ \ i = 1, 2, \ldots, N \right\}, \tag{4.3.16}$$

where $\omega = \omega(t) \in \mathbb{R}^q$ is called disturbance of the system, which is uncontrollable and unmeasurable, $T_i \in \mathbb{R}^{n\times q}$ is the disturbance gain, $y = y(t) \in \mathbb{R}^r$ is the output of the system, and $C_i \in \mathbb{R}^{n\times q}$ is the output matrix. The meanings of other notations in Inclusion (4.3.16) are the same as those of Inclusion (4.3.2). The control objective is to find feedback $u(t) = K(x(t))$, such that the effect of the disturbance ω to the state $x(t)$ or output $y(t)$ is as small as possible.

Theorem 4.3.4 If there exist matrices $P_j \in \mathbb{R}^{n\times n}$ and $F_j \in \mathbb{R}^{m\times n}$, with $P_j > 0$, $j = 1, 2, \ldots, J$, such that for all $i = 1, 2, \ldots, N$,

$$\begin{bmatrix} M_{ij} & T_i \\ T_i^T & -I \end{bmatrix} \le 0, \tag{4.3.17}$$

where $M_{ij} = A_i Q_j + Q_j A_i^T + B_i F_j + F_j^T B_i^T - \sum_{k=1}^{J} \lambda_{ijk} \left(Q_k - Q_j \right)$, $Q_j = P_j^{-1}$. If $\|\omega(t)\|_2^2 = \int_0^\infty \omega^T(t)\omega(t)dt \le 1$, then by the feedback $u(t) = F\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x(t)$, where $F\left(\gamma^*\right)$, $Q^{-1}\left(\gamma^*\right)$ are defined before Eq. (4.3.5), then the set L_c is attractive by the meaning that every trajectory of the closed-loop system starting from the origin will be restricted in L_c, i.e. with the initial condition $x(t) \in L_c$, for $t \in [0,\ \infty)$.

Proof Given $V_c(x)$ by Definition 4.2.1 in the above subsection, consider the derivative of $V_c(x)$ along the trajectories of the closed-loop system (4.3.16) with $u = F\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x$, we have

$$\dot{V}_c(x) \in \text{co}\left\{\frac{\partial}{\partial x}V_c(x)\Big[A_i x + B_i F\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x + T_i\omega\Big\}\right. .$$

First, let $x \in E_j$. Then we have $V_c(x) = V_c\left(x_j\right) = x_j^T P_j x_j = 1$ and $\frac{\partial V_c(x)}{\partial x} = \frac{\partial V_c\left(x_j\right)}{\partial x_j} = 2Q_j^{-1}x_j$. By Inequality (4.3.17), we have $M_{ij} + T_i T_i^T \leq 0$, note

$$M_{ij} = A_i Q_j + Q_j A_i^T + B_i F_j + F_j^T B_i^T - \sum_{k=1}^{J}\lambda_{ijk}\left(Q_k - Q_j\right)$$

multiplying M_{ij} from left and right sides by Q_j^{-1}, it yields

$$\begin{aligned} &Q_j^{-1}A_i + A_i^T Q_j^{-1} + Q_j^{-1}B_i F_j Q_j^{-1} + Q_j^{-1}F_j^T B_i^T Q_j^{-1} \\ &\quad - \sum_{k=1}^{J}\lambda_{ijk}Q_j^{-1}\left(Q_k - Q_j\right)Q_j^{-1} + Q_j^{-1}T_i T_i^T Q_j^{-1} \leq 0. \end{aligned}$$

If $x \in E_j$, then

$$\begin{aligned} &\frac{\partial}{\partial x}V_c(x)\left[A_i x + B_i F\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x + T_i\omega\right]\Big|_{x=x_j} \\ &\quad = 2x_j^T Q_j^{-1}\left[A_i x_j + B_i F_j Q_j^{-1}x_j + T_i\omega\right] \\ &\quad = x_j^T\left[Q_j^{-1}A_i + A_i^T Q_j^{-1} + Q_j^{-1}B_i F_j Q_j^{-1} + Q_j^{-1}F_j^T B_j^T Q_j^{-1}\right]x_j + 2x_j^T Q_j^{-1}T_i\omega. \end{aligned}$$

By the inequality (4.1.5), we have

$$2x_j^T Q_j^{-1}T_i\omega \leq x_j^T Q_j^{-1}T_i T_i^T Q_j^{-1}x_j + \omega^T\omega.$$

Thus, for $x \in E_j$, by Theorem 4.2.3, it holds

$$\dot{V}_c\left(x_j\right) \leq x_j^T\left(\sum_{k=1}^{J}\lambda_{ijk}Q_j^{-1}\left(Q_j - Q_k\right)Q_j^{-1}\right)x_j + \omega^T\omega \leq \omega^T\omega.$$

For $x \in \text{bd}\ L_c$, using the arguments as in Theorem 4.3.1, we have $x = \sum_{j=1}^{J_0}\gamma_j^* x_j$ where the meaning of J_0 can be found in the proof of Theorem 4.3.1. For $x_j \in E_j$, it yields

$$\begin{aligned}
\dot{V}_c(x) &= \frac{\partial}{\partial x} V_c(x)\left[A_i x + B_i F\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x + T_i \omega\right] \\
&= 2x^T Q^{-1}\left(\gamma^*\right)\left[\sum_{j=1}^{J_0} \gamma_j^* A_i x_j + \sum_{j=1}^{J_0} \gamma_j^* B_i F_j Q^{-1}\left(\gamma^*\right) x + T_i \omega\right] \\
&= 2\left[\sum_{j=1}^{J_0} \gamma_j^* x^T Q^{-1}\left(\gamma^*\right) A_i x_j + \sum_{j=1}^{J_0} \gamma_j^* x^T Q^{-1}\left(\gamma^*\right) B_i F_j Q_j^{-1} x_j\right] \\
&\quad + 2x^T Q^{-1}\left(\gamma^*\right) T_i \omega \\
&\le 2\left[\sum_{j=1}^{J_0} \gamma_j^* x_j^T Q_j^{-1} A_i x_j + \sum_{j=1}^{J_0} \gamma_j^* x_j^T Q_j^{-1} B_i F_j Q_j^{-1} x_j\right] \\
&\quad + x^T Q^{-1}\left(\gamma^*\right) T_i T_i^T Q^{-1}\left(\gamma^*\right) x + \omega^T \omega \\
&\le 2\left[\sum_{j=1}^{J_0} \gamma_j^* x_j^T \left(Q_j^{-1} A_i + Q_j^{-1} B_i F_j Q_j^{-1}\right) x_j\right] + \sum_{j=1}^{J_0} \gamma_j^* x_j^T Q_j^{-1} T_i T_i^T Q_j^{-1} x_j + \omega^T \omega.
\end{aligned}$$

According to the above discussion for $x \in E_j$, it is obtained similarly

$$\begin{aligned}
&2\left[\sum_{j=1}^{J_0} \gamma_j^* x_j^T \left(Q_j^{-1} A_i + Q_j^{-1} B_i F_j Q_j^{-1}\right) x_j\right] + \sum_{j=1}^{J_0} \gamma_j^* x_j^T Q_j^{-1} T_i T_i^T Q_j^{-1} x_j \\
&\qquad = 2\left[\sum_{j=1}^{J_0} \gamma_j^* x_j^T \left(Q_j^{-1} A_i + Q_j^{-1} B_i F_j Q_j^{-1} + Q_j^{-1} T_i T_i^T Q_j^{-1}\right) x_j\right] \\
&\qquad \le 0,
\end{aligned}$$

so

$$\dot{V}_c(x) \le \omega^T \omega. \tag{4.3.18}$$

By integrating the both sides of Inequality (4.3.18), it yields

$$V_c\left(x(t)\right) - V_c\left(x(0)\right) \le \int_0^t \omega^T\left(\tau\right) \omega\left(\tau\right) d\tau \le 1. \tag{4.3.19}$$

Since $V_c\left(x(0)\right) = V_c(0) = 0$, and $V_c\left(x(t)\right) \le 1$, it holds $x(t) \in L_c$. □

The theorem illustrates that if $\|\omega\| \le 1$ then the trajectories start from the origin will stay in L_c. The result can be extended. If $x(0) \neq 0$, then we have $V_c\left(x(t)\right) - V_c\left(x(0)\right) \le 1$. The inequality is equivalent to $x(t) \in L_c\left(\left(L_c x(0)\right) + 1\right)$. It supports that if the conditions of theorem are satisfied, the trajectories will not go too far from their original position although the disturbance exists.

The following theorem is about the attraction for the system existed disturbance.

Theorem 4.3.5 Consider System (4.3.16). If there exist matrices $P_j \in \mathbb{R}^{n\times n}$ and $F_j \in \mathbb{R}^{m\times n}$, with $P_j > 0, j = 1, 2, \ldots, J$, and $\beta > 0$, such that for all $i = 1, 2, \ldots, N$

$$\begin{bmatrix} M_{ij} + \beta Q_j & T_i \\ T_i^T & -I \end{bmatrix} \leq 0, \tag{4.3.20}$$

where M_{ij} is defined as that in Theorem 4.3.4, $Q_j = P_j^{-1}$. If $\|\omega(t)\|_2^2 = \int_0^\infty \omega^T(t)\omega(t)dt \leq 1$, then there exists the feedback $u(t) = F(\gamma^*)\,Q^{-1}(\gamma^*)\,x(t)$, such that the level set L_c of the closed-loop system is attractive.

Proof Inequality (4.3.20) implies $M_{ij} + \beta Q_j + T_i T_i^T \leq 0$. Multiplying it from left and right sides by Q_j^{-1}, it yields

$$\begin{aligned} &Q_j^{-1}A_i + A_i^T Q_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}F_j^TB_i^TQ_j^{-1} - \sum_{k=1}^{J}\lambda_{ijk}Q_j^{-1}\left(Q_k - Q_j\right)Q_j^{-1} \\ &\quad + Q_j^{-1}T_iT_i^TQ_j^{-1} \leq -\beta Q_j^{-1}. \end{aligned}$$

$V_c(x)$ is defined by Definition 4.2.1, then

$$\dot{V}_c(x) \in \mathrm{co}\left\{\frac{\partial}{\partial x}V_c(x)\Big[A_ix + B_iF\left(\gamma^*\right)Q^{-1}\left(\gamma^*\right)x + T_i\omega\right\}.$$

Using the arguments as done in Theorem 4.3.4, it can be deduced that for $x \in E_j$,

$$\begin{aligned} &\frac{\partial}{\partial x}V_c(x)\left[A_ix + B_iF\left(\gamma^*\right)Q^{-1}\left(\gamma^*\right)x + T_i\omega\right]\Big|_{x=x_j} \\ &\quad = x_j^T\left[Q_j^{-1}A_i + A_i^TQ_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}F_j^TB_j^TQ_j^{-1}\right]x_j + 2x_j^TQ_j^{-1}T_i\omega, \end{aligned}$$

and since,

$$2x_j^TQ_j^{-1}T_i\omega \leq x_j^TQ_j^{-1}T_iT_i^TQ_j^{-1}x_j + \omega^T\omega,$$

we have

$$\dot{V}\left(x(t)\right) \leq -\beta V\left(x(t)\right) + \omega^T(t)\omega(t). \tag{4.3.21}$$

Using the arguments as done in Theorem 4.3.4, it can be deduced that Inequality (4.3.21) holds for $x \in \operatorname{bd} L_c$. Furthermore, it can hold for every $x \in \mathbb{R}^n$. Consequently, we have

$$V(x(t)) \le e^{-\beta t} V(x(0)) + \int_0^t e^{-\beta(t-\tau)} \omega^T(\tau)\,\omega(\tau)\,d\tau.$$

Since $e^{-\beta(t-\tau)} \le 1 \quad (\tau \le t)$, it yields

$$V(x(t)) \le e^{-\beta t} V(x(0)) + \int_0^t \omega^T(\tau)\,\omega(\tau)\,d\tau \le e^{-\beta t} V(x(0)) + 1.$$

It implies $\overline{\lim_t} V(x(t)) \le 1$, i.e., $x(t)$ will trend to L_c. □

By the discussion of Corollary 4.3.1, if for all $j = 1, 2, \ldots, J;\ i = 1, 2, \ldots, N$, inequalities $M_{ij} + T_i T_i^T < 0$ hold, then there exists $\beta > 0$, such that $M_{ij} + \beta Q_j + T_i T_i^T \le 0$. Consequently, we have the following corollary.

Corollary 4.3.4 Consider system (4.3.16). If there exist matrices $P_j \in \mathbb{R}^{n\times n}$ and $F_j \in \mathbb{R}^{m\times n}$, with $P_j > 0, j = 1, 2, \cdots, J$, such that for all $i = 1, 2, \ldots, N$,

$$\begin{bmatrix} M_{ij} & T_i \\ T_i^T & -I \end{bmatrix} < 0,$$

where M_{ij} is defined in Theorem 4.3.4, and $Q_j = P_j^{-1}$. If $\|\omega(t)\|_2^2 = \int_0^\infty \omega^T(t)\omega(t)dt \le 1$; then there exists the feedback $u(t) = F(\gamma^*)\,Q^{-1}(\gamma^*)\,x(t)$, such that level set L_c of the closed-loop system is attractive. □

Corollary 4.3.4 can also be regarded as a corollary of Theorem 4.3.4. It provides a result if the“≤” is replaced by “<” in Inequality (4.3.17).

At the end of this section, we give a conclusion for disturbances rejection.

Theorem 4.3.6 Consider System (4.3.16). If there exist matrices $P_j \in \mathbb{R}^{n\times n}$ and $F_j \in \mathbb{R}^{m\times n}$, with $P_j > 0, j = 1, 2, \ldots, J$, and $\delta > 0$, such that for all $i = 1, 2, \ldots, N$,

$$\begin{bmatrix} M_{ij} & T_i & Q_j C_i^T \\ T_i^T & -I & 0 \\ C_i Q_j & 0 & -\delta^2 I \end{bmatrix} \le 0, \tag{4.3.22}$$

with $M_{ij} = A_iQ_j + Q_jA_i^T + B_iF_j + F_j^TB_i^T - \sum_{k=1}^{J} \lambda_{ijk} \left(Q_k - Q_j\right)$, $Q_j = P_j^{-1}$, then there exists a feedback $u(t) = F(\gamma^*)\, Q^{-1}\, (\gamma^*)\, x(t)$ such that the output $y(t)$ satisfies $\|y(t)\|_2 \leq \delta\|\omega(t)\|_2$ when the initial condition $x_0 = 0$.

Proof It can be seen from the proof of Theorem 4.3.4. The key is to derive some inequalities from the known conditions. By Inequality (4.3.22), and using Lemma 4.1.1, we have

$$M_{ij} - \begin{bmatrix} T_i & Q_jC_i^T \end{bmatrix} \begin{bmatrix} -I & 0 \\ 0 & -\delta^2 I \end{bmatrix}^{-1} \begin{bmatrix} T_i^T \\ C_iQ_j \end{bmatrix} = M_{ij} + T_iT_i^T + \delta^2 Q_jC_i^TC_iQ_j \leq 0.$$

Substituting M_{ij} into the above inequality, and multiplying Q_j^{-1} from left and right sides for the inequality, it yields

$$Q_j^{-1}A + A^TQ_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}F_j^TB_i^TQ_j^{-1} - \sum_{k=1}^{J} \lambda_{ijk} Q_j^{-1} \left(Q_k - Q_j\right) Q_j^{-1}$$
$$+ Q_j^{-1}T_iT_i^TQ_j^{-1} + C_i^TC_i \leq 0\ .$$

$V_c(x)$ is defined by Definition 4.2.1, we can obtain

$$\dot{V}_c(x) \in \text{co}\left\{ \frac{\partial}{\partial x} V_c(x) \Big[A_ix + B_iF\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x + T_i\omega \right\}\ .$$

By the arguments given in Theorem 4.3.4, it can obtain that for $x \in E_j$,

$$\begin{aligned} &\frac{\partial}{\partial x} V_c(x)\left[A_ix + B_iF\left(\gamma^*\right) Q^{-1}\left(\gamma^*\right) x + T_i\omega\right]\Big|_{x=x_j} \\ &= x_j^T\left[Q_j^{-1}A_i + A_i^TQ_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}F_j^TB_j^TQ_j^{-1}\right] x_j + 2x_j^TQ_j^{-1}T_i\omega \\ &= x_j^T\left[Q_j^{-1}A_i + A_i^TQ_j^{-1} + Q_j^{-1}B_iF_jQ_j^{-1} + Q_j^{-1}F_j^TB_j^TQ_j^{-1} + \delta^{-2}C_j^TC_j\right] x_j \\ &\quad + 2x_j^TQ_j^{-1}T_i\omega - \delta^{-2}x_j^TC_j^TC_jx_j\ . \end{aligned}$$

Applying the following inequality

$$2x_j^TQ_j^{-1}T_i\omega \leq x_j^TQ_j^{-1}T_iT_i^TQ_j^{-1}x_j + \omega^T\omega,$$

we have

$$\dot{V}\left(x(t)\right) - \delta^{-2}y(t)^Ty(t) + \omega^T(t)\omega(t) \leq 0. \tag{4.3.23}$$

It is not difficult to extend Inequality (4.3.23) to the case that $x \in \text{bd } L_c$, and then extend the conclusion to all $x \in \mathbb{R}^n$. By integrating on both sides of (4.3.23), it yields

$$V(x(t)) - V(x(0)) \leq \delta^{-2} \int_0^t y^T(\tau)\, y(\tau)\, d\tau - \int_0^t \omega^T(\tau)\, \omega(\tau)\, d\tau.$$

Since $V(x(t)) \geq 0$ and $V(x(0)) = 0$, it yields

$$\int_0^t \omega^T(\tau)\, \omega(\tau)\, d\tau \leq \delta^{-2} \int_0^t y^T(\tau)\, y(\tau)\, d\tau.$$

□

The control problems, such as its attractiveness, stabilization and disturbances injection for linear polytope systems, are discussed in this section. The method for the proofs is quite similar. All start from the set E_j, then extend to L_c and to $\mathbb{R}^n$ at the end. Since $V_c(x)$ does not depend on the expression of a detailed system, all kinds of linear polytope systems can be discussed along this line, it is an example of developing linear polytope systems with time delay in this section. The method can also be used to deal with the problem of tracking (Cai et al. 2012). As an exercise for the reader, disturbances rejection for linear polytope systems with time delay can be discussed.

Problems

1. Prove the conclusion of Theorem 4.3.5 under $\omega^T(t)\omega(t) \leq 1$. That is, if there exist matrices $P_j \in \mathbb{R}^{n\times n}$ and $F_j \in \mathbb{R}^{m\times n}$, with $P_j > 0, j = 1, 2, \ldots, J$, such that for every $i = 1, 2, \ldots, N$,

$$\begin{bmatrix} M_{ij} + \beta Q_j & T_i \\ T_i^T & -\beta I \end{bmatrix} \leq 0$$

holds where $M_{ij} = A_i Q_j + Q_j A_i^T + B_i F_j + F_j^T B_i^T - \sum_{k=1}^{J} \lambda_{ijk} \left(Q_k - Q_j\right)$, $Q_j = P_j^{-1}$. If $\omega^T(t)\omega(t) \leq 1$, then L_c is attractive under the feedback control $u(t) = F(\gamma^*)\, Q^{-1}(\gamma^*)\, x(t)$, and try to extend the conclusion to $L_c(\rho)$.

2. If the other conditions is unchanged in Theorem 4.3.5, Inequality (4.3.20) is changed as follows

$$\begin{bmatrix} M_{ij} + \beta Q_j & T_i \\ T_i^T & -\beta I \end{bmatrix} \leq 0,$$

then L_c is invariable.

3. Prove: There exists the linear feedback control $u = Kx$, such that the closed-loop system is asymptotically stable if and only if there exist solutions X and Y, where X is positive definite, for the following linear matrix inequality

$$AX + XA^T + BY + Y^T B^T < 0.$$

4. Prove Corollary 4.3.3 and Corollary 4.3.4.

4.4 Saturated Control for Linear Control Systems

At the beginning of this chapter, we have mentioned that the saturated control for linear systems results in the study for linear differential inclusions. We deal with the problem with a viewpoint of differential inclusions in this section.

4.4.1 Saturated Control Described by Set-Valued Mappings

Consider the linear multivariable system described as follows

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{4.4.1}$$

where $x(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$ are the state and the input of linear system (4.3.21), respectively, and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$. Without loss of generality, we assume that B is full column rank, i.e., rank $B = m$. Denote $B = [b_1\ b_2 \dots b_m\ ]$, where b_i is the ith column of B.

A single variable function $\sigma(u)$ is called as the symmetrical saturation function, or shortened for saturation function, if it takes the form as follows

$$\sigma(u) = \begin{cases} \overline{u} & \overline{u} \le u, \\ u & -\overline{u} \le u < \overline{u}, \\ -\overline{u} & u < -\overline{u}, \end{cases} \tag{4.4.2}$$

where the constant $\bar{u}$ is called saturation level. In Eq. (4.4.2), without loss of generality, we have fixed the slope of the linear section to be 1, because if it is not equal to 1, the gain can be transferred to the input matrix B.

A single variable function $dz(u)$ is called dead zone function, if $dz(u) = u - \sigma(u)$. Dead zone function and saturation function are two typical nonlinear functions. Because the gain of an actual system is always limited, the existence of the saturation phenomenon for real systems is very common. Similarly, there is the dead zone phenomenon as long as there is friction. Hence, it is very necessary to consider saturation nonlinearity and dead zone nonlinearity for the control system design.

If $u = [u_1\ u_2 \ldots u_m]^T \in \mathbb{R}^m$, the vector saturation function is defined as follows

$$\sigma(u) = \begin{bmatrix} \sigma(u_1) \\ \sigma(u_2) \\ \vdots \\ \sigma(u_m) \end{bmatrix},$$

where $\sigma(u_i)$, $i = 1, 2, \ldots, m$ are single variable functions defined by Eq. (4.4.2). It is not necessary for us to require that $\overline{u}_i$, $i = 1, 2, \ldots, m$ are identical, but without loss of generality, we assume that their slopes are fixed to be 1. We also assume $\sigma(u) \in \mathbb{R}^m$, i.e., $\sigma : \mathbb{R}^m \to \mathbb{R}^m$. Similarly, we can define the vector value dead zone nonlinear function, and the details are omitted.

In the design of linear control systems, state feedback $u(t) = Fx(t)$ is widely used for System (4.4.1). We have mentioned saturation phenomenon is very common for actual systems; hence, the feedback $u(t) = Fx(t)$ should be replaced by saturation control $u(t) = \sigma(Fx(t))$. The closed-loop system then becomes

$$\dot{x} = Ax + B\sigma(Fx). \tag{4.4.3}$$

Equation (4.4.3) is usually called the saturated system. Similarly, if the dead zone nonlinearity is considered, i.e., $u(t) = Fx - \sigma(Fx(t))$, then the closed-loop system is

$$\dot{x} = (A + BF)x - B\sigma(Fx). \tag{4.4.4}$$

If we denote $\overline{A} = A + BF$ and $\overline{B} = -B$, then System (4.4.1) has a similar form of Eq. (4.4.3). Therefore, the study for system (4.4.1) generated by dead zone nonlinear feedback can be transformed into that for the saturated system (4.4.3).

Assume that $P \in \mathbb{R}^{n\times n}$ is a positive definite matrix, and $V(x) = x^T Px$ is quadratic form. By the Lyapunov theory, System (4.4.3) is asymptotically stable if

$$\dot{V}(x) = \frac{\partial V}{\partial x}(Ax + B\sigma(Fx)) = 2x^T P(Ax + B\sigma(Fx)) < 0. \tag{4.4.5}$$

By the definition of layer set, it is enough that Inequality (4.4.2) holds at the level set bd L_P.

It is obvious that the saturation function is nonlinear, and it is not convenient for further discussion. We naturally hope that it can be transformed into a linear operation, the consideration leads to the introduction of the differential inclusion.

Let Δ be a set of diagonal matrices defined as follows

$$\Delta = \left\{D_\eta = \operatorname{diag}\left(d_{\eta 1}\ \ d_{\eta 2}\ \ \ldots d_{\eta m}\right),\ \ d_{\eta j} \in \{0, 1\}\right\}.$$

Δ is a finite set and has 2^m elements. If $D_\eta \in \Delta$, denote $D_\eta^- = I - D_\eta$, where I is the $m \times m$ identity matrix, then $D_\eta^- \in \Delta$. For any two matrices $F, G \in \mathbb{R}^{m\times n}$, $D_\eta F + D_\eta^- G$ is also an $m \times n$ matrix and is composed by F and G, that is, if $d_{\eta j_0} = 1$, then the j_0th row of $D_\eta F + D_\eta^- G$ is the j_0th row of F, otherwise, it is the j_0th row of G.

The following lemma gives a linear upper bound of the saturation function. Let $H \in \mathbb{R}^{m\times n}$ and h_j^T be the jth row of H. Recall the notation of L_H defined in Sect. 4.2,

$$L_H = \{x;\ \|Hx\|_\infty \le 1\} = \left\{x;\ \max_i \left|h_i^T x\right| \le 1\right\},$$

where $H = \left[h_1^T, \dots, h_m^T\right]^T$. L_H is a simplex, and its vertices are the intersections of hyperplanes $h_i^T x = \pm 1$, $i = 1, 2, \dots, m$.

Lemma 4.4.1 If $F,\ H \in \mathbb{R}^{m\times n}$, and $x \in L_H$, then there exists $D_\eta \in \Delta$, such that

$$\sigma(Fx) \le D_\eta Fx + D_\eta^- Hx,$$

where if $x, y \in \mathbb{R}^n$, then $x \le y$ means $x_i \le y_i$ for $i = 1, 2, \dots, n$.

Proof If we can prove $\sigma\left(f_i^T x\right) \le \max\left(f_i^T x, h_i^T x\right)$, for every $i = 1, 2, \dots, m$, then we choose $d_{\eta i} = 1$ for the case of $f_i^T x \ge h_i^T x$, and $d_{\eta i} = 0$ otherwise. It is direct to show this $D_\eta = \operatorname{diag}\left(d_{\eta i}\right)$ meets the requirement of the lemma. We now prove $\sigma\left(f_i^T x\right) \le \max\left(f_i^T x, h_i^T x\right)$, for every $i = 1, 2, \dots, m$.

If $f_i^T x \ge -1$, then $\sigma\left(f_i^T x\right) \le f_i^T x$. It is sufficient for us to consider the case of $f_i^T x < -1$. Thus, $\sigma\left(f_i^T x\right) = -1$. Because $x \in L_H$, it yields $\left|h_i^T x\right| \le 1$. It implies $\sigma\left(f_i^T x\right) \le h_i^T x$. In conclusion, $\sigma\left(f_i^T x\right) \le \max\left(f_i^T x, h_i^T x\right)$. □

By Lemma 4.4.1, the selection of matrix $D_\eta \in \Delta$ depends on x it means D_η is a function of x; consequently, it is more suitable to denote D_η by $D_\eta(x)$.

4.4.2 Stabilization by the Saturated Control

Based on Lemma 4.4.1, the following lemma can be directly obtained.

Lemma 4.4.2 Consider the saturated system (4.4.3). If $P \in \mathbb{R}^{n\times n}$, $F,\ H \in \mathbb{R}^{m\times n}$, and $x \in L_H$, then it holds

$$x^T P b_i \sigma\left(f_i^T x\right) \le \max\left\{x^T P b_i f_i^T x,\ x^T P b_i h_i^T x\right\},$$

where $b_i,\ \ i = 1, 2, \dots, m$ is the ith column of B, respectively,f_i^T and h_i^T are the ith rows of F and H, respectively.

Proof The conclusion is verified by consideration of the following four cases.

(1) If $x^T Pb_i \geq 0$ and $f_i^T x \leq -1$, then $x^T Pb_i \sigma\left(f_i^T x\right) = -x^T Pb_i$. Because $x \in L_H$, $\left|h_i^T x\right| \leq 1$, that is, $-1 \leq h_i^T x$. It follows $x^T Pb_i \sigma\left(f_i^T x\right) = -x^T Pb_i \leq x^T Pb_i h_i^T x$.

(2) If $x^T Pb_i \geq 0$ and $f_i^T x > -1$, then $\sigma\left(f_i^T x\right) \leq f_i^T x$. Consequently, $x^T Pb_i \sigma\left(f_i^T x\right) \leq x^T Pb_i f_i^T x$.

(3) If $x^T Pb_i \leq 0$ and $f_i^T x \geq 1$, then $x^T Pb_i \sigma\left(f_i^T x\right) = x^T Pb_i$. Because $x \in L_H$, $\left|h_i^T x\right| \leq 1$, that is, $1 \geq h_i^T x$. It follows $x^T Pb_i \sigma\left(f_i^T x\right) = x^T Pb_i \leq x^T Pb_i h_i^T x$.

(4) If $x^T Pb_i \leq 0$ and $f_i^T x < 1$, then $\sigma\left(f_i^T x\right) \geq f_i^T x$. Consequently, $x^T Pb_i \sigma\left(f_i^T x\right) \leq x^T Pb_i f_i^T x$.

In conclusion,

$$x^T Pb_i \sigma\left(f_i^T x\right) \leq \max\left\{x^T Pb_i f_i^T x,\ \ x^T Pb_i h_i^T x\right\}. \qquad \square$$

In the proof of the lemma, it does not need the assumption that P is positive definite. In what follows, the positive definiteness of matrix P is used to estimate the domain of attraction, and the meaning of the layer $L_P(\rho)$ can be found in Sect. 4.2.

Theorem 4.4.1 Consider System (4.4.3). If $P \in \mathbb{R}^{n\times n},\ \ P > 0$, there exists an $H \in \mathbb{R}^{m\times n}$, such that:

(1) $L_P(\rho) \subset L_H$;

(2) $\left(A + B\left(D_\eta F + D_\eta^- H\right)\right)^T P + P\left(A + B\left(D_\eta F + D_\eta^- H\right)\right) < 0$, for every $D_\eta \in \Delta$.

Then $L_P(\rho)$ is an invariable set with attractiveness.

Proof Let $V(x) = x^T Px$. The derivate of $V(x)$ along with the trajectory of system (4.4.3) satisfies

$$\dot{V}(x) = 2x^T P\left(Ax + B\sigma(Fx)\right) = 2x^T PAx + 2x^T PB\sigma(Fx). \tag{4.4.6}$$

By Lemma 4.4.2, if $x \in L_H$ then

$$x^T PB\sigma(Fx) = \sum_{i=1}^{m} x^T Pb_i \sigma\left(f_i^T x\right) \leq \sum_{i=1}^{m} \max\left(x^T Pb_i f_i^T x, x^T Pb_i h_i^T x\right). \tag{4.4.7}$$

Define a matrix $D_\eta = \operatorname{diag}\left(d_{\eta i},\ \ i = 1, 2, \ldots, m\right)$, let $d_{\eta i} = 1$, if $x^T Pb_i f_i^T x \geq x^T Pb_i h_i^T x$, otherwise, $d_{\eta i} = 0$, note that the matrix D_η depends on x, i.e., $D_\eta = D_\eta(x)$. Thus, we have

$$\sum_{i=1}^{m} \max\left(x^T Pb_i f_i^T x, x^T Pb_i h_i^T x\right) = x^T PB\left(D_\eta(x)F + D_\eta^-(x)H\right)x. \tag{4.4.8}$$

Substituting Eq. (4.4.8) into Inequality (4.4.7), and then substituting (4.4.7) into (4.4.6), it yields

$$\dot{V}(x) = 2x^T PAx + 2x^T PB\sigma(Fx) \leq 2x^T PAx + 2x^T PB\left(D_\eta(x)F + D_\eta^-(x)H\right)x.$$

A sufficient condition for $\dot{V}(x) < 0$ is that the matrix

$$\left(A + B\left(D_\eta F + D_\eta^- H\right)\right)^T P + P\left(A + B\left(D_\eta F + D_\eta^- H\right)\right)$$

is negative definite for every $D_\eta \in \Delta$. By the condition of theorem, we have

$$\dot{V}(x) \le x^T\left[\left(A + B\left(D_\eta F + D_\eta^- H\right)\right)^T P + P\left(A + B\left(D_\eta F + D_\eta^- H\right)\right)\right]x < 0$$

for $x \in L_H$. Since for all bd$L_P(\rho) \subset L_H$, $\dot{V}(x) < 0$, $x \in$ db$L_P(\rho)$. Therefore, $L_P(\rho)$ is an invariable set with attractiveness. □

There are two remarks about Theorem 4.4.1.

Remark 1 It is obvious $0 \in L_P(\rho)$ is an invariable set of the closed-loop system (4.4.3) for every $\rho > 0$. If there is a $\rho > 0$ such that the origin is the unique invariant in $L_P(\rho)$, then by the LaSalle invariance principle, the conditions of Theorem 4.4.1 ensure the closed-loop system (4.3.23) is locally asymptotically stable. □

Remark 2 In the conditions of Theorem 4.4.1, Condition (2) is essential. Because the requirement that $\left(A + B\left(D_\eta F + D_\eta^- H\right)\right)^T P + P\left(A + B\left(D_\eta F + D_\eta^- H\right)\right)$ is negative definite is equivalent to $\left(A + B\left(D_\eta F + D_\eta^- H\right)\right)$ is a Hurwitz matrix for every $D_\eta \in \Delta$. If there exists an H to satisfy this requirement, we can find a ρ, such that $L_P(\rho) \subset L_H$. Thus, we obtain a basin of attraction. □

Corollary 4.4.1 If the matrices F and H given in Theorem 4.4.1 can stabilize the system (4.4.1), then $A + BF$ and $A + BH$ are both Hurwitz matrices.

Proof By Condition (2) of Theorem 4.4.1, inequality

$$\left(A + B\left(D_\eta F + D_\eta^- H\right)\right)^T P + P\left(A + B\left(D_\eta F + D_\eta^- H\right)\right) < 0$$

holds for $D_\eta \in \Delta$. So if we choose $D_\eta = I$, then $D_\eta^- = 0$, the above inequality is $(A + BF)^T P + P(A + BF) < 0$. Consequently $A + BF$ is a Hurwitz matrix. Similarly, it can be verified that $A + BH$ is also a Hurwitz matrix. □

Corollary 4.4.2 Consider System (4.4.3). If $P \in \mathbb{R}^{n\times n}$, $P > 0$, there exists a $K \in \mathbb{R}^{m\times n}$, such that

(1) $L_P(\rho) \subset L_{KF}$;

(2) $\left(A + B\left(D_\eta F + D_\eta^- KF\right)\right)^T P + P\left(A + B\left(D_\eta F + D_\eta^- KF\right)\right) < 0$ for $D_\eta \in \Delta$.

Then $L_P(\rho)$ is an invariable set with attractiveness.

In general, for System (4.4.3), a feedback matrix F is designed firstly for stabilization. Corollary 4.4.2 shows that H can be determined by K to complete the saturated feedback.

The following theorem extends conclusions to the stabilization problem based on the convex hull Lyapunov function.

Theorem 4.4.2 If there exist positive definite matrices $P_j \in \mathbb{R}^{n\times n}$ and matrices $F_j, H_j \in \mathbb{R}^{m\times n}$, $j = 1, 2, \ldots, J$, and constant $\lambda_{\eta jk} \geq 0$, $\eta \in \{1, 2, 3, \ldots, 2^m\}$, $j, k \in \{1, 2, \ldots, J\}$, such that

(1) $L_{P_j} \subset L_{H_j}$;
(2) For every $\eta = 1,\ 2, \ldots,\ 2^m$ and $D_\eta \in \Delta$,

$$Q_jA^T+AQ_j+\left(D_\eta F_j+D_\eta^- H_jQ_j\right)^T B^T+B\left(D_\eta F_j+D_\eta^- H_jQ_j\right) < \sum_{k=1}^{J} \lambda_{\eta jk}\left(Q_k-Q_j\right),$$

with $Q_j = P_j^{-1}$.

Then $Q\left(\gamma^*\right)$ is the matrix constructed in Sect. 4.2, i.e., $Q\left(\gamma^*\right) = \sum\limits_{j=1}^{J} \gamma_j^* Q_j$, and, in addition, define $F\left(\gamma^*\right) = \sum\limits_{j=1}^{J} \gamma_j^* F_j$, then the saturated control

$$u = \sigma\left(F\left(\gamma^*\right) Q{\left(\gamma^*\right)}^{-1} x\right) \tag{4.4.9}$$

can make closed-loop System (4.4.3) is strongly asymptotically stable.

Proof The proof is similar to that of Theorem 4.4.1. The key is to construct $H\left(\gamma^*\right)$ such that $L_{P(\gamma^*)} \subset L_{H(\gamma^*)}$. By Lemma 4.2.1, the condition (1) given by this theorem is

$$\begin{bmatrix} 1 & z_{ij}^T \\ z_{ij} & Q_j \end{bmatrix} \geq 0\ , \quad i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, J, \tag{4.4.10}$$

where z_{ij}^T is the ith row of matrix H_jQ_j. Since $\gamma_j^* \geq 0$, $\sum\limits_{j}^{J} \gamma_j^* = 1$, by Inequality (4.4.10), it can be deduced that,

$$\begin{bmatrix} 1 & \sum\limits_{j=1}^{J} \gamma_j^* z_{ij}^T \\ \sum\limits_{j=1}^{J} \gamma_j^* z_{ij} & Q\left(\gamma^*\right) \end{bmatrix} \geq 0\ , \quad i = 1, 2, \ldots, m, \tag{4.4.11}$$

where $\sum_{j=1}^{J} \gamma_j^* z_{ij}^T$ is the ith row of $\sum_{j=1}^{J} \gamma_j^* H_j Q_j$. Denote $H\left(\gamma^*\right) = \sum_{k=1}^{J} \gamma_k^* H_k Q_k Q\left(\gamma^*\right)^{-1}$, then we have $H\left(\gamma^*\right) \in \mathbb{R}^{m\times n}$, and

$$\sum_{j=1}^{J} \gamma_j^* H_j Q_j = H\left(\gamma^*\right) Q\left(\gamma^*\right).$$

It leads to that $\sum_{j=1}^{J} \gamma_j^* z_{ij}^T$ is the ith row of $H(\gamma^*)Q(\gamma^*)$. So, by Lemma 4.2.1, Inequality (4.4.11) implies

$$L_{P(\gamma^*)} \subset L_{H(\gamma^*)}. \tag{4.4.12}$$

Multiplying with Q_j^{-1} for the inequality given by Condition (2) from left and right sides, respectively, it yields

$$\begin{aligned}
&\left[A + B\left(D_\eta F_j Q_j^{-1} + D_\eta^- H_j\right)\right]^T Q_j^{-1} + Q_j^{-1}\left[A + B\left(D_\eta F_j Q_j^{-1} + D_\eta^- H_j\right)\right] \\
&\quad < \sum_{k=1}^{J} \lambda_{\eta jk} Q_j^{-1}\left(Q_k - Q_j\right) Q_j^{-1},
\end{aligned} \tag{4.4.13}$$

for every $D_\eta \in \Delta$.

The set E_j is defined at the second section of this chapter. First, if $x = x_j \in E_j = \text{bd}\, L_c \cap \text{bd}\, L_{P_i}$, then we have

$$\begin{aligned}
\dot{x} = \dot{x}_j &= Ax_j + B\sigma\left(D_\eta F\left(\gamma^*\right) Q\left(\gamma^*\right)^{-1}\right) x_j \\
&\le Ax_j + B\left(D_\eta F\left(\gamma^*\right) Q\left(\gamma^*\right)^{-1} + D_\eta^- H\left(\gamma^*\right)\right) x_j \\
&= Ax_j + B\left(D_\eta F_j Q_j^{-1} + D_\eta^- H_j\right) x_j.
\end{aligned}$$

Note that the $D_\eta \in \Delta$ depends on $x = x_j$. By Inequality (4.4.13), it holds $\dot{V}(x) = \dot{V}\left(x_j\right) < 0$.

If $x \in \text{bd}\, L_c$, by the method described in the previous section, we assume that $\gamma_j^*(x) > 0$, for $j = 1, 2, \ldots, J_0$; and $\gamma_j^*(x) = 0$, for $j = J_0 + 1, \ldots, J$. By Theorem 4.2.3, it can be deduced $x = \sum_{j=1}^{J_0} \gamma_j^* x_j$, with $x_j \in E_j$, and $Q(\gamma^*)^{-1} x = Q_j^{-1} x_j$, $j = 1, 2, \ldots, J_0$.

In the proof of Theorem 4.3.1, it has been deduced $F(\gamma^*)\,Q(\gamma^*)^{-1}x = \sum_{j=1}^{J_0} \gamma_j^* F_j Q_j^{-1} x_j$. In addition,

$$H(\gamma^*)\,x = \left(\sum_{j=1}^{J} \gamma_j^* H_j Q_j Q(\gamma^*)^{-1} \right) x = \sum_{j=1}^{J_0} \gamma_j^* H_j Q_j Q_j^{-1} x_j = \sum_{j=1}^{J_0} \gamma_j^* H_j x_j.$$

Using the arguments did in Theorem 4.3.1, and by the above equality, we get

$$\begin{aligned} \dot{V}(x) &= \frac{\partial V}{\partial x}\Big(Ax + B\left(D_\eta F(\gamma^*)\,Q^{-1}(\gamma^*) + D_\eta^- H(\gamma^*)\right)\Big)x \\ &= \sum_{j=1}^{J_0} 2\gamma_j^* x_j^T Q_j \left(Ax_j + B\left(D_\eta F_j Q_j^{-1} + D_\eta^- H_j\right)\right) x_j, \end{aligned}$$

for $x \in \text{bd } L_c$. By omitting the detailed proof, we have $\dot{V}(x) < 0$, for $x \in \text{bd } L_c$. That is, under the saturated control $\sigma\left(F(\gamma^*)\,Q(\gamma^*)^{-1}x\right)$, the closed system (4.4.3) is asymptotically stable. □

As Corollary 4.4.1, we can assert that $A + BF_jQ_j^{-1}$ and $A + BH_j$ are all Hurwitz matrices, that is, $F_jQ_j^{-1}$ and H_j can all stabilize (A, B).

From the proofs given in this section, readers can understand the general steps to deal with the saturated feedback for a linear system by using set-valued mapping. If readers compare the method applied in this section with that in Sect. 4.3, they may find that we have added a constraint condition $L_P(\rho) \subset L_H$. The condition makes us to apply Lemma 4.4.1, we then get an upper boundary for the saturated mapping. If readers are interest in the design of saturated feedback by using the method of differential inclusions, they can try to deal with the polytope system $\dot{x} \in \text{co}\{A_i x + B_i \sigma(u); i = 1, 2, \ldots, N\}$ as an exercise.

4.4.3 Disturbance Rejection by the Saturated Control

In this section, we consider disturbance rejection under the saturated control for the linear system based on the method developed in previous section. Consider a linear system suffered from disturbance

$$\begin{aligned} \dot{x} &= Ax + Bu + T\omega, \\ y &= Cx, \end{aligned} \tag{4.4.14}$$

where $y \in \mathbb{R}^r$ is the output of the system, $C \in \mathbb{R}^{r\times n}$ is the output matrix, $\omega \in \mathbb{R}^q$ is the disturbance of the system, and $T \in \mathbb{R}^{n\times q}$ is the disturbance gain. Assume the disturbance ω is bound. At first, let us consider the saturation control of linear feedback $u = Fx$, the closed-loop system is

$$\begin{aligned} \dot{x} &= Ax + B\sigma(Fx) + T\omega, \\ y &= Cx. \end{aligned} \tag{4.4.15}$$

Theorem 4.4.3 Consider System (4.4.15). Assume the disturbance ω satisfies that $\omega^T(t)\omega(t) \leq 1$ for $t \in [0, \ \infty)$. If there exist $P \in \mathbb{R}^{n\times n}, \ \ P > 0$ and $F, \ H \in \mathbb{R}^{m\times n}$, such that

(1) $L_P(\rho) \subset L_H$ for some $\rho > 0$;
(2) There is a $\delta > 0$ such that for every $D_\eta \in \Delta$, the following inequality holds

$$\begin{bmatrix} \left(A + B\left(D_\eta F + D_\eta^- H\right)\right)^T P + P\left(A + B\left(D_\eta F + D_\eta^- H\right)\right) + \frac{\delta}{\rho}Q & T \\ T^T & -\delta I \end{bmatrix} < 0.$$

Then $L_P(\rho)$ is an invariable set with attraction.

Proof Let $x \in L_P(\rho)$. Since $L_P(\rho) \subset L_H$, by Lemma 4.4.2, there exists a $D_\eta \in \Delta$ such that $PB\sigma(Fx) \leq PB\left(D_\eta F + D_\eta^- H\right)x$.

Let $V(x) = x^T Px$, the derivative of $V(x)$ along with the trajectory of System (4.4.15) is

$$\begin{aligned} \dot{V}(x) &= \frac{\partial V}{\partial x}(Ax + B\sigma(Fx) + T\omega) \\ &= 2x^T P(Ax + B\sigma(Fx) + T\omega) \\ &\leq 2x^T PAx + 2x^T PB\left(D_\eta F + D_\eta^- H\right)x + 2x^T PT\omega\ . \end{aligned}$$

By Inequality (4.1.5), we have

$$2x^T PT\omega \leq \frac{1}{\delta}x^T PTT^T Px + \delta\omega^T\omega \leq \frac{1}{\delta}x^T PTT^T Px + \delta,$$

where we have used $\omega^T\omega \leq 1$.

Using Lemma 4.1.1, the condition (2) of the theorem implies

$$AQ + QA^T + B\left(D_\eta F + D_\eta^- H\right)Q + Q\left(D_\eta F + D_\eta^- H\right)^T B^T + \frac{\delta}{\rho}Q + \frac{1}{\delta}TT^T < 0.$$

Multiplying the above inequality by $Q^{-1} = P$ from left and right sides, respectively, it yields

$$PA + A^T P + PB\left(D_\eta F + D_\eta^- H\right) + \left(D_\eta F + D_\eta^- H\right)^T B^T P + \frac{\delta}{\rho} P + \frac{1}{\delta} PTT^T P < 0. \tag{4.4.16}$$

Thus, we have

$$\begin{aligned}\dot{V}(x) &\leq 2x^T PAx + 2x^T PB\left(D_\eta F + D_\eta^- H\right)x + \frac{1}{\delta} x^T PTT^T Px + \delta \\ &< -\frac{\delta}{\rho} x^T Px + \delta\ .\end{aligned}$$

$x \in L_P(\rho)$ implies $x^T Px \leq \rho$, hence $\dot{V}(x) < 0$, that is, $L_P(\rho)$ is an invariable set with attraction. □

Theorem 4.4.3 can be extended to the convex hull quadratic form, we only list the conclusion as follows, the proof is left to readers.

Corollary 4.4.3 Consider System (4.4.15). Assume there exist $P_j \in \mathbb{R}^{n\times n},\ P_j > 0$, $j = 1, 2, \ldots, J$, and $F_j,\ H_j \in \mathbb{R}^{m\times n}$, such that

(1) $L_{P_j} \subset L_{H_j},\ \ j = 1, 2, \ldots, J$;
(2) There exist δ >0 and $\lambda_{\eta jk} > 0$ where $\eta \in \{1, 2, 3, \ldots, 2^m\}$ and $j, k \in \{1, 2, \ldots, J\}$ such that for every $D_\eta \in \Delta$, for $j = 1, 2, \ldots, J$, the following inequality holds

$$\begin{bmatrix} M_{\eta j} + \delta Q_j & T \\ T^T & -\delta I \end{bmatrix} < 0,$$

where $M_\eta = Q_j A^T + AQ_j + \left(D_\eta F_j + D_\eta^- H_j Q_j\right)^T B^T + B\left(D_\eta F_j + D_\eta^- H_j Q_j\right) - \sum_{k=1}^{J} \lambda_{\eta jk}\left(Q_k - Q_j\right)$. Then L_c is an invariable set with attraction. □

A set $L \subset \mathbb{R}^n$ is said to be locally ultimate attractive, if there is a δ >0 for every initial condition $\|x_0\| \leq \delta$ the trajectory of System (4.4.15), $x(t)$, will satisfy $x(t) \in L$ for all $t > T$, where the T may depend on δ. A set $L \subset \mathbb{R}^n$ is said to be ultimate attractive if δ can be large arbitrarily.

Theorem 4.4.4 Consider system (4.4.15). Assume the disturbance satisfies $\|\omega(t)\| \leq 1$. If there exist $P \in \mathbb{R}^{n\times n},\ \ P > 0$, and $F,\ H \in \mathbb{R}^{m\times n}$, such that

(1) $L_P \subset L_H$;
(2) There exists a $\delta > 0$, such that for every $D_\eta \in \Delta$, the following inequality holds

$$\begin{bmatrix} \left(A+B\left(D_\eta F+D_\eta^- H\right)\right)^T P+P\left(A+B\left(D_\eta F+D_\eta^- H\right)\right)+\delta Q & T \\ T^T & -\delta I \end{bmatrix} \leq 0,$$

where $Q=P^{-1}$. Then L_c is ultimate attractive.

Proof Using the same argument as in Theorem 4.4.2, let $V(x)=x^T P x$, then the derivative of $V(x)$ along with the trajectory of System (4.4.15) satisfies

$$\dot{V}(x) \leq 2x^T PAx+2x^T PB\left(D_\eta F+D_\eta^- H\right)x+2x^T PT\omega .$$

By Inequality (4.1.5), it yields

$$2x^T PT\omega \leq x^T PTT^T Px+\omega^T\omega .$$

Using the condition (2) and Lemma 4.1.1, we have

$$\dot{V}\left(x(t)\right) \leq -\delta V\left(x(t)\right)+\omega^T(t)\omega(t).$$

Since

$$V\left(x(t)\right) \leq e^{-\delta t}V\left(x(0)\right)+\int_0^t e^{-\delta(t-\tau)}\omega^T\left(\tau\right)\omega\left(\tau\right)d\tau \leq e^{-\delta t}V\left(x(0)\right)+1,$$

we get $\overline{\lim_t} V\left(x(t)\right) \leq 1$, that is, L_c is ultimate attractive. □

Theorem 4.4.3 can also be extended to the convex hull quadratic function. We list it as a corollary.

Corollary 4.4.4 Consider System (4.4.15). Assume the disturbance satisfies $\|\omega(t)\| \leq 1$. If there exist $P_j \in \mathbb{R}^{n\times n},\ \ P_j>0$, and $F_j,\ H_j \in \mathbb{R}^{m\times n}, j=1,2,\ldots,J$, such that

(1) $L_{P_j} \subset L_{H_j}, j=1,2,\ldots,J$;
(2) There exist $\delta>0$ and $\lambda_{\eta jk}>0$ for $\eta \in \{1,2,3,\ldots,2^m\}$ and $j,k \in \{1,2,\ldots,J\}$ such that for every $D_\eta \in \Delta$, and $j=1,2,\ldots,J$, the following inequality holds

$$\begin{bmatrix} M_{\eta j}+\delta Q_j & T \\ T^T & -I \end{bmatrix} < 0,$$

where $M_{\eta j}=Q_j A^T+AQ_j+\left(D_\eta F_j+D_\eta^- H_j Q_j\right)^T B^T+B\left(D_\eta F_j+D_\eta^- H_j Q_j\right)-\sum_{k=1}^J \lambda_{\eta jk}\left(Q_k-Q_j\right)$. Then L_c is ultimate attractive. □

The final conclusion is about disturbance rejection. Using the same argument as in Theorem 4.3.6, we can prove Theorem 4.4.5. Corollary 4.4.5 can be proved using the same argument as in Corollary 4.4.3. The proofs are omitted.

Theorem 4.4.5 Consider the system (4.4.15). If there exist $P \in \mathbb{R}^{n \times n}$, $P > 0$, and $F, H \in \mathbb{R}^{m \times n}$, such that

(1) $L_P(\rho) \subset L_H$;
(2) There exists a $\delta > 0$, such that for every $D_\eta \in \Delta$, the following inequality holds

$$\begin{bmatrix} M_\eta & T & QC \\ T^T & -I & 0 \\ CQ & 0 & -\delta^2 I \end{bmatrix} \leq 0,$$

where $M_\eta = AQ + QA^T + B\left(D_\eta F + D_\eta^- H\right) + \left(D_\eta F + D_\eta^- H\right)^T B^T$, then the output $y(t)$ satisfies $\|y(t)\| \leq \delta \|\omega(t)\|$ provided that the initial value $x(0) = 0$. □

Corollary 4.4.5 Consider the system (4.4.15). If there exist $P_j \in \mathbb{R}^{n \times n}$, $P_j > 0$, and $F_j, H_j \in \mathbb{R}^{m \times n}, j = 1, 2, \ldots, J$, such that

(1) $L_{P_j} \subset L_{H_j}, j = 1, 2, \ldots, J$;
(2) There exist $\delta > 0$ and $\lambda_{\eta jk} > 0$ where $\eta \in \{1, 2, 3, \ldots, 2^m\}$ and $j, k \in \{1, 2, \ldots, J\}$ such that for every $D_\eta \in \Delta$ and $j = 1, 2, \ldots, J$, the following inequality holds

$$\begin{bmatrix} M_{\eta j} & T & Q_j C \\ T^T & -I & 0 \\ CQ_j & 0 & -\delta^2 I \end{bmatrix} \leq 0,$$

where $M_{\eta j} = AQ_j + Q_j A^T + B\left(D_\eta F_j + D_\eta^- H_j Q_j\right) + \left(D_\eta F_j + D_\eta^- H_j Q_j\right)^T B^T$, then the output $y(t)$ satisfies $\|y(t)\| \leq \delta \|\omega(t)\|$ provided that the initial value $x(0) = 0$. □

Problems

1. Consider the dead zone nonlinearity

$$dz(u) = \begin{cases} u - \overline{u}, & u \geq \overline{u}, \\ 0, & -\overline{u} \leq u < \overline{u}, \\ u + \overline{u}, & u < -\overline{u} \circ \end{cases}$$

where $\overline{u} > 0$ is called the dead zone boundary. Prove:

(1) $dz(u) \in \text{co}\{0, u - v\}$, where $v \in [-\overline{u}, \overline{u}]$;

(2) If we restrict us to consider the part $dz(u) \geq 0$ only, then $dz(u) \in \text{co}\,\{0, u\}$, that is, $dz(u) = \delta u$ where δ depends on argument u, i.e., $\delta = \delta(u)$. It means $dz(u) = \delta u$ is a nonlinear function.

(3) Furthermore, if $u \in \mathbb{R}^n$, define

$$dz(u) = \begin{bmatrix} dz\,(u_1) \\ dz\,(u_2) \\ \vdots \\ dz\,(u_m) \end{bmatrix}.$$

If we consider only the part of $dz\,(u_i) \geq 0,\ \ i = 1, 2, \ldots, m$, then $dz(y) \in \text{co}\left\{D_\eta u, D_\eta \in \Delta\right\}$, where Δ and D_η are those defined at the beginning of this section.

2. Consider the control system $\dot{x} = Ax + Bu$, and the dead zone nonlinear control $u = dz(Kx)$. Using the method given in Problem 1, transform this control problem into a control problem for the linear differential inclusion, and give the corresponding conclusion of Lemma 4.4.1.
3. Under the conditions of Theorem 4.4.1, prove that $x = 0$ is the unique invariable set in $L_P(\rho)$.
4. If the condition (2) of Theorem 4.4.1 holds, then the problem of finding the largest ρ such that $L_P\,(\rho) \subset L_H$ is called as the problem of largest invariant set in L_H. Applying Schur complement Theorem, give a matrix inequality method for finding the largest invariant set in L_H.
5. Assume $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $b = \begin{bmatrix} 0 \\ 5 \end{bmatrix}$, $F = \begin{bmatrix} -2 & -1 \end{bmatrix}$; $P = \begin{bmatrix} 4 & 1 \\ 1 & 0.5 \end{bmatrix}$, $H = [0.4 \ \ -1]$.

   Try to find the largest ρ such that $L_P(\rho)$ is ultimate attractive. (Note: in this case $\sigma(Fx) = \sigma\,(-2x_1 - x_2)$)
6. Assume $A = \begin{bmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{bmatrix}$, $b = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$, $T = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}$; if $P = \begin{bmatrix} 0.0752 & -0.0566 \\ -0.0566 & 0.1331 \end{bmatrix}$, $F = [-0.1125 \quad -0.2987]$. Check the conclusion of Theorem 4.4.3 to hold.
7. Prove Corollary 4.4.3 and Corollary 4.4.4.
8. Prove Theorem 4.4.5 and Corollary 4.4.5.
9. Assume $A = \begin{bmatrix} 0 & -0.5 \\ 1 & 1.5 \end{bmatrix}$, $b = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$, using, respectively, $F_1 = [0.9471 \quad 1.6000]$ and $F_2 = [-0.1600 \quad 1.6000]$ as the gain of saturated control, show that the closed-loop system is stable. There are

$$P_1 = \begin{bmatrix} 1.6245 & -1.5364 \\ -1.5364 & 15.3639 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 5.9393 & -0.2561 \\ -0.2561 & 2.5601 \end{bmatrix}$$

respectively. Try to find $Q\,(\gamma^*)$ and $F\,(\gamma^*)$.

References

Cai X, Huang J, Xie Q (2012) Output tracking and disturbance rejection of linear differential inclusion systems [J]. Int J Syst Sci 43(13):2072–2078

Filippov AF (1988) Differential equations with discontinuous right-hand sides [M]. Kluwer Academic Publishers, Dordrecht

Hu T, Lin Z (2004) Properties of the composite quadratic Lyapunov functions [J]. IEEE Trans AC 49(7):1162–1167

Hu T, Teel A, Zaccarian L (2006) Stability and performance for saturated systems via quadratic and nonquadratic Lyapunov functions [J]. IEEE Trans AC 51(11):1770–1780

Chapter 5
Luré Differential Inclusion Systems

In the former chapter, control problems of linear polytope systems are considered by using the convex hull Lyapunov function. The linear polytope system is a convex combination of a finite set of finite linear systems. Involved set-valued mapping in the differential inclusion is the convex combination; hence, convex theory can be applied to deal with these control problems. In this chapter, the Luré differential inclusion system and its relative control problems are considered. This kind of differential inclusions is different from linear convex hulls; the set-valued mapping satisfies so-called sector condition, i.e., the image of the set-valued mapping is in a cone. Hence, it is a naturally nonlinear mapping.

5.1 Luré Systems

As an introduction, we present some basic materials for Luré systems and the positive realness of functions. The system has been widely researched in 1950s, and the positive realness was a very useful tool in 1960s in the research of the design of adaptive control of linear systems.

5.1.1 Luré Systems and Absolute Stability

The Luré system is a kind of nonlinear system. Its block diagram is shown in Fig. 5.1. From Fig. 5.1, it is clear that the system has a feedback structure. Its forward path is a linear system *G* and its feedback is a nonlinear system *N*. In the classical theory of nonlinear control systems, many systems are assumed to hold

Z. Han et al., *Theory of Control Systems Described by Differential Inclusions*,
Springer Tracts in Mechanical Engineering, DOI 10.1007/978-3-662-49245-1_5

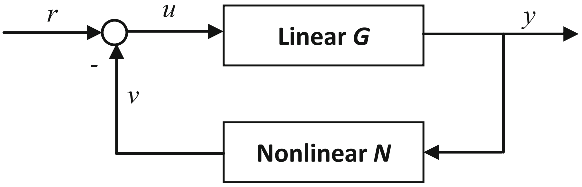


Fig. 5.1 A class of nonlinear control systems

such a configuration. Many research studies have been done to deal with the methods of transforming equivalently a nonlinear system into this configuration.

In Fig. 5.1, the input of system is the reference signal r and the output of the system is y which is also the input of nonlinear part. The input of linear part is the control signal u which is an algebraic sum of r and v, the reference and the output of nonlinear part, respectively. Without loss of generality, we assume the feedback is negative. We also assume the dimension of v is equal to that of y.

The linear part G is assumed to be a dynamic system described as follows

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t). \end{cases} \tag{5.1.1}$$

In the system of Eq. (5.1.1), all symbols have the same meanings as those in Eq. (4.1.1) except $Du(t)$. $Du(t)$ is called the direct transportation term, and D is the gain of the direct transportation. The transfer function of System (5.1.1) is

$$G(s) = C(sI - A)^{-1}B + D. \tag{5.1.2}$$

$G(s)$ is a matrix of rational functions. If $D = 0$, every element in matrix $G(s)$ is a strictly proper rational function, i.e., the degree of the numerator is less than that of denominator. If $D \neq 0$, then there are certainly some elements in rational matrix $G(s)$ whose degrees of numerators are equal to those of denominators. Such a rational matrix is called proper.

The nonlinear part N is a static process described by

$$v(t) = v\,(t, y, \tau) \tag{5.1.3}$$

where $v(t, y, \tau)$ is a nonlinear function which may be time-varying. t stands for the time, and $\tau > 0$ is a constant to indicate the delayed time. In the next section, $v(t, y, \tau)$ is a set-valued mapping. $v(t, y, \tau)$ is always supposed to be continuous for t and y.

The following inequality

$$\int_{t_0}^{t} v^T\,(\tau)\,y\,(\tau)d\tau \geq -r_0^2 \tag{5.1.4}$$

is valid for any time t, where t_0 is the initial time and $r_0 \neq 0$ is a constant, and it may depend on the initial state x_0 and t_0. In general, Inequality (5.1.4) is called Popov integral inequality. If $v(t)$ and $y(t)$ satisfy Inequality (5.1.4), then we say they satisfy the Popov condition, or the nonlinear part N satisfies the Popov condition.

It is not required that the integration in the left side of Inequality (5.1.4) is convergent as $t \to \infty$. If the integration converges or tends to infinity, then Inequality (5.1.4) always holds.

Definition 5.1.1 Consider the system described in Fig. 5.1. Let $r(t) = 0$, $v(t)$ and $y(t)$ satisfy Popov integral Inequality (5.1.4). The system is said to be absolutely stable, if there exists a constant K such that

$$\|x\left(t, x_0, t_0\right)\| \leq K\left(\|x_0\| + r_0\right), \tag{5.1.5}$$

for any initial time t_0 and initial state x_0, where $x(t, x_0, t_0)$ is the state of System (5.1.1). Furthermore, the system is said to be absolutely asymptotically stable if $\lim_{t\to\infty} x\left(t, x_0, t_0\right) = 0$. □

The absolute stability is defined for the whole system. In the terminology *absolute stability*, the word *stability* is used to describe the property of the state $x(t)$, which is a solution of differential equation, and the word *absolute* is introduced for the nonlinear part N. It means that we only require the function $v(t, y, \tau)$ to satisfy Inequality (5.1.4), but we do not care the specific form of $v(t, y, \tau)$.

The SISO system with the structure described by Fig. 5.1 was firstly considered by Popov, but in stand of Inequality (5.1.4), he considered the following sector condition:

$$0 \leq k_1 y^2 \leq y v\left(t, y, \tau\right) \leq k_2 y^2, \quad 0 \leq k_1 < k_2. \tag{5.1.6}$$

Inequality (5.1.6) means that, in the v-y space, the image of $v = v\left(t, y, \tau\right)$ lies in a sector of the first and/or the third quadrants. Inequality (5.1.6) obviously implies Inequality (5.1.4). At that time, Popov investigated the stability of the system. In order to distinguish the conventional definition of stability, he called it absolute stability because of being regardless the specific form of $v(t, y, \tau)$. Later on, some researchers called the system is hyperstable if the nonlinear part N satisfies Inequality (5.1.4) and absolute stability for Inequality (5.1.6). But this book does not distinguish the two concepts, and call them all absolute stability.

Let $u(t) = r(t) - v(t)$. Then the closed-loop system combined by linear System (5.1.1) with nonlinear mapping Eq. (5.1.3) can be described as

$$\begin{cases} \dot{x}(t) = Ax(t) - Bv(t) + Br(t), \\ v(t) = v\left(t, y, \tau\right), \\ y(t) = Cx(t) - Dv(t) + Dr(t), \end{cases} \tag{5.1.7}$$

where $v(t)$ and $y(t)$ satisfy Popov integral Inequality (5.1.6). The system described by Eq. (5.1.7) is called Luré system.

If $v(t, y, \tau)$ is a set-valued mapping, then System (5.1.7) becomes

$$\begin{cases} \dot{x}(t) = Ax(t) - B\omega(t) + Br(t), \\ \omega(t) \in v\left(t, y, \tau\right), \\ y(t) = Cx(t) - D\omega(t) + Dr(t). \end{cases} \tag{5.1.8}$$

(5.1.8) is called the Luré differential inclusion. A detailed definition will be given in the next section.

Theorem 5.1.1 Let the linear System (5.1.1) be both controllable and observable and $r(t) \equiv 0$.

1. If System (5.1.7) is absolutely stable, then (a) all eigenvalues of A stay at the left open complex plane, (b) every eigenvalue on the complex axis is simple;
2. If System (5.1.7) is absolutely asymptotically stable, then A is a Hurwitz matrix.

Proof If $r(t) \equiv 0$, then Inequality (5.1.4) is equivalent to

$$\int_{t_0}^{t} u^{\mathrm{T}}\left(\tau\right) y\left(\tau\right) d\tau \le r_0^2.$$

Let $u(t) \equiv 0$, and $x(0) = x_0 \neq 0$, the above inequality always holds. And the state response is $x(t) = e^{At}x_0$. If the eigenvalue of A is $\sigma + \mathrm{j}\omega$, where $\sigma > 0$, then there always exists x_0 such that the state $x(t) = e^{At}x_0$, which contains the term $ce^{(\sigma+\mathrm{j}\omega)t}$ with a nonzero vector c. Thus, $\|x(t)\| \to \infty$, the system is not absolutely stable. If there exist multiple poles $\mathrm{j}\omega$, then there exists x_0 such that $x(t) = e^{At}x_0$ contains the term $c(t)e^{\mathrm{j}\omega t}$, where $c(t)$ is a nonzero degree polynomial with respect to t. Thus, $\|x(t)\| \to \infty$. The similar proof can also be obtained for the absolutely asymptotical stability. Here we just omit it. □

Theorem 5.1.1 illustrates the necessary condition for the absolute (absolutely asymptotical) stability of System (5.1.7). It is the same as that condition of System (5.1.1).

5.1.2 Positive Realness and the Positive Realness Lemma

In the research of absolute stability, positive realness plays a very important role. This section will introduce the positive realness briefly; more details can be found in the textbooks for adaptive control system or nonlinear control system (Rochafellar 1970; Smirnov 2002).

Consider a function $f : \mathbb{C} \to \mathbb{C}$, where $\mathbb{C}$ is the set of complex numbers. f is called a real function, if $f(\mathbb{R}) \subset \mathbb{R}$, i.e., f maps real number to real number. Let $s = \sigma + \mathrm{j}\omega \in \mathbb{C}$ be a complex number, where $\mathrm{j}^2 = -1$ is the unit of imaginary number, then $f(s)$ can be written as $f(s) = \mathrm{Re}\, f(s) + \mathrm{j}\mathrm{Im} f(s)$, where $\mathrm{Re} f(s)$ and $\mathrm{Im} f(s)$ are the real part and imaginary parts of $f(s)$, respectively, and they are all real numbers for any $s \in \mathbb{C}$. For a real function f, it holds that $f(\bar{s}) = \mathrm{Re}\, f(s) - \mathrm{j}\mathrm{Im} f(s)$, where $\bar{s}$ is the conjugate complex number of s, i.e., $\bar{s} = \overline{\sigma + \mathrm{j}\omega} = \sigma - \mathrm{j}\omega$.

Definition 5.1.2 A function $f : \mathbb{C} \to \mathbb{C}$ is a positive real function, if (1) f is a real function; (2) $\mathrm{Re}\, f(s) \geq 0$ if $\sigma > 0$.

$f : \mathbb{C} \to \mathbb{C}$ is a strictly positive real function, if (1) f is a real function and (2) there exists a $\lambda > 0$ such that $\mathrm{Re}\, f(s) \geq 0$ for $\sigma \geq -\lambda$. □

Roughly speaking, a positive real function maps the real axis to the real axis and also maps the right complex plane to the right complex plane.

According to the above definition, if $f_1(s)$ and $f_2(s)$ are both positive real functions, then $f_1(s) + f_2(s)$ and $f_1(f_2(s))$ are also positive real functions; $a\, f_1(s)$ is positive real function if a is a positive constant; $[f_1(s)]^{-1}$ is also a positive real function whenever $[f_1(s)]^{-1}$ exists. However, $f_1(s) \cdot f_2(s)$ may not be a positive real function. The following theorems present the properties of positive real functions and strictly positive real functions. They can be proved by the theory of complex analysis and are omitted here.

Theorem 5.1.2 A real function $f(s)$ is positive real if and only if the following conditions hold simultaneously:

1. If $s = \sigma + \mathrm{j}\omega$ is a pole of $f(s)$, then $\sigma \leq 0$.
2. If $s = \mathrm{j}\omega$ is a pole of $f(s)$, then it is simple and its residue is positive.
3. If $s = \mathrm{j}\omega$ is not the pole of $f(s)$, then $\mathrm{Re}\, f(\mathrm{j}\omega) > 0$. □

Theorem 5.1.3 A real function $f(s)$ is strictly positive real if and only if the following conditions hold simultaneously:

1. If $s = \sigma + \mathrm{j}\omega$ is a pole of $f(s)$, then $\sigma < 0$.
2. For every $s = \mathrm{j}\omega$, $\mathrm{Re}\, f(\mathrm{j}\omega) > 0$. □

The first condition of Theorem 5.1.2 means that there is no pole in the right open half complex plane, i.e., $f(s)$ is analytic in the right open half complex plane. If real part of $f(s)$ is positive, then $[f(s)]^{-1}$ has also positive real part; thus, the zeros of $f(s)$ do not lie in the right open half complex plane. The fact means that the zeros and poles of a positive real function do not lie in the right open half complex plane. Similarly, the zeros and poles of a strictly positive real function do not lie in the closed right half complex plane.

If the above discussion is applied in the real rational function $f(s) = n(s)/d(s)$, then more properties of $f(s)$ can be obtained.

Theorem 5.1.4 If a real rational function $f(s) = n(s)/d(s)$ is positive real, then $|\deg n(s) - \deg d(s)| \leq 1$.

Proof Assume that $m = \deg n(s) \geq \deg d(s) = n$. We have

$$f(s) = \frac{n(s)}{d(s)} = p_0 s^{m-n} + p_1 s^{m-n-1} + \cdots + \frac{\overline{n}(s)}{d(s)},$$

then j∞ is an $m - n$ multiple pole of $f(s)$. By (2) of Theorem 5.1.2, a pole in imaginary axis is simple; thus $1 \geq m - n \geq 0$. If $m \leq n$, we will consider $[f(s)]^{-1}$. By using a similar discussion, we can obtain that $0 \leq n - m \leq 1$. Thus, the theorem is proved. □

Theorem 5.1.5 If a real rational function $f(s) = n(s)/d(s)$ is strictly positive real, then $n(s)$, $d(s)$ and $n(s) + d(s)$ are all Hurwitz polynomials.

Proof Applying Condition (1) of Theorem 5.1.3 to $f(s)$, $[f(s)]^{-1}$ and $1 + f(s)$, these conclusions can be proved directly. □

Theorems 5.1.4 and 5.1.5 provide simple necessary conditions for the positive real rational functions. They are very useful.

Now the definition of positive realness will be extended to the matrix case.

Let $F(s) : \mathbb{C} \to \mathbb{C}^{m\times m}$ be an $m \times m$ complex matrix. $F(s)$ is a real matrix if $F(\mathbb{R}) \subset \mathbb{R}^{m\times m}$.

Definition 5.1.3 Let $F(s)$ be a real matrix. If

$$F(s) + F^*(s) \geq 0, \tag{5.1.9}$$

for a complex number s which holds $\operatorname{Re} s > 0$, then $F(s)$ is a positive real matrix, where $F^*(s)$ is the conjugate and transpose matrix of $F(s)$.

If there exists a $\lambda > 0$ such that for any complex number s with $\operatorname{Re} s > -\lambda$,

$$F(s) + F^*(s) \geq 0,$$

then $F(s)$ is called strictly positive real matrix. □

In the matrix theory, if $H(s) = H^*(s)$ for any complex s, then the $H(s)$ is called as Hermite matrix. The concept of Hermite matrix is an extended version of symmetry matrix which is defined in real field. The main feature is that the eigenvalues of a Hermite matrix are all real. Since $H^*(s) = H^{\mathrm{T}}(\overline{s})$, $F(s) + F^*(s)$ can also be written as $F(s) + F^{\mathrm{T}}(\overline{s})$. For simplicity, it is, sometimes, denoted by $\operatorname{He} F(s)$.

The following theorems are reversions of Theorems 5.1.2 and 5.1.3 in matrix case, respectively; the proofs are omitted.

Theorem 5.1.6 A real matrix function $F(s)$ is positive real if and only if the following conditions hold simultaneously:

(1) $F(s)$ is analytic at the left open half plane.

(2) If $s = \mathrm{j}\omega$ is a pole of some element of $F(s)$ then $\operatorname*{res}_{s=\mathrm{j}\omega} F(s) + \operatorname*{res}_{s=\mathrm{j}\omega} F^*(s) \geq 0$.

(3) If $s = \mathrm{j}\omega$ is not a pole of $F(s)$, then $F(\mathrm{j}\omega) + F^*(\mathrm{j}\omega) \geq 0$. □

Theorem 5.1.7 A real matrix function $F(s)$ is strictly positive real if and only if the following conditions hold simultaneously:

(1) $F(s)$ is analytic at the left closed half plane.
(2) For any $s = \mathrm{j}\omega$, $F(\mathrm{j}\omega) + F^*(\mathrm{j}\omega) > 0$. □

Theorems 5.1.3 and 5.1.4 can both be extended to rational function matrices. In order not to derive from the main topic, we will leave them to the readers in the exercises, and these conclusions are needed in the latter of this chapter.

Now we return back to the linear control system Eq. (5.1.1). The System (5.1.1) is positive real or strictly positive real if its transfer function $G(s)$ is positive real or strictly positive real. In this book, we always suppose that System (5.1.1) is both controllable and observable, and the degree of the denominator of its transfer function is n.

Theorem 5.1.8 Suppose System (5.1.1) is both controllable and observable, then it is positive real if and only if there exist positive definite matrix P and matrices K and L with compatible dimensions, such that

$$PA + A^T P = -LL^T, \tag{5.1.10}$$

$$K^T L^T + B^T P = C, \tag{5.1.11}$$

$$K^T K = D + D^T. \tag{5.1.12}$$

Theorem 5.1.8 was proved by B.D.O Anderson in 1967, and it is also called the positive real lemma (Anderson 1967). If $D = 0$, then $K = 0$ and Eq. (5.1.11) disappears, and Eq. (5.1.11) becomes $B^T P = C$. This is common form in this chapter. The following corollary gives the necessary condition under which (5.5.1) is positive real.

Corollary 5.1.1 Suppose System (5.1.1) is both controllable and observable and $D = 0$, then CB is positive definite if the system is positive real.

Proof Since $B^T P = C$, we have $B^T PB = CB$. Thus, CB is symmetry and semipositive definite. It needs only to prove that the rank of CB is of full rank. If $\operatorname{rank}(CB) < m$, then there exists a $u \neq 0$ such that $CBu = 0$, and $u^T B^T PBu = 0$. Consider the fact that B has the full column rank, hence $Bu \neq 0$, and $u^T B^T PBu > 0$, it contracts to $u^T B^T PBu = 0$. □

It is noted that Corollary 5.1.1 is essential for the feedback positive realness.

If Eq. (5.1.10) is rewritten as $PA + A^T P = -LL^T - Q$ for a positive definite matrix Q. The equation and Eqs. (5.1.11) and (5.1.12) give the necessary and sufficient conditions for the strictly positive realness condition.

An alternative form of Theorem 5.1.8 is given in the exercises; these conclusions are useful for the proof of Theorem 5.1.8.

5.1.3 Criterion for Absolute Stability

In this subsection, we prove that the absolute stability is equivalent to positive realness for the system described by Fig. 5.1. The following result of Laplace transformation is needed in the proof.

Let $f(t)$, $t \in [0, \infty)$, be a real function. If there exists a $\sigma > 0$ such that $\int_0^\infty f(t)e^{-st}dt$ converges for every s with $\mathrm{Re}\, s \geq \sigma$. Then integration is called the Laplace transformation of $f(t)$ and denoted by $F(s)$. Usually, t in $f(t)$ is called time domain variable and s in $F(s)$ is the frequency domain variable, and $F(s)$ is the image function and $f(t)$ is the original function. If $F(s)$ is known, $f(t)$ can be computed as follows

$$f(t) = \frac{1}{2\pi \mathrm{j}} \int_{-\infty}^{\infty} F(\sigma + \mathrm{j}\omega)\, e^{(\sigma + \mathrm{j}\omega)t} d\omega,$$

at each continuous point of $f(t)$. The equation is called the inversion formula of Laplace transformation. If $F(s)$ exists when $s \to \infty$, by residual theory, we can obtain another equation

$$f(t) = \sum_k \mathrm{Re\ s}\, F(p_k)\, e^{p_k t}, \tag{5.1.13}$$

where p_k is the pole of $F(s)$.

Theorem 5.1.9 Consider the system described by Fig. 5.1. If the nonlinear part satisfies Popov integral Inequality (5.1.4), then the closed-loop system is absolutely stable if and only if the linear part Eq. (5.1.1) is positive real, i.e., the transfer function matrix Eq. (5.1.2) is positive real.

Proof Sufficiency. Let P be the positive definite matrix satisfying Theorem 5.1.8. Let $V(x) = x^T Px$ be the Lyapunov candidate function. Considering the derivative of $V(x)$ along with the trajectory of the closed-loop System (5.1.7) with $r(t) \equiv 0$, we can get that

$$\begin{aligned}
\dot{V}(x) &= (Ax - B\nu)^T Px + x^T P (Ax - B\nu) \\
&= x^T \left(A^T P + PA\right) x - \nu^T B^T Px - x^T PB\nu \\
&= -x^T L^T Lx - 2\nu^T \left(C - K^T L^T\right) x \\
&= -x^T L^T Lx + 2\nu^T K^T L^T x - \nu^T K^T K\nu - 2\nu^T Cx + \nu^T K^T K\nu \\
&= -\|Lx - K\nu\|^2 - \nu^T Cx - x^T C^T \nu + \nu^T D\nu + \nu^T D^T \nu \\
&= -\|Lx - K\nu\|^2 - 2\nu^T y \\
&\leq -2\nu^T y.
\end{aligned}$$

Integrating both sides results in

$$\int_0^t \dot{V} d\tau = x^T(t)Px(t) - x^T(0)Px(0) \leq -2\int_0^t v^T(\tau)\, y(\tau)\, d\tau \leq 2r_0^2.$$

Let λ_M and λ_m be the maximum and minimum eigenvalues of P. It follows from the above inequality that

$$\lambda_m \|x(t)\|^2 \leq \lambda_M \|x(0)\|^2 + 2r_0^2 \leq K\left(\|x(0)\|^2 + r_0^2\right) \leq K(\|x(0)\| + r_0)^2,$$

where $K = \max\{\lambda_M, 2\}$. It means that Inequality (5.1.5) holds.

Necessity. It is proved by using Definition 5.1.2. By Theorem 5.1.2, $G(s)$ is analytic at the right open half plane.

If there exists a complex number $\sigma_0 + \mathrm{j}\omega_0$ with $\sigma_0 > 0$ such that $G(\sigma_0 + \mathrm{j}\omega_0) + G^*(\sigma_0 + \mathrm{j}\omega_0)$ is not semipositive definite, i.e., there exists $u_0 \in \mathbb{C}^m$ with $\|u_0\| = 1$ such that

$$u_0^* \left[G(\sigma_0 + \mathrm{j}\omega_0) + G^*(\sigma_0 + \mathrm{j}\omega_0)\right] u_0 < 0.$$

Choose an input $u(t) = \operatorname{Re} u_0 e^{(\sigma_0 + \mathrm{j}\omega_0)t}$. Since the system described by $G(s)$ is asymptotically stable, by Eq. (5.1.13), we have

$$y(t) = \operatorname{Re} G(\sigma_0 + j\omega_0)\, u_0 e^{(\sigma_0 + j\omega_0)t}. \tag{5.1.14}$$

In view of $u^{\mathrm{T}}(t)y(t) = \operatorname{Re} u_0^{\mathrm{T}} e^{(\sigma_0 + \mathrm{j}\omega_0)t} \operatorname{Re} G(\sigma_0 + j\omega_0)\, u_0 e^{(\sigma_0 + \mathrm{j}\omega_0)t}$. It can be proved that $\int_0^t u^{\mathrm{T}}(\tau)\, y(\tau) d\tau \rightarrow \infty$; hence, Popov integral Inequality (5.1.4) holds. However, $Cx(t) = y(t) - Du(t) = \operatorname{Re} C(\sigma_0 + \mathrm{j}\omega_0 - A)^{-1} B u_0 e^{(\sigma_0 + \mathrm{j}\omega_0)t}$, i.e., $\sup \|x(t)\| \rightarrow \infty$. This contradicts with the absolute stability. □

By a similar discussion to Theorem 5.1.9, we can have the following theorem; the detailed proof is omitted.

Theorem 5.1.10 Consider the system described by Fig. 5.1. If the nonlinear part satisfies Popov integral Inequality (5.1.4), then the closed-loop system is absolutely asymptotically stable if and only if the linear part Eq. (5.1.1) is strictly positive real, or the transfer function matrix (5.1.2) is strictly positive real. □

Problems

1. Let System (5.1.1) be both controllable and observable. Then it is positive real if and only if there exist positive definite matrix P, semipositive definite matrices Q and R and matrix S such that

$$
\begin{aligned}
&PA + A^{\mathrm{T}}P = -Q,\\
&B^{\mathrm{T}}P + S^{\mathrm{T}} = C,\\
&D + D^{\mathrm{T}} = R,\\
&\text{with} \quad \begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix} > 0.
\end{aligned}
$$

2. Let System (5.1.1) be both controllable and observable. Then it is strictly positive real if and only if there exist positive definite matrix P, Q, R and matrix S such that

$$
\begin{aligned}
&PA + A^{\mathrm{T}}P = -Q,\\
&B^{\mathrm{T}}P + S^{\mathrm{T}} = C,\\
&D + D^{\mathrm{T}} = R,\\
&\text{with} \quad \begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix} > 0.
\end{aligned}
$$

3. Prove Theorem 3.1.9.
4. Let $W(s)$ be an $n \times n$ rational function matrix. $W(s) = N(s)D^{-1}(s)$ is a right coprime factorization and rank $[D(s)\ \ N(s)] = n$ for any $s \in \mathbb{C}$. Then the following conclusions hold:

   (1) $W(s)$ is positive real if and only if $D^*(s)N(s)+N^*(s)D(s)$ is positive definite.
   (2) If rank $W(s) = n$ for s with $\operatorname{Re} s > 0$, then $W(s)$ is positive real if and only if $W^{-1}(s)$ is positive real.
   (3) If $W(s)$ is positive real, then $\det N(s)$ and $\det D(s)$ are Hurwitz polynomials.
   (4) If $W(s)$ is positive real, then $N(s)[D(s)+\alpha N(s)]^{-1}$ and $[N(s)+\alpha D(s)]\,D^{-1}(s)$ are both positive real, where $\alpha > 0$ is a real number.
   (5) If $D(s)$ is column proper, then $|\alpha_i - \beta_i| \leq 1$, where α_i and β_i are the ith column degree of $N(s)$ and $D(s)$ (Let $\partial_i D(s)$ be column degree of the ith column of $D(s)$, $D(s)$ is column proper if $\deg(\det D(s)) = \sum \partial_i D(s)$).

   The above conclusion is also valid when left coprime factorization is considered.
5. Prove Theorem 5.1.10.

5.2 Stabilization of Luré Differential Inclusion Systems

The Luré system was introduced in the former section; in that section, we have concluded the system is stable if and only if its linear part is positive real. This section tries to extend this conclusion to Luré differential inclusion systems.

5.2.1 An Example of the Luré Differential Inclusion System

Before we deal with the stabilization of Luré differential inclusion, we would like to present an example which was introduced by J.C.A. de Bruin and his colleagues.

An object may move if it is subjected to external force. The object has to suffer from friction force whether or not it moves. We have mentioned in Sect. 2.3 that the friction is a quite complicated phenomenon. The friction coefficient which determines the value of friction force has a nonlinear relationship respected to the velocity of object. We have pointed out at that section if the static friction is considered, then it is suitable to be described by set-valued mapping.

J.C.A. de Bruin and colleagues have designed an experimental device to illustrate the necessity of introduction of Luré differential inclusion (de Bruim et al. 2009). The experimental device is given in Fig. 5.2. The input of the device is the voltage which makes the upper DC motor rotate. The DC motor drives the upper disk through gear box, and the upper disk is connected to lower disk via a flexible cable.

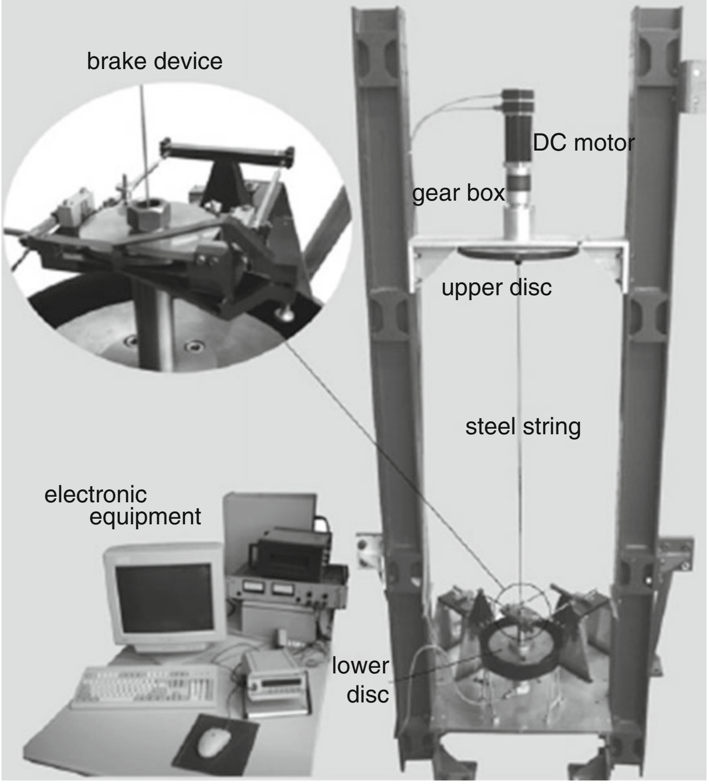


Fig. 5.2 The experimental device of flexible rotation (de Bruim et al. 2009)

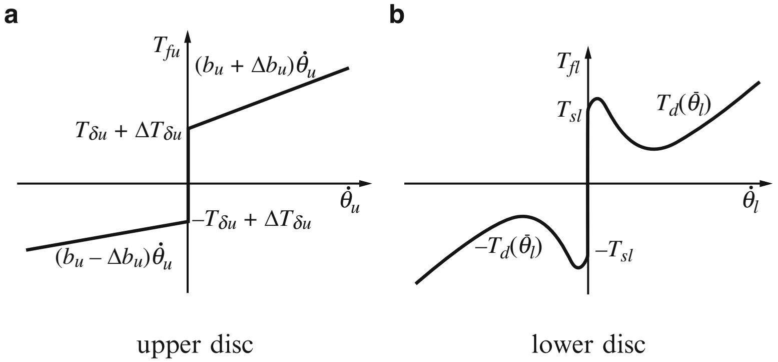


Fig. 5.3 Dead zone nonlinearity relationship between torque and angular velocity. (**a**) upper disk. (**b**) lower disk

An oil seal is set up at the tail of the flexible cable to make all kinds of friction more significant, while the angular velocity of the lower disk is recorded by the computer as output of the device.

It is obviousthat the velocity is feasible to be measured, while the torque is difficult to be measured. Figure 5.3 presents the relations of angular velocities and torques where the angular velocity is argument, then the torque is the dependent variable. Figure 5.3a gives the image of the upper disk and Fig. 5.3b the lower disk. These pictures show that they are set-valued mappings. By moment balance, it is easy to obtain that

$$\begin{aligned} &J_u\ddot{\theta}_u + k_\theta\left(\theta_u - \theta_l\right) + b\left(\dot{\theta}_u - \dot{\theta}_l\right) + T_{fu}\left(\dot{\theta}_u\right) - k_m u = 0, \\ &J_l\ddot{\theta}_l - k_\theta\left(\theta_u - \theta_l\right) - b\left(\dot{\theta}_u - \dot{\theta}_l\right) + T_{fl}\left(\dot{\theta}_l\right) = 0, \end{aligned} \tag{5.2.1}$$

where variable subscripted with u stands for the variable of upper disk, and variable subscripted with l stands for the lower disk, J is the moment of inertia, $k_\theta\left(\theta_u - \theta_l\right) + b\left(\dot{\theta}_u - \dot{\theta}_l\right)$ is twisting torque, and θ is angular displacement. Resistance torque T_f can be computed, respectively, by the following set-valued mappings

$$T_{fu}\left(\dot{\theta}_u\right) \in \begin{cases} T_{cu}\left(\dot{\theta}_u\right)\operatorname{sgn}\left(\dot{\theta}_u\right), & \dot{\theta}_u \neq 0, \\ \left[-T_{su} + \Delta T_{su}, T_{su} + \Delta T_{su}\right], & \dot{\theta}_u = 0, \end{cases}$$

$$T_{fl}\left(\dot{\theta}_l\right) \in \begin{cases} T_{cl}\left(\dot{\theta}_l\right)\operatorname{sgn}\left(\dot{\theta}_l\right), & \dot{\theta}_l \neq 0, \\ \left[-T_{sl}, T_{sl}\right], & \dot{\theta}_l = 0, \end{cases}$$

where $T_{cu}\left(\dot{\theta}_u\right)$ and $T_{cl}\left(\dot{\theta}_u\right)$ is the nonlinear terms. Let $x = \left[\theta_u - \theta_l \; \dot{\theta}_u \; \dot{\theta}_l\right]^T$, Eq. (5.2.1) can be written as

$$\begin{cases} \dot{x}(t) = Ax(t) + G\omega(t) + Bu(t), \\ z(t) = Hx(t), \\ y(t) = Cx(t), \\ \omega(t) \in -\varphi(z). \end{cases}$$

where $z = \left[\dot{\theta}_u \; \dot{\theta}_l\right]^T$, $\omega = \left[T_{fu}(z_1) \; T_{fl}(z_2)\right]^T$. The above expression is a Luré differential inclusion system. From the experimental results, it is suitable to describe the device by the differential inclusion model.

5.2.2 Stabilization of Luré Differential Inclusion Systems

Let us consider the Luré differential inclusion System (5.1.8). Assume that $D = 0$ firstly, how to extend the conclusion to the case of $D \neq 0$ will be pointed out. In order to be clear, the block diagram is shown in Fig. 5.4. In the discussion of this chapter, the block diagram is a very useful tool.

The Luré differential inclusion system is described as follows:

$$\begin{cases} \dot{x}(t) = Ax(t) - B\omega(t) + Br(t), \\ \omega(t) \in v\left(t, y, \tau\right), \\ y(t) = Cx(t), \end{cases} \tag{5.2.2}$$

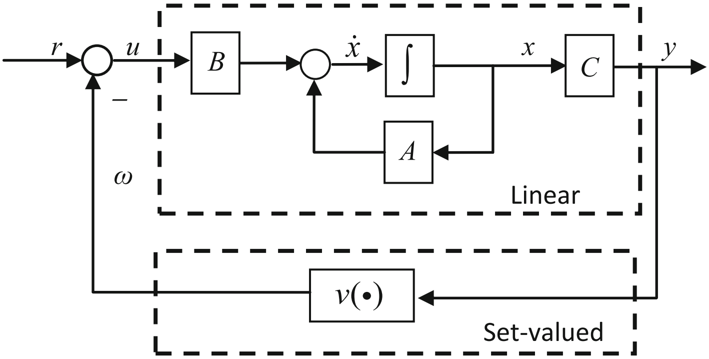


Fig. 5.4 The Luré differential inclusion system

where r, u, x, y and ω are the reference, control input, state, output of the system and output of the set-valued mapping, respectively. The dimensions of r, u, y and ω are all r, and the dimension of x is n. t is the time variable and τ is the delayed time, respectively. The set-valued function $\nu(t, y, \tau)$ is supposed to be monotone with respect to y. It means according to the definition given in Sect. 2.6, for any $y_i \in \mathbb{R}^m, \ i = 1, 2$, and $\omega_i \in \nu\,(t, y_i, \tau)$, $\langle(\omega_1 - \omega_2)\,, (y_1 - y_2)\rangle \geq 0$. In order to guarantee that the equilibrium of the system is the original, it is always assumed that $0 \in \nu\,(t, 0, \tau)$. This assumption is not loss of the generality because we can use coordinate transformation to realize it. By the monotonicity of the set-valued mapping, we then obtain that $\langle\omega, y\rangle \geq 0$, i.e.,

$$y^{\mathrm{T}}\nu\,(t, y, \tau) \geq 0. \tag{5.2.3}$$

It is obvious that Inequality (5.2.3) implies Popov integral Inequality (5.1.4). By Theorems 5.1.8 and 5.1.9, we can get the following conclusion. Detailed proof is omitted.

Theorem 5.2.1 Consider the Luré differential inclusion system Eq. (5.2.2). Suppose the system satisfies the following conditions:

1. $\nu(y)$ is monotone and $0 \in \nu\,(t, 0, \tau)$.
2. (C, A, B) is both controllable and observable.

If the transfer function of linear part $G(s) = C(sI - A)^{-1}B$ is positive real, then the system is stable; if $G(s) = C(sI - A)^{-1}B$ is strictly positive real, then the system is asymptotically stable. □

Theorem 5.2.1 still holds when $D \neq 0$.

5.2.3 Zeroes and Relative Degree of Control Systems

This subsection introduces two fundamental concepts used in linear control systems which play very important roles in Sect. 5.3.

5.2.3.1 The Zeroes of System

By Theorem 5.2.1, the stability of the Luré differential inclusion System (5.2.1) is equivalent to the positive realness or strictly positive realness of linear part (C, A, B). Thus, the stabilization design can be transformed to make the linear part positive real or strictly positive real which is called feedback positive realness in this book. This subsection mainly focuses on the output feedback, i.e., $r = Ky$. By the block diagram, the control of linear part is $u = Ky - \omega$.

The problem of feedback positive realness is to find feedback gain K such that $(C, A + BKC, B)$ is positive real or strictly positive real. In order to discuss feedback positive realness, the zeroes of a system are introduced. The zeroes of a system include finite zeroes and infinite zeroes. The concept of finite zeroes is presented firstly.

Let $W(s)$ be an $r \times m$ rational function matrix, and assume the rank of $W(s)$ is $\min(r, m)$.[1] Let $s_0 \in \mathbb{C}$ be a complex number. Then $W(s_0)$ be a complex matrix. If $\operatorname{rank} W(s_0) < \min(r, m)$, then s_0 is said to be a finite zero of $W(s)$, and finite zero is also simplified as zero. If (C, A, B) is a realization of $W(s)$, then s_0 is called a finite zero of the system (C, A, B). The properties of finite zeroes are as follows (Chen 1984).

1. Let $W(s) = N_r(s)D_r^{-1}(s) = D_l^{-1}(s)N_l(s)$ be a right coprime factorization and a left coprime factorization of $W(s)$, respectively, then s_0 is a finite zero of $W(s)$ if and only if $\operatorname{rank} N_r(s_0) = \operatorname{rank} N_l(s_0) < \min(r, m)$.
2. If (C, A, B) is a minimal realization of $W(s)$, then s_0 is a finite zero of $W(s)$ if and only if
$$\operatorname{rank}\begin{bmatrix} s_0 I - A & B \\ -C & 0 \end{bmatrix} < n + \min(r, m).$$
3. If s_0 is a finite zero of (C, A, B), then s_0 is also the finite zero of $\left(CT^{-1}, T(A + BF + GC)T^{-1}, TB\right)$, where T, F and G are matrices with compatible dimensions, and T is invertible. The conclusion illustrates that the finite zeroes are invariant under the coordinate transformation, state feedback and output injection.

The multiplicity of a finite zero can be also defined. Let $W(s)$ be an $r \times m$ rational function matrix, and denote $q = \min(r, m)$. The right coprime factorization of $W(s)$ is $W(s) = N_r(s)D_r^{-1}(s)$. If s_0 is a finite zero of $W(s)$, then $s - s_0$ is the common factor of all qth-order determinants of $N_r(s)$; if $(s - s_0)^p$ is the common factor of all qth-order determinants of $N_r(s)$ but $(s - s_0)^{p+1}$ is not, then s_0 is a p-multiplicity finite zero of $W(s)$.

In the control theory, linear system (C, A, B) is called a minimal phase system, if the zeroes of the system are all in the left half complex plane, or, the zeroes are all stable.

5.2.3.2 Relative Degree of System

Infinite zero is relevant to relative degree of system. Let us start with the single-variable system.

Consider a rational function $W(s) = n(s)/d(s)$. If $\deg n(s) < \deg d(s)$, then $n(s)/d(s) \to 0$, when $s \to \infty$. The fact means that infinite is also the zero of the rational function. In order to distinguish from finite zero, this kind of zero is called infinite zero. The multiplicity of infinite zero can also be defined. If $\deg d(s) - \deg n(s) = p > 0$, then the infinite zero is called p-multiplicity of the rational function. Infinite zero is very explicit in the theory of automatic control.

[1] The rank of rational matrix is defined on the field of rational functions. Denote that $q = \operatorname{rank} W(s)$, then there exists at least one $q \times q$ sub-matrix of $W(s)$ whose determinant is not equal to 0.

The multiplicity of infinite zero is called relative degree in the theory of linear systems. The specific meaning is as follows. Let us consider single-variable system (c^{T}, A, b), and the state space is described as:

$$\dot{x} = Ax + bu,$$
$$y = c^{\mathrm{T}}x.$$

without loss of generality, we assume that (c^{T}, A, b) takes the Brunovsky controllable canonical form:

$$c^{\mathrm{T}} = \left[\beta_n \ \beta_{n-1} \cdots \beta_1\right], \quad A = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 1 \\ -\alpha_n & -\alpha_{n-1} & \cdots & -\alpha_2 & -\alpha_1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{5.2.4}$$

Its transfer function is

$$G(s) = \frac{\beta_1 s^{n-1} + \cdots + \beta_{n-1}s + \beta_n}{s^n + \alpha_1 s^{n-1} + \cdots + \alpha_{n-1}s + \alpha_n}.$$

When we assume that the (c^{T}, A, b) is both controllable and observable, the numerator and the denominator are relatively prime. If the relative degree of $G(s)$ is p, then the degree of the numerator is $n - p$, and $\beta_1 = \cdots = \beta_{p-1} = 0$ but $\beta_p \neq 0$. It is easy to verify that $c^{\mathrm{T}}b = 0, \ c^{\mathrm{T}}Ab = 0, \ldots c^{\mathrm{T}}A^{p-2}b = 0, \ c^{\mathrm{T}}A^{p-1}b \neq 0$. On the other word, taking the derivative of the output y, $\dot{y} = c^{\mathrm{T}}\dot{x} = c^{\mathrm{T}}(Ax + bu) = c^{\mathrm{T}}Ax$, $\ddot{y} = c^{\mathrm{T}}A\dot{x} = c^{\mathrm{T}}A(Ax + bu) = c^{\mathrm{T}}A^2x$. We then conclude that in the expression of $y^{(p-1)}$ the control input u does not appear; and in the p-th derivative of y, $y^{(p)}$, we have $y^{(p)} = c^{\mathrm{T}}A^p x + c^{\mathrm{T}}A^{p-1}bu$, u appears for the first time.

It is easy to conclude that the relative degree is equal to p if and only if $\lim\limits_{s\to\infty} s^{p-1}G(s) = 0$; $\lim\limits_{s\to\infty} s^p G(s) \neq 0$. Although the controllable canonical Eq. (5.2.4) is used to explain that u appears firstly in the p-derivative of output y, i.e., $y^{(p)}$, it does not rely on the specific form of the system matrix.

We now extend the definition of relative degree to the rational matrix. Let $W(s)$ be an $r \times m$ rational function matrix, $W(s)$ can be written as

$$W(s) = \begin{bmatrix} w_1^{\mathrm{T}}(s) \\ w_2^{\mathrm{T}}(s) \\ \vdots \\ w_r^{\mathrm{T}}(s) \end{bmatrix}, \quad \text{for} \quad i = 1, 2, \ldots, r,$$

where $w_i^{\mathrm{T}}(s)$ is an m-dimension row vector, and its elements are rational fractions.

If $\lim\limits_{s\to\infty} s^j w_i^{\mathrm{T}}(s) = 0, j = 0, 1, 2, \ldots, p_i - 1$; $\lim\limits_{s\to\infty} s^{p_i} w_i^{\mathrm{T}}(s) \neq 0$, and

$$\operatorname{rank}\begin{bmatrix} \lim\limits_{s\to\infty} s^{p_1} w_1^{\mathrm{T}}(s) \\ \lim\limits_{s\to\infty} s^{p_2} w_2^{\mathrm{T}}(s) \\ \vdots \\ \lim\limits_{s\to\infty} s^{p_r} w_r^{\mathrm{T}}(s) \end{bmatrix} = \min(r, m),$$

then the ith relative degree of $W(s)$ is p_i, the set $\{p_1, p_2, \ldots, p_r\}$ is called the set of relative degrees of $W(s)$, and $p = \min\{p_1, p_2, \ldots, p_r\}$ is the relative degree of $W(s)$.

Let us consider the multivariable system (C, A, B), the state space is described by

$$\dot{x} = Ax + Bu,$$
$$y = Cx.$$

Assume that $m = r$. Denote the ith row of C by c_i^{T}, then the transfer function matrix is

$$G(s) = \begin{bmatrix} c_1^{\mathrm{T}}(sI - A)^{-1}B \\ c_2^{\mathrm{T}}(sI - A)^{-1}B \\ \vdots \\ c_r^{\mathrm{T}}(sI - A)^{-1}B \end{bmatrix}.$$

If $\{p_1, p_2, \ldots, p_r\}$ is the relative degree set of $G(s)$, then u does not appears in $\dot{y}_i, \ldots, y_i^{(p_i-1)}$ but in $y_i^{(p_i)}$, and we denote that

$$\Pi = \begin{bmatrix} c_1^{\mathrm{T}} A^{p_1} B \\ c_2^{\mathrm{T}} A^{p_2} B \\ \vdots \\ c_r^{\mathrm{T}} A^{p_r} B \end{bmatrix},$$

then Π is invertible. Note that the invertibility of Π is additional requirement for the definition of the relative degree in multivariable systems.

The relative degree is defined by transfer function. Thus, it is also invariant under coordinate transformation and state feedback except for the sequence.

Suppose the input and output are all m-dimension. If $\det(CB) \neq 0$, then the elements in the set of relative degrees are all 1. By coordinate transformation, the system can be changed into

$$\begin{bmatrix} A_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & 0 \\ C_1 & C_2 & 0 \end{bmatrix},$$

where B_1 is an $m \times m$ invertible matrix. $\det(CB) \neq 0$ implies that C_1 is also invertible. If we make another state transformation with the transformation matrix $T = \begin{bmatrix} I & C_1^{-1}C_2 \\ 0 & I \end{bmatrix}$, i.e., $T^{-1} = \begin{bmatrix} I & -C_1^{-1}C_2 \\ 0 & I \end{bmatrix}$, the system then takes the form of

$$\begin{bmatrix} \overline{A}_{11} & \overline{A}_{12} & B_1 \\ \overline{A}_{21} & \overline{A}_{22} & 0 \\ C_1 & 0 & 0 \end{bmatrix}, \tag{5.2.5}$$

where C_1 and B_1 are the same as the previous expression, but $\bar{A}_{21}$ and $\bar{A}_{22}$ may be different from A_{21} and A_{22}, respectively. It should be noted that the matrix (5.2.5) is obtained depending on relative degree, but not depending on the observability. (5.2.5) will be often used in the sequel.

5.2.3.3 Hurwitz Vectors

In this part, we will give a new definition.

Let $h \in \mathbb{R}^n$, $h = \begin{bmatrix} h_1 & h_2 & \cdots & h_n \end{bmatrix}^T$, h is called a Hurwitz vector if

$$h(x) = h_1 x^{n-1} + h_2 x^{n-2} + \cdots + h_{n-1}x + h_n$$

is a Hurwitz polynomial. We do not suppose that $h_1 \neq 0$ in the Hurwitz vector. If $h_1 \neq 0$, then none of $h_i, i = 2, 3, \ldots, n$ cannot be equal to 0, and their signs are the same as that of h_1. In order to be simple, we assume that $h_1 > 0$. Furthermore, if $h_1 = h_2 = \cdots = h_{i-1} = 0$, but $h_i \neq 0$, then we assume that $h_i > 0$, and the signs of $h_{i+1}, \ldots, h_n$ have to be the same as that of h_i.

The following conclusion is about the Hurwitz vector.

Lemma 5.2.1 Let $h = \begin{bmatrix} h_1 & h_2 & \cdots & h_n \end{bmatrix} \in \mathbb{R}^n$ be a Hurwitz vector. If $h_1 = h_2 = \cdots = h_{i-1} = 0$ but $h_i \neq 0$, and A has the Brunovsky controllable canonical (5.2.4), then there exists a $\lambda \in \mathbb{R}$ such that $\widehat{h} = (A + \lambda I)\, h$ is a Hurwitz vector, and $\widehat{h}_{i-1} \neq 0$ in $\hat{h}$.

Proof Denote that $h(s) = h_i s^{n-i} + h_{i+1}s^{n-i-1} + \cdots + h_{n-1}s + h_n$. It is direct to verify that the polynomial corresponding to $\hat{h}$ is

$$\widehat{h}(s) = (s + \lambda)\, h(s) - a_{n-i+1}h_i - a_{n-i}h_{i+1} - \cdots - a_1 h_n.$$

Denote that $\rho = a_{n-i+1}h_i + a_{n-i}h_{i+1} + \cdots + a_1 h_n$. Consider the equation $\widehat{h}(s) = sh(s) + \lambda h(s) - \rho = 0$. It can be written as

$$\frac{h(s)}{sh(s) - \rho} = -\frac{1}{\lambda}.$$

It is noted that the degree of denominator polynomial is $n-i+1$, and the degree of nominator polynomial is $n-i$. By the theory of root locus, when $\lambda \to \infty$, $n-i$ roots converge to the zeroes of $h(s)$, and one root converges to $-\infty$. Therefore, if $h(s)$ is a Hurwitz polynomial, and λ is large enough, then $n-i+1$ roots of $\hat{h}(s)$ all stay at the left half plane. □

Lemma 5.2.1 is proved under the assumption that A has the Brunovsky controllable canonical. In fact, the conclusion is also valid when A has other forms, which is left to readers as an exercise.

Lemma 5.2.1 can be used repeatedly. If h is a Hurwitz vector, $h_1 = h_2 = \cdots = h_{i-1} = 0$, and $h_i \neq 0$, then there always exist $\lambda_1, \ldots, \lambda_{n-i} > 0$, such that $\widehat{h} = (A+\lambda_{n-1}I)\cdots(A+\lambda_1 I)\, h$, and $\hat{h}$ is a Hurwitz vector, whose first component is not zero.

5.2.4 *Feedback Positive Realness*

As pointed out in the former subsection, the key of stabilization for Luré differential inclusion System (5.2.2) is to make the linear part positive real or strictly positive real. Two useful conclusions for feedback positive realness will be given in this subsection. Let us consider the single-variable system firstly.

Lemmas 5.2.2 Consider the single-variable system (c^{T}, A, b) which is both controllable and observable. There exists an output feedback $u = ky$ such that $\left(c^{\mathrm{T}}, A + kbc^{\mathrm{T}}, b\right)$ is strictly positive real if and only if the following two conditions hold.

1. $c^{\mathrm{T}} b > 0$.
2. The system is minimum phase.

Proof Sufficiency. If $c^T b \neq 0$, then the relative degree is 1. (c^T, A, b) can be written in the form of (5.2.5), i.e.,

$$c^T = \begin{bmatrix} 1 \; 0 \cdots 0 \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{5.2.6}$$

where $A_{11} \in \mathbb{R}$, $b_1 \in \mathbb{R}$, and $b_1 > 0$. By using the matrix theory, we can obtain that

$$\begin{bmatrix} s - A_{11} & -A_{12} \\ -A_{21} & sI - A_{22} \end{bmatrix}^{-1} = \frac{1}{\det(sI - A)} \begin{bmatrix} \det(sI - A_{22}) & * \\ * & * \end{bmatrix},$$

where $*$ stands for the element which does not affect the result of the following computation. We also know that

$$c^T(sI-A)^{-1}b = \frac{\det(sI-A_{22})\,b_1}{\det(sI-A)}.$$

The Condition 2 of Lemma 5.2.2 implies that $\det(sI-A_{22})$ is a Hurwitz polynomial, i.e., A_{22} is Hurwitz matrix. Then, for any positive definite matrix Q_{22}, there exists positive definite matrix P_{22} such that $P_{22}A_{22}+A_{22}^TP_{22}=-Q_{22}$.

Let $P_{11}=b_1^{-1}$. Then $P_{11}b_1=1$. By choosing

$$Q_{21}=Q_{12}^T=-\left(P_{22}A_{21}+A_{12}^TP_{11}\right),$$

then we have

$$\begin{aligned}
&\begin{bmatrix} P_{11} & 0 \\ 0 & P_{22}\end{bmatrix}\begin{bmatrix} A_{11}+kb_1 & A_{12} \\ A_{21} & A_{22}\end{bmatrix}+\begin{bmatrix} A_{11}+kb_1 & A_{21}^T \\ A_{12}^{\mathrm{T}} & A_{22}^T\end{bmatrix}\begin{bmatrix} P_{11} & 0 \\ 0 & P_{22}\end{bmatrix}\\
&\quad=\begin{bmatrix} 2P_{11}(A_{11}+kb_1) & P_{11}A_{12}+A_{21}^TP_{22} \\ A_{12}^{\mathrm{T}}P_{11}+P_{22}A_{21} & P_{22}A_{22}+A_{22}^TP_{22}\end{bmatrix}\\
&\quad=\begin{bmatrix} 2b_1^{-1}(A_{11}+kb_1) & -Q_{21}^T \\ -Q_{21} & -Q_{22}\end{bmatrix}.
\end{aligned}$$

k is chosen as

$$k<-\frac{1}{2}\left(2b_1^{-1}A_{11}+Q_{21}^TQ_{22}^{-1}Q_{21}\right),$$

then by Schur complements (Lemma 4.1.1)

$$\begin{bmatrix} 2b_1^{-1}(A_{11}+kb_1) & -Q_{21}^T \\ -Q_{21} & -Q_{22}\end{bmatrix}<0.$$

Thus, if $P=\begin{bmatrix} P_{11} & 0 \\ 0 & P_{22}\end{bmatrix}$, $Q=\begin{bmatrix} -2(A_{11}+kb_1) & Q_{21}^{\mathrm{T}} \\ Q_{21} & Q_{22}\end{bmatrix}$, then P and Q are both positive definite, and the following equations hold

$$\begin{aligned}
&P\left(A+kbc^T\right)+\left(A+kbc^T\right)^TP=-Q,\\
&Pb=c.
\end{aligned}\tag{5.2.7}$$

By the statements after Theorem 5.1.8, $\left(c^T, A+kbc^T, b\right)$ is strictly positive definite.

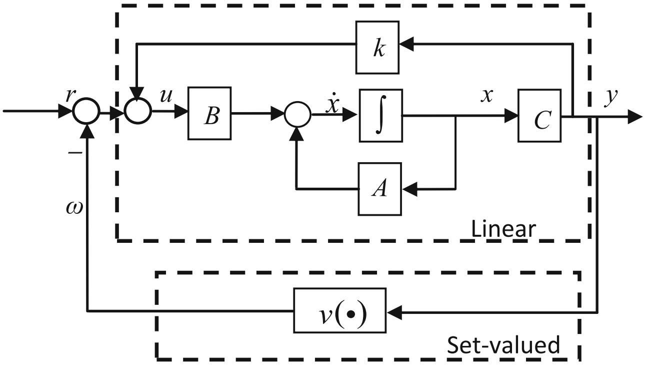


Fig. 5.5 The closed system for feedback positive realness

Necessity. By Corollary 5.1.1, $c^T b > 0$. According to Theorem 5.1.5, the numerator of transfer function is a Hurwitz polynomial; thus, the finite zeroes of the system are all stable, i.e., the system is minimum phase. □

The following corollary gives the frequency form of Lemma 5.2.2.

Corollary 5.2.1 Suppose the system (c^T, A, b) is both controllable and observable, the transfer function is

$$G(s) = \frac{\beta_1 s^{n-1} + \cdots + \beta_{n-1} s + \beta_n}{s^n + \alpha_1 s^{n-1} + \cdots + \alpha_{n-1} s + \alpha_n}.$$

There exists the output feedback $u = ky$ such that the closed-loop system $\left(c^T, A + kbc^T, b\right)$ is strictly positive real if and only if $\left[\beta_1\ \beta_2 \cdots \beta_n\right]$ is a Hurwitz vector, and the leading coefficient $\beta_1 > 0$. □

The scheme of feedback positive realness presented in Lemma 5.2.2 and Corollary 5.2.1 is drawn in Fig. 5.5. The linear part within the dashed line is strictly positive real; hence, by Theorem 5.2.1, the closed-loop system is asymptotically stable. However, it may be fail since confluent point of the output feedback $u = ky$ is behind the confluent point of ω. If the original system which consists of linear part and set-valued mapping is indivisible, then we cannot insert the feedback behind that confluent point of ω. Fortunately, the system drawn in Fig. 5.5 is equivalent to that in Fig. 5.6 by block diagram operation. The scheme is then realizable. The feedback is outside of the original system.

We now extend Lemma 5.2.2 to the multivariable case. Let $A \in \mathbb{R}^{n\times n}$. A is decomposed as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

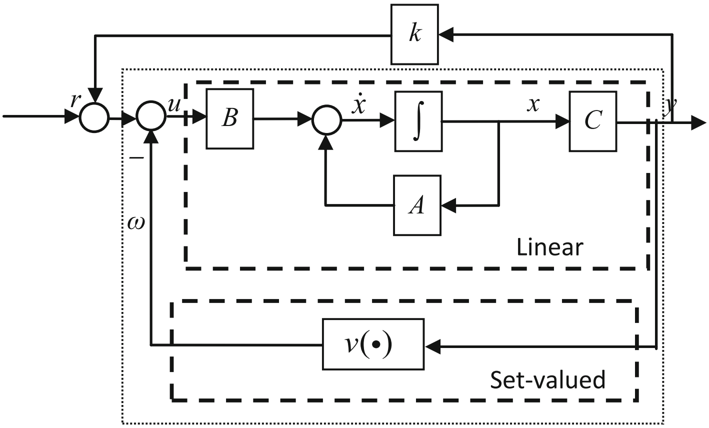


Fig. 5.6 The revised structure of feedback system

where $A_{11} \in \mathbb{R}^{n_1 \times n_1}$. If A_{22} is invertible, then A can be written as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I & A_{12}A_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ A_{22}^{-1}A_{21} & I \end{bmatrix},$$

where two Is are all identity matrices, but their dimensions may be different. According to the decomposition, if A is invertible, then we have that

$$A^{-1} = \begin{bmatrix} I & 0 \\ -A_{22}^{-1}A_{21} & I \end{bmatrix} \begin{bmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} & 0 \\ 0 & A_{22}^{-1} \end{bmatrix} \begin{bmatrix} I & -A_{12}A_{22}^{-1} \\ 0 & I \end{bmatrix}. \quad (5.2.8)$$

Theorem 5.2.2 Consider the multivariable system (C, A, B), which is both controllable and observable. There exists an output feedback $u = Ky$ such that $(C,\ A + BKC,\ B)$ is strictly positive real if and only if the following two conditions hold.

(1) CB is positive definite.
(2) The system is minimum phase.

Proof Sufficiency. Since CB is invertible, (C, A, B) can be written in the form of (5.2.4), i.e.,

$$C = \begin{bmatrix} C_1 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix},$$

where $C_1, A_{11}, B_1 \in \mathbb{R}^{m \times m}$ and C_1, B_1 are invertible. Specially, if $B_1 = I$, then C_1 is positive definite, and the feedback is chosen as $u = ky, k \in \mathbb{R}$. The transfer function $G(s) = C(sI - A)^{-1}B$, by Eq. (5.2.8), has the following form

$$\begin{aligned}G(s) &= \begin{bmatrix} C_1 & 0 \end{bmatrix}\begin{bmatrix} (sI - A_{22}) & -A_{12} \\ -A_{21} & (sI - A_{22}) \end{bmatrix}^{-1}\begin{bmatrix} B_1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} C_1 & 0 \end{bmatrix}\begin{bmatrix} I & 0 \\ (sI - A_{22})^{-1}A_{21} & I \end{bmatrix} \\ &\quad\times\begin{bmatrix} \left[sI - A_{11} - A_{12}(sI - A_{22})^{-1}A_{21}\right]^{-1} & 0 \\ 0 & (sI - A_{22})^{-1} \end{bmatrix}\begin{bmatrix} I & A_{12}(sI - A_{22})^{-1} \\ 0 & I \end{bmatrix}\begin{bmatrix} B_1 \\ 0 \end{bmatrix} \\ &= C_1\left[sI - A_{11} - A_{12}(sI - A_{22})^{-1}A_{21}\right]^{-1},\end{aligned}$$

where two Is are identity matrices with appropriate dimensions. Since $(sI - A_{22})^{-1} = (\text{Adj}\,(sI - A_{22}))\,/\,(\det(sI - A_{22}))$, where $\text{Adj}\,(sI - A_{22})$ is adjoint matrix of $(sI - A_{22})$ and is a polynomial matrix, then

$$G(s) = [C_1 \det(sI - A_{22})]\,[sI - A_{11} - A_{12}\text{Adj}\,(sI - A_{22})\,A_{21}]^{-1} \tag{5.2.9}$$

is a left composition of $G(s)$. By Property 1 given behind the definition of finite zero, the zeroes of system are all the roots of $\det(sI - A_{22})$. Condition 2 of Theorem 5.2.2 implies that A_{22} is a Hurwitz matrix; thus, for any positive definite matrix $Q_{22} \in \mathbb{R}^{(n-m)\times(n-m)}$, there always exists a positive definite matrix P_{22} such that $P_{22}A_{22} + A_{22}^TP_{22} = -Q_{22}$.

Denote that $P_{11} = C_1$ and $Q_{21} = Q_{12}^T = -\left(P_{22}A_{21} + A_{12}^TP_{11}\right)$. Consider

$$\begin{aligned}&\begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix}\begin{bmatrix} A_{11} + kC_1 & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} A_{11}^T + kC_1 & A_{21}^T \\ A_{12}^T & A_{22}^T \end{bmatrix}\begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix} \\ &\quad= \begin{bmatrix} P_{11}A_{11} + A_{11}^TP_{11} + 2kC_1^2 & P_{11}A_{12} + A_{21}^TP_{22} \\ A_{12}^TP_{11} + P_{22}A_{21} & P_{22}A_{22} + A_{22}^TP_{22} \end{bmatrix} \\ &\quad= \begin{bmatrix} P_{11}A_{11} + A_{11}^TP_{11} + 2kC_1^2 & -Q_{21}^T \\ -Q_{21} & -Q_{22} \end{bmatrix}.\end{aligned}$$

Thus, if $P_{11}A_{11} + A_{11}^TP_{11} + 2kC_1^2 - Q_{21}^{\mathrm{T}}Q_{22}^{-1}Q_{21} < 0$, then

$$\begin{bmatrix} P_{11}A_{11} + A_{11}^TP_{11} + 2kC_1^2 & -Q_{21}^{\mathrm{T}} \\ -Q_{21} & -Q_{22} \end{bmatrix} < 0.$$

The above inequality is feasible since $C_1^{-2}\left(Q_{21}^TQ_{22}^{-1}Q_{21} - P_{11}A_{11} - A_{11}^TP_{11}\right)$ is symmetric matrix. Denote the minimum eigenvalue of the matrix by $\lambda_{\min}$, and if $k < (1/2)\,\lambda_{\min}$, then $P_{11}A_{11} + A_{11}^TP_{11} + 2kC_1^2 - Q_{21}^TQ_{22}^{-1}Q_{21} < 0$. By Lemma 4.1.1, we obtain

$$\begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix}\begin{bmatrix} A_{11} + kC_1 & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} A_{11}^T + kC_1 & A_{21}^T \\ A_{12}^T & A_{22}^T \end{bmatrix}\begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix} < 0,$$

and we also have $\begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix} \begin{bmatrix} B_1 \\ 0 \end{bmatrix} = \begin{bmatrix} C_1^T \\ 0 \end{bmatrix}$. According to Theorem 5.1.8, $\left(C \; A + kBC \; B \right)$ is strictly positive real.

Necessity. If $(C, A + BKC, B)$ is strictly positive real, the Condition 1 is verified by Corollary 5.1.1. By Problem 5.1.4 (3), if the left prime composition of $G_K(s) = C(sI - A - BKC)^{-1}B$ is made, $G_K(s) = N_r(s)D_r^{-1}(s)$, then $\det N_r(s)$ is a Hurwitz polynomial, i.e., the zeroes of $(C, A + BKC, B)$ are all stable. By the Property 3 of zeroes, the zeroes of $(C, A + BKC, B)$ are the same as that of (C, A, B); thus, (C, A, B) is minimum phase. □

The same problem has been considered in Huang et al. (1999) by an alternative approach. The frequency criterion for multivariable system is left in Problems of this section, and the readers can derive it by the same method used in this section.

It is obvious that the feedback of multivariable system can be realized by the scheme given in Fig. 5.6.

5.2.5 Feedback Stabilization – Single-Variable Systems

Theorem 5.2.2 provides an approach of stabilization when the relative degree of system is one. The aim of the subsection aims at extending the result to the case where the relative degree is larger than one. Firstly, PD compensation for output is used to change the relative degree of the system, and then stabilization is achieved by Theorem 5.2.2. Secondly, state feedback is employed to change the relative degree of the system. This result establishes a basis for the design of state observer in the next section.

The stabilization for the system whose relative degree is larger than 1 will be discussed in the following two subsections, respectively. Firstly, we consider the single-variable system (c^T, A, b); secondly, we consider multivariable system (C, A, B) at the next subsection.

If the relative degree of the linear part is 2, then the transfer function is

$$G(s) = \frac{\beta_2 s^{n-2} + \cdots + \beta_{n-1}s + \beta_n}{s^n + \alpha_1 s^{n-1} + \cdots + \alpha_{n-1}s + \alpha_n}.$$

By Corollary 5.2.1, there dose not exist output feedback $u = ky$ such that the transfer function of closed-loop system is strictly positive real. Hence, an auxiliary output is introduced as $z(t) = \lambda y(t) + \dot{y}(t)$, i.e., $z(s) = (\lambda + s)\, y(s)$. It is equivalent to cascade a PD control in the output term. The transfer function from $u(s)$ to $z(s)$ becomes

$$G_z(s) = \frac{z(s)}{u(s)} = \frac{(s + \lambda)\left(\beta_2 s^{n-2} + \cdots + \beta_{n-1}s + \beta_n\right)}{s^n + \alpha_1 s^{n-1} + \cdots + \alpha_{n-1}s + \alpha_n}.$$

The relative degree is 1. If we choose $\lambda > 0$, and the system is minimum phase and $\beta_2, \beta_3, \ldots, \beta_n$ are all positive, then there exists output feedback $u = kz$ such that the DI system is absolutely stable by Corollary 5.2.1.

The design method can be extended to the arbitrary relative degree case, i.e., $i > 1$. Design the polynomial $\lambda(s) = (\lambda_1 + s)\cdots(\lambda_{i-1} + s)$ with degree of $i-1$,[2] where $\lambda_p > 0, p = 1, 2, \ldots, i-1$, then the transfer function from u to z is $G_z(s) = \lambda(s)G(s)$, and its relative degree is 1. There always exists output feedback such that the closed-loop system is strictly positive real by Corollary 5.2.1. Thus, the whole Luré differential inclusion system is strongly asymptotically stable. The above conclusion is summed up in the following corollary.

Corollary 5.2.2 Consider the single-variable system Luré differential inclusion system Eq. (5.2.2), if the numerator of the transfer function of linear part (c^{T}, A, b) is a Hurwitz polynomial, and the leading coefficient is positive, then there always exists a real polynomial $\lambda(s)$ and output feedback $u = kz$ where $z(s) = \lambda(s)y(s)$, such that by the feedback, closed-loop system is positive real. □

Remark Since zeroes of the numerator polynomial are the same as that of the system, the condition is just that the linear part (c^T, A, b) is minimum phase. □

The stabilization scheme presented in Corollary 5.2.2 can be realized by Fig. 5.7. In Fig. 5.7, we require that a compensator can be inserted to the system, otherwise we cannot achieve the positive realness of the linear part.

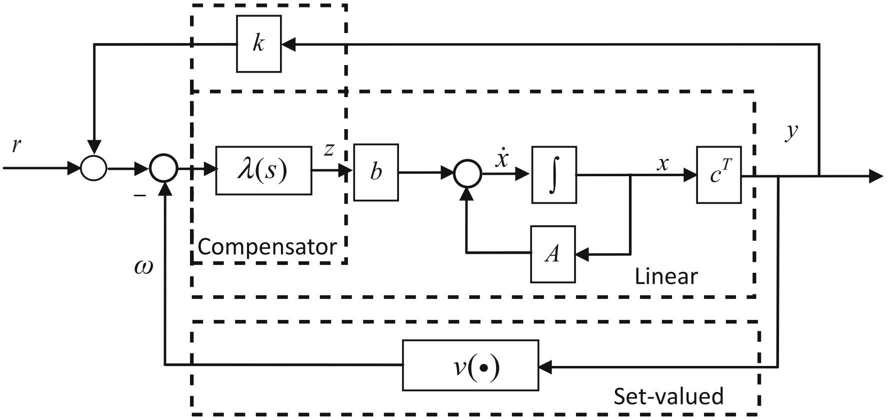


Fig. 5.7 The compensating scheme of Corollary 5.2.2

[2] In fact, it only needs the condition that $\lambda(s)$ is Hurwitz polynomial with first coefficient one. This assumption is convenient to the proof.

5.2.6 Feedback Stabilization – Multivariable Systems

Let us turn to multivariable systems. Based on the result of single-variable systems, the discussion for the multivariable systems becomes simple. Consider Luré differential inclusion system Eq. (5.2.2), its linear part is (C, A, B). Denote that $B = \begin{bmatrix} b_1 \ b_2 \cdots b_m \end{bmatrix}$, where $b_i \in \mathbb{R}^n$, $i = 1, 2, \ldots, m$, then the transfer function matrix is

$$G(s) = \begin{bmatrix} C(sI - A)^{-1} b_1 \ C(sI - A)^{-1} b_2 \cdots C(sI - A)^{-1} b_m \end{bmatrix}.$$

For the sake of simplicity, we denote that $G_i(s) = C(sI - A)^{-1} b_i$.

By a similar method introduced in Sect. 5.2.3, we can define the relative degree for column vector for $G_i(s)$. If $\lim\limits_{s \to \infty} s^j G_i(s) = 0, j = 0, 1, 2, \ldots, q_i - 1$; $\lim\limits_{s \to \infty} s^{q_i} G_i(s) \neq 0$, and

$$\operatorname{rank} \left[\lim_{s \to \infty} s^{q_1} G_1(s) \ \lim_{s \to \infty} s^{q_2} G_2(s) \cdots \lim_{s \to \infty} s^{q_m} G_m(s) \right] = \min(r, m),$$

then the ith column relative degree of $G(s)$ is q_i, the set $\{q_1, q_2, \ldots, q_m\}$ is called the set of column relative degree of $G(s)$, and denote

$$\overline{\Pi} = \begin{bmatrix} CA^{p_1} b_1 \ CA^{p_2} b_2 \cdots CA^{p_m} b_m \end{bmatrix}.$$

It is easy to see that the column relative degree holds the same properties as the relative degree.

Let the ith column relative degree be q_i, then for any polynomial with $q_i - 1$ degree $\Lambda_i(s) = (\lambda_{i,1} + s) \cdots (\lambda_{i,q_i - 1} + s)$, we have $\lim\limits_{s \to \infty} s c_i^{\mathrm{T}} \Lambda_i(s)(sI - A)^{-1} B = c_i^{\mathrm{T}} A^{p_i} B$. Since $\overline{\Pi}$ is invertible, the relative degrees of

$$\Lambda(s) G(s) = \begin{bmatrix} C(sI - A)^{-1} b_1 \Lambda_1(s) \ C(sI - A)^{-1} b_2 \Lambda_2(s) \cdots C(sI - A)^{-1} b_m \Lambda_m(s) \end{bmatrix}$$

are all 1, where $\Lambda(s) = \operatorname{diag}\{\Lambda_1(s) \ \cdots \ \Lambda_r(s)\}$.

If $G(s) = D_l^{-1}(s) N_l(s)$ is a left coprime factorization, then there exists $\Lambda(s)$ such that $\Lambda(s) N_l(s)$ is left coprime to $D_l(s)$; thus, the finite zeroes of $N_l(s)\Lambda(s)$ are the same as those of $\Lambda(s) G(s)$. The minimal realization of $\Lambda(s) G(s)$ is

$$\begin{aligned} \dot{\xi} &= \mathfrak{A} \xi + \mathfrak{B} u, \\ \zeta &= \mathfrak{C} \xi. \end{aligned}$$

The relative degree is 1 and $\mathfrak{C}\mathfrak{B} = \overline{\Pi}$, and the zeroes of the system are the zeroes of $N_l(s)\Lambda(s)$. By Theorem 5.2.2, we can conclude the following result.

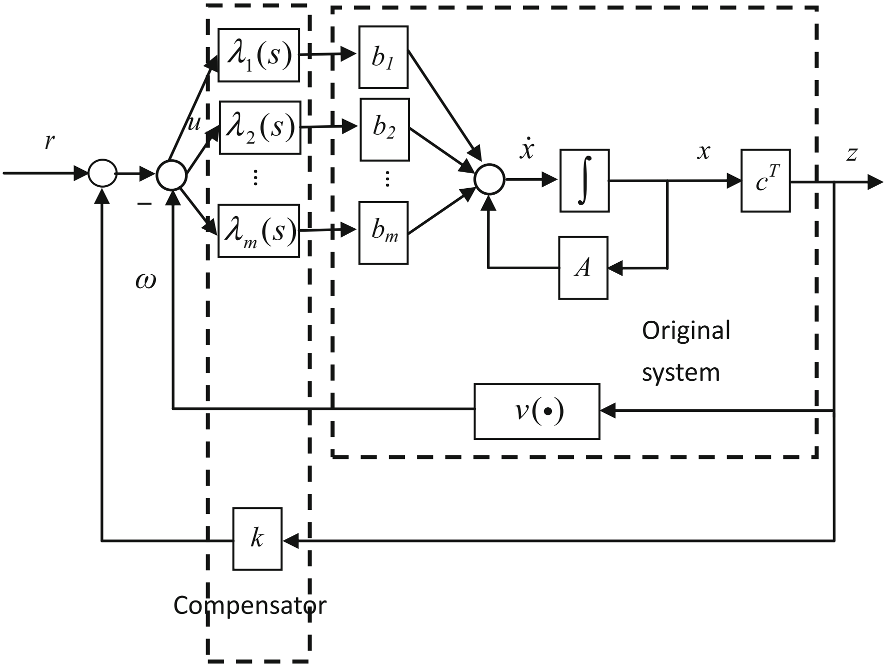


Fig. 5.8 The compensation of Theorem 5.2.3

Theorem 5.2.3 Consider the Luré differential inclusion system Eq. (5.2.2). If the set-valued mapping $v(\cdot)$ is monotone, the zeroes of linear part (C, A, B) are all stable, and the decoupling matrix $\overline{\Pi}$ is positive definite, then there exists a feedforward $\overline{B} = B\Lambda(s)$ and feedback $u(t) = kz(t)$ such that the closed system is asymptotically stable. □

The compensation can be realized by the scheme given in Fig. 5.8

Problems

1. Let $M(s)$ be a $r \times m$ polynomial matrix and $M(s) = [m_1(s), \dots, m_m(s)]$, where $m_i(s)$ is the ith column of $M(s)$. The maximal degree of the r polynomials in $m_i(s)$ is the column degree of the ith column of $M(s)$ and denoted by d_i, then $m_i(s)$ can be written as $m_i(s) = m_{d_1}s^{d_1} + m_{d_1-1}s^{d_1-1} + \cdots + m_0$, where $m_{d_i} \in \mathbb{R}^r$, m_{d_i} is called the leading coefficient of the ith column of . Denote that $M_D = [m_{d_1} \cdots m_{d_m}] \in \mathbb{R}^{r\times m}$, and M_D is called the column degree matrix of $M(s)$. If rank $M_D < \min(r, m)$, please prove that there exists unimodular matrices $A(s)$ and $B(s)$ with compatible dimensions such that the rank of column degree matrix of $A(s)M(s)$ and rank of column degree matrix of $M(s)B(s)$ are $\min(r, m)$. (This transformation is called regular transformation.)

2. Suppose (C,A,B) is a both controllable and observable system, and $r = m$. Let $\{p_1,p_2,\ldots,p_r\}$ and $\{q_1,q_2,\ldots,q_m\}$ be the sets of relative degree and column relative degree, respectively. Then

   (1) $\{p_1,p_2,\ldots,p_r\} = \{q_1,q_2,\ldots,q_m\}$;
   (2) $\Pi = \overline{\Pi}$.

3. Let

$$A = \begin{bmatrix} 0 & 2 & 1 \\ 2 & 0 & 1 \\ 2 & 1 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 5 & 2 \\ -1 & 2 & 1 \end{bmatrix}.$$

   (1) Compute that the zeroes of the system and then determine whether the system is minimum phase system or not.
   (2) Please design the feedback law such that the system is strictly positive real.

4. Let

$$A = \begin{bmatrix} 0 & 2 & -1 \\ 2 & 0 & 1 \\ 2 & 1 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 4 & 2 \\ -1 & 2 & 1 \end{bmatrix}.$$

   (1) Compute that the zeroes of the system and then determine whether the system is minimum phase system or not
   (2) Design $\Delta(s) = \begin{bmatrix} \lambda_1(s) & 0 \\ 0 & \lambda_2(s) \end{bmatrix}$ such that $C(sI-A)^{-1}B\Delta(s)$ is minimum phase.
   (3) Design a feedback law such that the system is strictly real positive.

5. Suppose (C,A,B) is a both controllable and observable system. If there exists a state feedback $u = Fx$ such that the closed-loop system $(C, A + BF, B)$ is strictly positive real, then the sufficient and necessary conditions are

   (1) $CB > 0$;
   (2) (C,A,B) is minimum phase.

   The problem shows for the problem of feedback positive realness, the conditions of state feedback are equal to those for output feedback.

6. Proof that the set of relative degrees is invariant under coordinate transformation, state feedback and output insertion, respectively.

5.3 Luenberger Observers and Separated Design

The observer design of control system is always a hot topic. Since the Luré differential inclusion system contains the set-valued function which leads that the output includes the uncertainty, the observer design for the system becomes

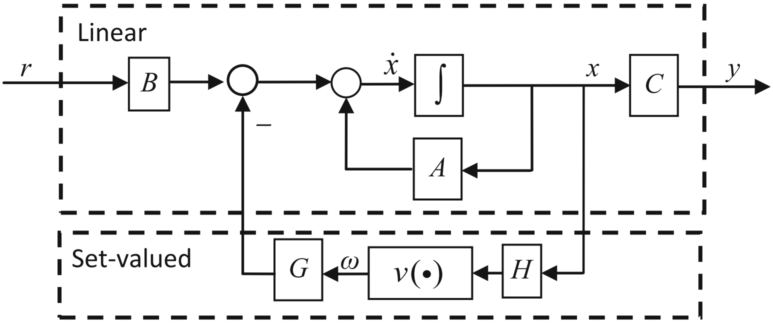


Fig. 5.9 Luré differential inclusion system

different. The observer design problem for the Luré differential inclusion system is studied in this section, a full-order observer is designed firstly, then the separation design principle is investigated.

Consider the following Luré differential inclusion system (Fig. 5.9),

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Br(t) - G\omega(t),\\ \omega(t) &\in \nu\left(Hx(t)\right),\\ y(t) &= Cx(t),\end{aligned} \tag{5.3.1}$$

where $x(t) \in \mathbb{R}^n$, $r(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^r$ are the state, input and output of the system, respectively, $\nu(\cdot)$ is a set-valued mapping and $\omega(t) \in \mathbb{R}^q$ is the output of the set-valued mapping. A, B, C are the given matrices with compatible dimensions. In order to consider more general case, an input matrix $H \in \mathbb{R}^{q\times n}$ and an output matrix $G \in \mathbb{R}^{n\times q}$ are added to the set-valued mapping. Without loss of generality, it is assumed that B and G are of full column ranks and C is of full row rank. Later, we will prove that H is of full row rank if there exists a Luenberger observer for Inc. (5.3.1). Otherwise, G is of full column rank if H is of full row rank. If $G = B$ and $H = C$, then Inc. (5.3.1) reduces to Inc. (5.2.2). Thus, Inc. (5.3.1) can be treated as a special case of Inc. (5.2.2). In the Luré differential inclusion system, set-valued mapping $\nu(\cdot)$ is always assumed to be monotone, i.e., if $\omega_i \in \nu\left(Hx_i\right)$, $i = 1, 2$, then $\langle \omega_1 - \omega_2, Hx_1 - Hx_2\rangle \geq 0$.

5.3.1 Well-Posedness

In Inc. (5.3.1), the control input $r(t)$ is supposed to be a piecewise continuous function. Thus, the solution of Cauchy problem of the linear differential equation $\dot{x}(t) = Ax(t) + Br(t)$ always exists for every $x(0) = x_0$. Note that the left derivative is not equal to the right derivative on those discontinuous points. The discontinuity can be treated by the way presented in Sect. 2.3, where we can consider the Filippov solution for the equation $\dot{x}(t) = Ax(t) + Br(t)$ for a discontinuous function $r(t)$. Filippov solution always exists. The two solutions are consistent except at those discontinuous points.

When we consider the solution of Cauchy problem of the following differential inclusion,

$$\dot{x}(t) = Ax(t) + Br(t) - G\omega(t),$$

$$\omega(t) \in \nu\left(Hx(t)\right), \quad x(0) = x_0,$$

where $\omega \in \nu(\cdot)$, the existence of the solution is lack of theory support. The existence theorems for the differential inclusion given in Chap. 2 of this book require the set-valued mappings are Lipschitz and closed. When the set-valued mapping $\nu(\cdot)$ is only monotone, the existence of the differential inclusion did not be researched, even for the linear inclusion system like $\dot{x}(t) = Ax(t) + Br(t) - G\omega(t)$.

The existence for the differential inclusion system is also called well-posedness. In the observer design for the differential inclusion system, there exist two kinds of method to treat the well-posedness: One is to assume that for any $\omega(t) \in \nu\left(Hx(t)\right)$, the solution exists for the Cauchy problem of differential equation $\dot{x}(t) = Ax(t) + Br(t) - G\omega(t)$ with $x(0) = x_0$, or assume that set-valued mapping $\nu(\cdot)$ is closed, convex and Lipschitz. The another one is to assume that the set-valued mapping $\nu(\cdot)$ is maximal monotone; thus, the existence of solution of $\dot{x}(t) = Ax(t)+Br(t)-G\omega(t)$ with $x(0) = x_0$ is guaranteed; moreover, the uniqueness of the solution is also guaranteed except a subset whose measure is zero, i.e., with mathematical terminology, the solution exists almost everywhere.

The following lemma is essential to the further study.

Lemma 5.3.1 If the set-valued mapping $\nu(\cdot)$: $\mathbb{R}^m \to \mathbb{R}^m$ is maximal monotone, the mapping $x \mapsto H^{\mathrm{T}}\nu\left(Hx + h\right)$, $(\mathbb{R}^n \to \mathbb{R}^n)$ is also maximal monotone, where $x \in \mathbb{R}^n$, $h \in \mathbb{R}^m$, $H \in \mathbb{R}^{m\times n}$ and H are of full row rank.

Proof Let $y_i \in H^T\nu\left(Hx_i + h\right), \ i = 1, 2$, there exists $\omega_i \in \nu\left(Hx_i + h\right)$ such that $y_i = H^T\omega_i$. Since H is full of row rank, ω_i is unique when y_i is determined. Consider that

$$\langle y_1 - y_2,\ x_1 - x_2\rangle = \left\langle H^T\omega_1 - H^T\omega_2, x_1 - x_2\right\rangle = \langle \omega_1 - \omega_2,\ Hx_1 - Hx_2\rangle\,,$$

Because

$$\langle \omega_1 - \omega_2,\ (Hx_1 + h) - (Hx_2 + h)\rangle = \langle \omega_1 - \omega_2,\ Hx_1 - Hx_2\rangle \geq 0.$$

Thus, $\langle y_1 - y_2,\ x_1 - x_2\rangle \geq 0$, i.e., $x \mapsto H^T\nu\left(Hx + h\right)$ is monotone.

Now we prove that $x \mapsto H^T\nu\left(Hx + h\right)$ is maximal. It is obvious that $\nu\left(Hx + h\right)$ is maximal if $\nu(Hx)$ is maximal. Thus, it is necessary to prove that $x \mapsto H^T\nu(Hx)$ is maximal.

Let $y_1 \in H^T \nu(Hx_1)$. If there are $y_2 \in \mathrm{Im} H^T$ and a $x_2 \in \mathbb{R}^n$ such that $\langle y_1 - y_2, x_1 - x_2 \rangle \geq 0$. Since $y_2 \in \mathrm{Im} H^T$, there exists $\omega_2 \in \mathbb{R}^m$ such that $y_2 = H^T \omega_2$. By $y_1 \in H^T \nu(Hx_1)$, there exists a $\omega_1 \in \nu(Hx_1)$ such that $y_1 = H^T \omega_1$. Then $\langle y_1 - y_2, x_1 - x_2 \rangle \geq 0$ implies $\langle \omega_1 - \omega_2, Hx_1 - Hx_2 \rangle \geq 0$. $\nu(Hx)$ is maximal; thus, $\omega_2 \in \nu(Hx_2)$, i.e., $y_2 \in H^{\mathrm{T}} \nu(Hx_2)$. □

In the sequel, the differential inclusion system Inc. (5.3.1) is always assumed to be well posed, which contains the case where $\nu(\cdot)$ is maximal monotone.

5.3.2 The Luenberger State Observer

Consider the following observer

$$\begin{aligned} &\dot{\widehat{x}}(t) = (A - LC)\widehat{x}(t) + Br(t) - G\widehat{\omega}(t) + Ly(t), \\ &\widehat{\omega}(t) \in \nu\left(H\widehat{x}(t)\right), \end{aligned} \tag{5.3.2}$$

where $\widehat{x}$ is the estimated state and $\widehat{\omega}$ is the output of the set-valued mapping in the observer. $L \in \mathbb{R}^{n \times r}$ is called the observer gain and used to make $A-LC$ be a Hurwitz matrix. By the theory of linear system, if (A, C) is observable, there always exists an L such that $A-LC$ is a Hurwitz matrix. Figure 5.10 gives the structure of the observer and the connection with the original system to be observed. Denote the solution sets of Inclusions (5.3.1) and (5.3.2) by $S_{[1]}(x_0, r(t))$ and $S_{[2]}\left(\widehat{x}_0, r(t)\right)$, where $r(t)$ input, and x_0 and $\widehat{x}_0$ are the initial conditions x_0 for Incs. (5.3.1) and (5.3.2), respectively. If $x_0 \neq \widehat{x}_0$, for two solutions $x_1(t) \in S_{[1]}\left(x_0, r(t)\right)$ and $x_2(t) \in S_{[2]}\left(\widehat{x}_0, r(t)\right)$ selected arbitrarily, we have

$$\|x_1(t) - x_2(t)\| \to 0 \quad (t \to \infty),$$

then Inc. (5.3.2) is called a full-order Luenberger observer of Inc. (5.3.1).

Subtracting Inc. (5.3.2) from Inc. (5.3.1), we obtain that

$$\begin{aligned} &\dot{x} - \dot{\widehat{x}} = (A - LC)\left(x - \widehat{x}\right) + G\omega - G\widehat{\omega}, \\ &\omega \in \nu(Hx), \\ &\widehat{\omega} \in \nu\left(H\widehat{x}\right), \\ &y = Cx. \end{aligned} \tag{5.3.3}$$

Denote that $e = x - \widehat{x}$, e is called the observation error. Then the first three expressions In Inc. (5.3.3) can be rewritten as

$$\begin{aligned} \dot{e} &= (A - LC)\,e + G\left(\omega - \widehat{\omega}\right), \\ \omega &\in \nu(Hx), \quad \widehat{\omega} \in \nu\left(H\widehat{x}\right). \end{aligned} \tag{5.3.4}$$

The target of the observer design is to construct an L such that $e \to 0 \ (t \to \infty)$.

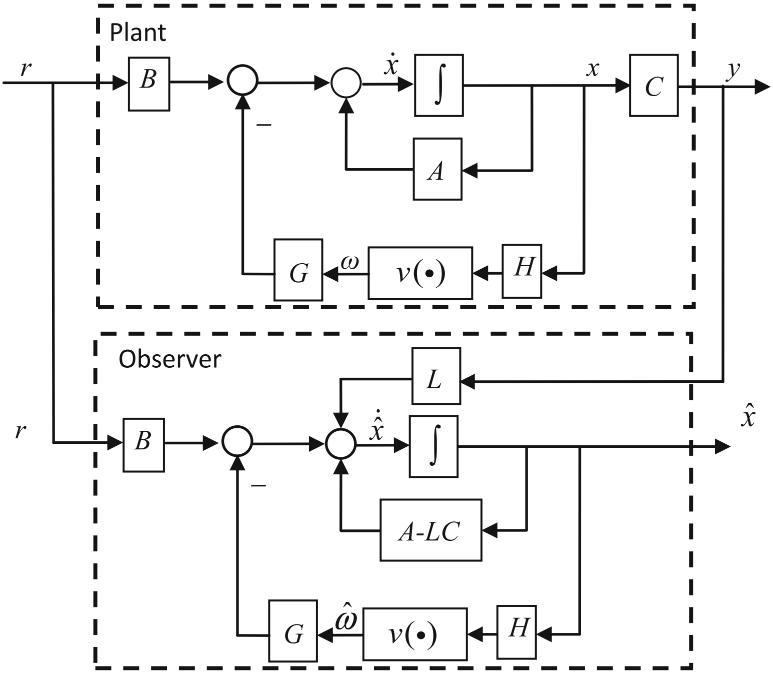


Fig. 5.10 The plant and observer

Theorem 5.3.1 Consider Incs. (5.3.1) and (5.3.2), if

1. There exists an $L \in \mathbb{R}^{n\times r}$ such that $(H, A - LC, G)$ is controllable, observable and strictly positive real.
2. $\nu(\cdot)$ is monotone.
3. Inc. (5.3.4) is well posed, i.e., its solution exists.

Then Inc. (5.3.2) is a Luenberger observer for Inc. (5.3.1).

Proof Let $z = Hx$ and $\widehat{z} = H\widehat{x}$ be the auxiliary output of the plant Inc. (5.3.1) and the observer Inc. (5.3.2), respectively. Since $\nu(\cdot)$ is monotone, then $\langle \omega - \widehat{\omega},\ z - \widehat{z}\rangle \geq 0$. Denote that $\mu = \omega - \widehat{\omega}$ and $\zeta = H\,(x - \widehat{x})$, then Inc. (5.3.4) is

$$\begin{aligned} \dot{e} &= (A - LC)\, e + G\mu, \\ \zeta &= He, \\ \mu &\in \nu(Hx) - \nu\,(H\widehat{x})\,. \end{aligned}$$

Let $\upsilon\,(e, x) = \nu(Hx) - \nu\,(H\widehat{x}) = \nu(Hx) - \nu\,(H\,(x - e))$. Obviously, $0 \in \upsilon\,(0, x)$ for every x. μ and ζ are treated as the input and output of linear part, respectively, $\langle \omega - \widehat{\omega},\ z - \widehat{z}\rangle \geq 0$ implies that $\langle \mu,\ \zeta\rangle \geq 0$. Thus, by Theorem 5.1.9, the proof is completed. □

By Theorem 5.3.1, the key of Luenberger observer design is to find an L such that $(H,\ A-LC,\ G)$ is strictly positive real. Differing from the system (C,A,B) considered in Sect. 5.2, where the feedback is realized by linking C and B. However, the system studied now is (H,A,G), but the feedback is not realized by H and G. Since there are three matrices C,H,G involved, the research may be more complicated. In what follows, we always suppose that (C,A,B) and (H,A,G) are both controllable and observable. The following lemma is very useful for the investigation of positive realness.

Lemma 5.3.2 Let P be an $m\times m$ real symmetry matrix, $U\in\mathbb{R}^{m\times n}$ and $V\in\mathbb{R}^{m\times l}$ are two matrices with full column ranks. Denote the orthogonal complement matrices of U and V by $U_\perp$[3] and $V_\perp$, respectively. Then there exists a $Q\in\mathbb{R}^{n\times l}$ such that

$$P+UQV^T+VQ^TU^T<0 \tag{5.3.5}$$

if and only if

$$(U_\perp)^TPU_\perp<0,\quad (V_\perp)^TPV_\perp<0.$$

The importance of Lemma 5.3.1 lies that U and V both appear in Inequality (5.3.5), but in the conditions of the lemma, $U_\perp$ and $V_\perp$ appear in two independent inequalities, and the conditions are much more easy to check. Lemma 5.3.1 is called positive real form of Parrott Theorem, the proof for it can be referred to Ly et al. (1994) and Huang et al. (1999). We omit it here, and the general form of Q is also presented in the reference.

It should be noted that if $m=n$, then U is of full column rank; it implies that U is an invertible matrix, so $U_\perp=0$. Meanwhile, the condition $(U_\perp)^TPU_\perp<0$ should be deleted. The fact can also be derived from the general Finsker Theorem. This is because if U is an invertible matrix, and let $K=UQ$, then Q can be determined uniquely by K. Let $K=kV$, then $P+UQV^T+VQ^TU^T=P+2kVV^T$. By Finsker theorem, there exists k such that $P+2kVV^T<0$ if and only if $V_\perp^TPV_\perp<0$.

Lemma 5.3.3 If G is of full column rank, then there exists an L such that $(H,\ A-LC,\ G)$ is strictly positive real if and only if the following hold simultaneously:

1. $HG>0$; hence, H is of full row rank.
2. There exists an $X\in\mathbb{R}^{(n-q)\times(n-q)}$ such that $G_\perp XG_\perp^T>0$ and

$$C_\perp\left[\left(H^T(HG)^{-1}H+G_\perp XG_\perp^T\right)A+A^T\left(H^T(HG)^{-1}H+G_\perp XG_\perp^T\right)\right]C_\perp^T<0.$$

[3] $U_\perp$ is an $(n-m)\times n$ matrix, such that $\begin{bmatrix}U\\U_\perp\end{bmatrix}$ is invertible and $UU_\perp^T=0$.

Proof If $(H,\ A-LC,\ G)$ is strictly positive real, then there exists a positive definite matrix P such that

$$\begin{aligned} &P(A-LC)+(A-LC)^TP<0,\\ &G^TP=H. \end{aligned} \tag{5.3.6}$$

Since G is full of column rank, $HG=G^TPG>0$, then Condition 1 is proved.

Let P be the solution of equation $G^TP=H$. Then P has the form of

$$P=H^T(HG)^{-1}H+G_\perp XG_\perp^T, \tag{5.3.7}$$

where $X\in\mathbb{R}^{(n-q)\times(n-q)}$ is any positive definite matrix. From the first one of Inequality (5.3.6), we have

$$PA+A^TP-PLC-(LC)^TP<0. \tag{5.3.8}$$

By Lemma 5.3.2, the sufficient and necessary condition of Inequality (5.3.8) is that[4]

$$C_\perp\left(PA+A^TP\right)C_\perp^T<0.$$

Substituting P by Eq. (5.3.7), we can obtain that

$$C_\perp\left[\left(H^T(HG)^{-1}H+G_\perp XG_\perp^T\right)A+A^T\left(H^T(HG)^{-1}H+G_\perp XG_\perp^T\right)\right]C_\perp^T<0.$$

Thus, the necessary condition is proved.

On the other hand, if the conditions of the theorem all hold, a P can be obtained such that $G^TP=H$. And from the second condition, it holds that $C_\perp\left(PA+A^TP\right)C_\perp^T<0$. By Lemma 5.3.2, there exists L such that Inequality (5.3.8) holds, i.e., $(H,A-LC,G)$ is strictly positive real. □

It is interesting that the only unknown variable is the matrix X in the second condition of the above theorem, and the inequality is linear with respect to X. Thus, it can be solved by LMI tools. In the exercise, we give a simple method to solve the inequality $C_\perp\left(PA+A^TP\right)C_\perp^T<0$, but it is independent of $G^TP=H$.

By Lemma 5.3.2, the further result can be obtained. Consider

$$C_\perp\left[\left(H^T(HG)^{-1}H+G_\perp XG_\perp^T\right)A+A^T\left(H^T(HG)^{-1}H+G_\perp XG_\perp^T\right)\right]C_\perp^T<0.$$

[4] In order to be simple, since C is full of row rank, $C_\perp$ is $(n-r)\times n$ matrix, and $\begin{bmatrix} C\\ C_\perp\end{bmatrix}$ is invertible with $CC_\perp^T=0$.

In order to be simpler, we denote that $M = C_\perp G_\perp, N = H^T(HG)^{-1}H$, then $M \in \mathbb{R}^{(n-r)\times(n-q)}$, $N \in \mathbb{R}^{n\times n}$ and N are semipositive definite matrix. By these notations, the above inequality becomes

$$C_\perp \left(NA + A^T N\right) C_\perp^T + MXG_\perp^T AC_\perp^T + C_\perp A^T G_\perp XM^T < 0. \tag{5.3.9}$$

By using Lemma 5.3.2 again, the sufficient and necessary condition for Inequality (5.3.9) is

$$M_\perp^T C_\perp \left(NA + A^T N\right) C_\perp^T M_\perp < 0, \quad \left(C_\perp A^T G_\perp\right)_\perp^T C_\perp \left(NA + A^T N\right) C_\perp^T C_\perp A^T G_\perp < 0.$$

Thus, we can get the following theorem.

Theorem 5.3.2 If G is of full column rank, then there exists an L such that $(H, A - LC, G)$ is strictly positive real if and only if the following conditions hold simultaneously.

1. $HG > 0$; hence, H is of full row rank.
2. $M_\perp^T C_\perp \left(NA + A^T N\right) C_\perp^T M_\perp < 0, \quad \left(C_\perp A^T G_\perp\right)_\perp^T C_\perp \left(NA + A^T N\right) C_\perp^T C_\perp A^T G_\perp < 0,$

where $M = C_\perp G_\perp, \quad N = H^T(HG)^{-1}H.$ □

Lemma 5.3.4 Let $A \in \mathbb{R}^{n\times n}$, and let $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{r\times n}$ be of full row rank and of full column rank, respectively, then for $K \in \mathbb{R}^{m\times r}$, the following equation holds

$$\max_{K\in \mathrm{R}^{m\times r}} \operatorname{rank}(A + BKC) = \min\left\{\operatorname{rank}[A \;\; B], \;\; \operatorname{rank}\begin{bmatrix} A \\ C \end{bmatrix}\right\}.$$

Lemma 5.3.4 is usually found in the exercises in the theory of matrix, the proof for it is referred to Han (1993).

A property $\mathfrak{P}$ is called an open property in $\mathbb{R}^n$. If a property $\mathfrak{P}$ holds for an $x_0 \in \mathbb{R}^n$, then there exists a neighborhood O of x_0 such that the property $\mathfrak{P}$ holds for every $x \in O$. By the definition, the above property of the maximal rank of $A + BKC$ for $K \in \mathbb{R}^{m\times r}$ is an open property.

Another familiar open property is: Let $s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n$ be a Hurwitz polynomial. Then the vector $[1 \; a_1 \; a_2 \; \cdots \; a_n]^T \in \mathbb{R}^{n+1}$ is a Hurwitz vector. A Hurwitz vector satisfies open property, i.e., there exits $\varepsilon > 0$, $(1+\delta)s^n + (a_1+\delta)s^{n-1} + \cdots + (a_{n-1}+\delta)s + a_n + \delta$ for any $\delta \in [-\varepsilon, \; \varepsilon]$ is a Hurwitz polynomial; especially, $[1 \;\; a_1+\delta \;\; a_2+\delta \;\; \cdots \;\; a_n+\delta]^T$ is always a Hurwitz vector. Similarly, if $[a_1 \; a_2 \; \cdots \; a_n; \; b_1 \; b_2 \; \cdots \; b_n]^T$ is treated as an $\mathbb{R}^{2n}$ vector and consists of coefficients of rational function $\left(b_1 s^{n-1} + \cdots + b_{n-1}s + b_n\right) / \left(s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n\right)$, then the strictly positive realness of rational function $\left(b_1 s^{n-1} + \cdots + b_{n-1}s + b_n\right) / \left(s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n\right)$ is open property in $\mathbb{R}^{2n}$. The result can be extended. Let $G(s)$ be an $r \times m$ rational function matrix, the vector $g_0 \in \mathbb{R}^l$ consists of the coefficients of $G(s)$. If $G(s)$ is a strictly positive real matrix, then there exists a sufficient small neighborhood $G \subset \mathbb{R}^l$ of g_0, and if we change it to $r \times m$ rational function matrix according to the above rule for any $g \in G$, then the rational function matrix is also strictly positive real.

If there exists an L_0 such that $(H, A - L_0C, G)$ is strictly positive real, then there must exist L in any neighborhood of L_0 such that the rank $A - LC$ can get the maximal rank and $(H, A - LC, G)$ is also strictly positive real. In what follows, we always keep in mind that the condition that $(H, A - LC, G)$ is strictly positive real and the condition that the rank of $A - LC$ is maximal hold simultaneously. The necessary condition is given by the following theorem, and it is easy to be verified.

Theorem 5.3.3 The necessary condition for the fact that there exits L such $(H, A - LC, G)$ is strictly positive real is that $\left(\begin{bmatrix} H \\ C \end{bmatrix}, A, G\right)$ has and only has $n - q$ stable zeros.

Proof Since $(H, A - LC, G)$ is strictly positive real, it is minimum phase, i.e., $(H, A - LC, G)$ has and only has $n - q$ stable zeros. Thus, there exist $n - q$ complex number $s_1, s_2, \ldots, s_{n-q}$ with negative real part such that

$$\operatorname{rank}\begin{bmatrix} s_jI - A + LC & G \\ H & 0 \end{bmatrix} < n + q, j = 1, 2, \ldots, n - q.$$

We also assume that the ranks of $\begin{bmatrix} s_jI - A + LC & G \\ H & 0 \end{bmatrix}$ are maximal for all s_j. Thus, by Lemma 5.3.3, rank $\begin{bmatrix} s_jI - A & G \\ H & 0 \\ C & 0 \end{bmatrix} < n + q$. □

The conditions of Theorem 5.3.3 can be solved by standard program; thus, it is easy to be verified. The following corollary may be more simple.

Corollary 5.3.1 If $(H, A - LC, G)$ is strictly positive real, then $\begin{bmatrix} H \\ C \end{bmatrix}$ is linear dependent with respect to column. □

Corollary 5.3.1 is a direct result of Theorem 5.3.3.

5.3.3 State Feedback Based on Observer

In this subsection, we continue to consider the stabilization problem for the Luré differential inclusion system, and the model is described in Sect. 5.2, i.e., $B = G,\ \ H = C$ in Inclusion (5.3.1).

The approach is presented in Fig. 5.5, and the stabilization is realized by state feedback. It is also pointed out that the output of observer is used for feedback since the state is not available. Here, we will give a detail discussion. Due to the limitation of space, the single-variable system is mainly studied. The readers can try to extend the result to multivariable systems, and it is a natural extension.

For the system whose state is not available, if the input u is available of the linear part in Fig. 5.5, the condition that (c^T, A) is observable guarantees the existence of the asymptotical observer.

$$\dot{\widehat{x}} = \left(A - lc^T\right)\widehat{x} + bu + ly,$$

where l is a vector such that $A - lc^T$ is a Hurwitz matrix, and $x - \widehat{x} \to 0\ (t \to \infty)$. Replacing x by $\widehat{x}$, i.e., $u = kc_e^T\widehat{x}$, the closed-loop system can be obtained (see Fig. 5.6).

$$\begin{aligned}
&\begin{bmatrix} \dot{x} \\ \dot{e} \end{bmatrix} = \begin{bmatrix} A + kbc_e^T & -kbc_e^T \\ 0 & A - lc^T \end{bmatrix}\begin{bmatrix} x \\ e \end{bmatrix} + \begin{bmatrix} b \\ 0 \end{bmatrix} u + \begin{bmatrix} b \\ 0 \end{bmatrix} \omega, \\
&\omega \in \nu(x), \\
&y = c^{\mathrm{T}}x,
\end{aligned}$$

where $e = x - \widehat{x}$ is the error. Since e is not controllable, it is easy to verify that the transfer function from u to y is

$$c^T\left(sI - A - kbc_e^T\right)^{-1}b,$$

and it is strictly positive real.

The stabilization approach based observer is given in Fig. 5.11. Recall the stabilization condition in the former subsection: If the system (c^T, A, b) is minimum phase and well-posed, the set-valued mapping $\nu(\cdot)$ is monotone and the input u of linear part can be obtained, then the stabilization for the differential inclusion system can be achieved by Luenberger observer and linear feedback. From Fig. 5.11, we can see that c_e^T and k are in the series form; thus, they can be combined into one block.

If the input u of linear part is not available, then the observer contains the set-valued mapping; its form is given in Inc. (5.3.2), i.e.,

$$\begin{aligned}
&\dot{\widehat{x}} = \left(A - lc^T\right)\widehat{x} + br - b\widehat{\omega} + ly, \\
&\widehat{\omega} \in \nu(x).
\end{aligned}$$

The whole system becomes

$$\begin{aligned}
&\begin{bmatrix} \dot{x} \\ \dot{e} \end{bmatrix} = \begin{bmatrix} A + kbc_e^T & -kbc_e^T \\ 0 & A - lc^T \end{bmatrix}\begin{bmatrix} x \\ e \end{bmatrix} + \begin{bmatrix} b \\ 0 \end{bmatrix} r + \begin{bmatrix} b & 0 \\ b & -b \end{bmatrix}\begin{bmatrix} \omega \\ \widehat{\omega} \end{bmatrix}, \\
&\omega \in \nu(x), \quad \widehat{\omega} \in \nu\left(\widehat{x}\right), \\
&y = c^T x.
\end{aligned}$$

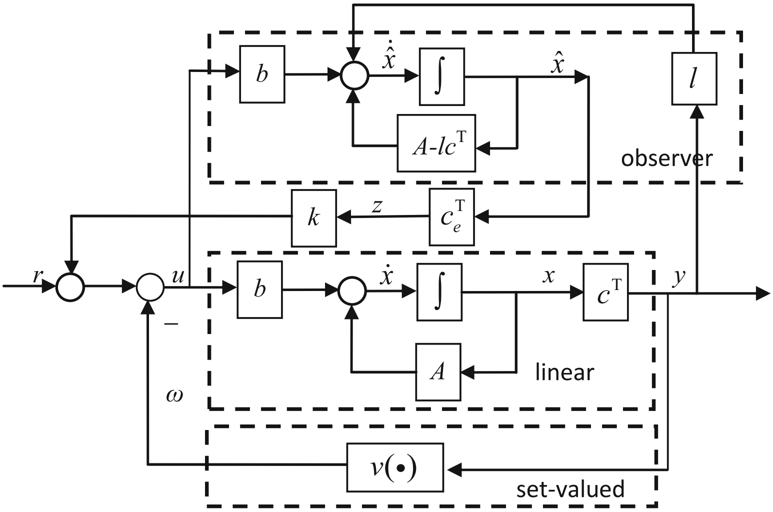


Fig. 5.11 Stabilization design-based observer

There exists k such that the system is stable. The discussion can be concluded as the following theorem.

Theorem 5.3.4 Consider the Luré differential inclusion system Inc. (5.2.2). If the set-valued mapping $v(\cdot)$ is monotone, the system is well posed, and the linear part (c^{T}, A, b) is minimum phase, then there always exists asymptotical observer Inc. (5.3.2) and linear feedback based on the observer, such that the closed-loop system is absolutely stable. □

The conditions of Theorem 5.3.4 are that the zeroes of linear part are stable and the set-valued mapping is monotone. The condition is very relaxed for the Luré differential inclusion systems.

By the similar treatment as that in the single-variable system, the output feedback matrix $\Lambda(s)$ can be replaced by constant matrix when the state can be obtained. Furthermore, if the input u of linear part is available, there always exists asymptotical observer, and the feedback-based observer can make the linear part strictly positive real, then the closed-loop Luré differential inclusion system is absolutely stable. The conclusion is very similar to that of stabilization of multivariable system.

By the discussion in this subsection, it is also illustrated that the separation principle for the Luré differential inclusion system is also valid. Thus, the observer and feedback can be designed separately, and stabilization can be realized by the combination of them.

5.3.4 Reduce-Order Luenberger Observer

In the theory of linear system, reduced-order observer is a specific topic. Let us consider the Luré differential inclusion system Inc. (5.3.1). The output $y = Cx$ can be written as $y = \begin{bmatrix} I \; 0 \end{bmatrix} \; x$ by an appropriate coordinate transformation, where I is the $r \times r$ identity matrix; thus, y is the former r dimension component of the state, and we need not design observer to reconstruct this r dimension substate but only design an $(n-r)$ order observer to reconstruct the left $(n-r)$-dimension substate. We call this kind of observer as reduced-order observer since its order is $(n-r)$. The main research topic of reduced-order observer is that if there exists a full-order observer, whether or not the reduced-order observer also exists. The fundamental conclusion is given as Theorem 5.3.5.

Consider the Luré differential inclusion system Inc. (5.3.1), there always exists coordinate transformation such that the resulting system has the following form.

$$\begin{aligned} &\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} \; A_{12} \\ A_{21} \; A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} r - \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \omega, \\ &\omega \in \nu(Hx) = \nu\left(H_1 x_1 + H_2 x_2\right), \\ &y = \begin{bmatrix} I_r \; 0 \end{bmatrix} x = x_1. \end{aligned} \tag{5.3.10}$$

Note that by the coordinate transformation, the state x and matrices A, B, G, H, etc., are all changed; hence, they should be written by different symbols, e.g. $\overline{x}$, $\overline{A}, \overline{B}, \overline{G}, \overline{H}$, etc. However, for the sake of simplicity, we still apply these original symbols, i.e., x and A, B, G, H, etc. The readers should keep in mind that the specific matrices are changed although the symbols are unchanged.

Theorem 5.3.5 Consider the Luré differential inclusion System (5.3.1). If the conditions of Theorem 5.3.1 all hold, then a reduced-order observer exists.

Proof If the conditions of Theorem 5.3.1 are all satisfied, then an Luenberger observer exists, it is equivalent to exist a matrix L such that $(H, A-LC, G)$ is strictly positive real. Hence, there exists a positive definite matrix P such that

$$\begin{aligned} &P\left(A-LC\right) + \left(A-LC\right)^T P < 0, \\ &PG = H^T. \end{aligned}$$

According to the decomposition done in Inc. (5.3.10), we obtain from the above expressions that

$$\begin{aligned} &\begin{bmatrix} * & * \\ * & P_{12}A_{12} + P_{22}A_{22} + A_{12}^T P_{12}^T + A_{22}^T P_{22} \end{bmatrix} < 0, \\ &\begin{bmatrix} P_{11}G_1 + P_{12}G_2 \\ P_{12}^{\mathrm{T}} G_1 + P_{22}G_2 \end{bmatrix} = \begin{bmatrix} H_1^T \\ H_2^T \end{bmatrix}, \end{aligned}$$

where $*$ stands for some determined elements which do not affect the result of computation. Thus, $P_{12}A_{12} + P_{22}A_{22} + A_{12}^{\mathrm{T}} P_{12}^{\mathrm{T}} + A_{22}^{\mathrm{T}} P_{22} < 0$ and $P_{12}^{\mathrm{T}} G_1 + P_{22}G_2 = H_2^{\mathrm{T}}$.

Let $K = -P_{22}^{-1} P_{12}$. Then

$$\begin{aligned} & P_{22}\left(A_{22}-KA_{12}\right)+\left(A_{22}-KA_{12}\right)^TP_{22} \\ & \quad = P_{22}A_{22}+P_{12}A_{12}+A_{22}^TP_{22}+A_{12}^TP_{12}^T \\ & \quad < 0, \end{aligned} \tag{5.3.11}$$

and

$$P_{22}\left(G_2-KG_1\right)=H_2^T. \tag{5.3.12}$$

Construct a system

$$\begin{aligned} & \dot{\widehat{z}}=\left(A_{22}-KA_{12}\right)\widehat{z}+\left(G_2-KG_1\right)\widehat{\omega}+\left(G_2-KG_1\right)r+\left[\left(A_{21}-KA_{11}\right)+\left(A_{22}-KA_{12}\right)K\right]y, \\ & \quad \widehat{\omega}\in\nu\left(H_2\widehat{z}+\left(H_1-H_2K\right)y\right), \\ & \widehat{x}_2=\widehat{z}+Ky, \end{aligned} \tag{5.3.13}$$

where r and y are the inputs, $\widehat{x}_2$ is the output. We now verify Inc. (5.3.13) is a reduced-order observer, i.e., to show that $\widehat{x}_2$ converges to x_2 asymptotically.

By (5.3.10), we obtain that

$$\dot{x}_2-K\dot{y}=\left(A_{22}-KA_{12}\right)x_1+\left(G_2-KG_1\right)\omega+\left(G_2-KG_1\right)r+\left(A_{21}-KA_{11}\right)y. \tag{5.3.14}$$

Denote that $z=x_2-Kx_1=x_2-Ky$ and $e=z-\widehat{z}$. Subtracting Eq. (5.3.14) from Inc. (5.3.13) yields

$$\begin{aligned} \dot{e} &= \left(A_{22}-KA_{12}\right)e+\left(G_2-KG_1\right)\left(\omega-\widehat{\omega}\right), \\ \omega &\in \nu\left(\left(H_1+H_2K\right)y+H_2z\right), \\ \widehat{\omega} &\in \nu\left(\left(H_1+H_2K\right)y+H_2\widehat{z}\right) . \end{aligned} \tag{5.3.15}$$

Following the treatment did in the proof of Theorem 5.3.1, let $\mu=\omega-\widehat{\omega}$ and $\zeta=H_2e$, then Inc. (5.3.15) becomes

$$\begin{aligned} & \dot{e}=\left(A_{22}-KA_2\right)e+\left(G_2-KG_1\right)\mu, \\ & \zeta=H_2e, \\ & \mu\in\nu\left(\left(H_1+H_2K\right)y+H_2z\right)-\nu\left(\left(H_1+H_2K\right)y+H_2\left(z-e\right)\right) . \end{aligned}$$

Inequalities (5.3.11) and (5.3.12) mean that $\left(H_2,\left(A_{22}-KA_{12}\right),\left(G_2-KG_1\right)\right)$ is strictly positive real, and the monotonicity of $\nu(\cdot)$ implies that the input and output of set-valued mappings satisfy the Popov inequality; thus, $e\to 0$, i.e., $\widehat{x}_2\to x_2$. □

Since the order of Inc. (5.3.13) is $n-r<n$, it is a reduced-order observer.

Problems

1. Let μ and ν denote the positive and negative inertial indices of the positive definite matrix P, respectively, then $\mathbb{R}^n=P_+\oplus P_-$ is orthogonal decomposition

of linear space $\mathbb{R}^n$, where $\dim P_+ = \mu$; when $x \in P_+,\ x \neq 0$, it holds that $x^T P x > 0$; and when $\dim P_- = \nu$ and $x \in P_-,\ x \neq 0$, it holds $x^T P x < 0$. Please prove that $V_\perp^T P V_\perp < 0$ if and only if $x = x_+ + x_-, x_+ \in P_+, x_- \in P_-, x_+^T P x_+ + x_-^T P x_- < 0$ for any $x \in \operatorname{span} V_\perp$.

2. Please prove that when $CB > 0$, the general solution of the equation $B^T P = C$ has the following form:

$$P = C^T (CB)^{-1} C + B_\perp X B_\perp^T,$$

where $X \in \mathbb{R}^{(n-m)\times(n-m)}$ and $B_\perp X B_\perp^T$ are positive definite.

3. Consider the Luré differential inclusion system

$$\begin{aligned}
\dot{x} &= \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix} x + \begin{bmatrix} 0 \\ -1 \end{bmatrix} r - \begin{bmatrix} 1 \\ -1 \end{bmatrix} \omega, \\
\omega &= \begin{cases} t+1, & t \in (0, \infty), \\ [-1, 1], & t = 0, \\ -t-1, & t \in (-\infty, 0), \end{cases} \\
y &= \begin{bmatrix} 1 & 0 \end{bmatrix} x.
\end{aligned}$$

Please design a full-order and a reduced-order Luenberger observers for it, respectively.

4. Consider the inequality

$$C_\perp \left[\left(H^T (HG)^{-1} H + G_\perp X G_\perp^T \right) A + A^T \left(H^T (HG)^{-1} H + G_\perp X G_\perp^T \right) \right] C_\perp^T < 0.$$

If the conditions of Theorem 5.3.2 all hold. By Applying Eq. (5.2.4) to (H, A, G), try to derive the simple form and conditions for the above inequality.

5. If $C_\perp = \begin{bmatrix} C_2 & 0 \end{bmatrix}$, where C_2 is invertible matrix. Accordingly, A can be decomposed as:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

then there exists a positive definite matrix P such that $C_\perp \left(PA + A^T P \right) C_\perp^T < 0$ if and only if $(A_{21},\ A_{22})$ is observable.

6. If the observer is designed as

$$\begin{aligned}
\dot{\widehat{x}}(t) &= (A - LC)\widehat{x}(t) + Br(t) - G\widehat{\omega}(t) + Ly(t), \\
\widehat{\omega}(t) &\in \nu \left(H\widehat{x}(t) + KC\widehat{x}(t) - Ky \right),
\end{aligned}$$

then how can Theorem 5.3.1 be modified? Comparing with observer Inc. (5.3.2), an extra matrix K. Thus, the freedom degree of design increases.

7. Please prove Theorem 5.3.4.

5.4 Linear Observers of Luré Differential Inclusion Systems

The Luenberger observer for the Luré differential inclusion system is designed in the former section. In the design of reduced-order observer, we have pointed out that the information of the states is contained in the output y, and the information of ω is also contained in y. Then, can ω be solved from y, and does the observer not contain the set-valued mapping? The topic will be discussed in this section, and we try to design a novel observer for the Luré differential inclusion system which does not contain set-valued function. Since this kind of observer is different from Luenberger observer, we call it linear observer in this book.

Consider the following Luré differential inclusion system described as follows,

$$\begin{aligned} \dot{x} &= Ax + G\omega + Bu, \\ \omega &\in \nu(x), \\ y &= Cx, \end{aligned} \tag{5.4.1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^r$ are state, input and output of the system, respectively, and $\omega \in \mathbb{R}^q$ is output of set-valued mapping, and A, B, G, C are given matrices with compatible dimensions. The condition of set-valued $\nu(x)$ is just required to guarantee that the system Inc. (5.4.1) is well posed. It should be noted that although $\nu(x)$ is set-valued mapping, when the system works for a determined moment, $\omega(t) \in v(x(t))$ is a determined selection, then the state $x(t)$ takes a determined value, so is the output $y(t)$. $y(t)$ has to contain some information of $\omega(t)$. The fact will be applied to design observer.

We start with a lemma which will be used frequently in this section.

Lemma 5.4.1 Consider the following linear system which holds a feedback of derivative of output.

$$\begin{aligned} \dot{x} &= Mx + N\dot{y} + Bu, \\ y &= Cx, \end{aligned} \tag{5.4.2}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^r$, M, N, B, C are the given matrices with compatible dimensions. If (C, M) is detectable, then there always exists an asymptotically linear observer for the system Eq. (5.4.2).

Proof Consider the following system,

$$\dot{\widehat{x}} = M\widehat{x} + N\dot{y} + Bu + L\left(y - C\widehat{x}\right), \tag{5.4.3}$$

where M, N, B, C are the same as those in Eq. (5.4.2), and the observer gain L will be designed later. Eq. (5.4.3) is exactly the Luenberger observer discussed in the last section.

Let $\zeta = \widehat{x} - Ny$. Then Eq. (5.4.3) becomes

$$\begin{aligned}\dot{\zeta} &= M(\zeta + Ny) + Bu + Ly - LC(\zeta + Ny)\\ &= (M - LC)\zeta + (MN + L - LCN)y + Bu,\end{aligned} \tag{5.4.4}$$

$$\widehat{x} = \zeta + Ny.$$

It is noted that $\dot{y}$ does not appear in the right-hand side of Eq. (5.4.4). Since (C, M) is detectable, there exists an L such that $M - LC$ is a Hurwitz matrix. Now we prove that Eq. (5.4.4) is an observer of the system described by Eq. (5.4.2).

Since Eq. (5.4.4) is exact the same as Eq. (5.4.3), subtracting Eq. (5.4.3) from Eq. (5.4.2) yields $\dot{e} = (M - LC)e$, where $e = x - \widehat{x}$. $M - LC$ is a Hurwitz matrix, we obtain that $e(t) \to 0$ $(t \to \infty)$ for any initial condition e_0. □

If the first equation in Eq. (5.4.2) is written as $(I - NC)\dot{x} = Mx + Bu$, then when $(I - NC)$ is invertible and it becomes $\dot{x} = (I - NC)^{-1}Mx + (I - NC)^{-1}Bu$, a Luenberger observer can be designed. However, the approach of Lemma 5.4.1 does not apply such a design thought; it is novel for the observer design. It is also stated in Lemma 5.4.1 that if $N\dot{y}$ does not appear in the state equation, then Ny also does not appear in the output equation of linear observer. The observer Eq. (5.4.3) reduces to the Luenberger observer.

5.4.1 Single-Variable Systems

We start with single-variable system. When $m = r = q = 1$, the system Eq. (5.4.1) becomes

$$\begin{aligned}&\dot{x} = Ax + g\omega + bu,\\ &\omega \in \nu(x),\\ &y = c^{\mathrm{T}}x.\end{aligned} \tag{5.4.5}$$

Assume that the system (c^{T}, A, g) is both controllable and observable and is minimum phase with relative degree one. Then, there always exists proper coordinate transformation such that (c^{T}, A, g) has the following form

$$c^{T} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \end{bmatrix}, A = \begin{bmatrix} -a_1 & 1 & \cdots & 0 & 0\\ -a_2 & 0 & \cdots & 0 & 0\\ \vdots & \vdots & \ldots & \vdots & \vdots\\ -a_{n-1} & 0 & \cdots & 0 & 1\\ -a_n & 0 & \cdots & 0 & 0 \end{bmatrix}, \ g = \begin{bmatrix} g_1\\ g_2\\ \vdots\\ g_{n-1}\\ g_n \end{bmatrix}, \tag{5.4.6}$$

where $g \in \mathbb{R}^n$ is a Hurwitz vector with $g_1 \neq 0$.

Taking the derivative of both sides of the third equation in Eq. (5.4.5), we have

$$\dot{y} = c^T\dot{x} = c^T\,(Ax + g\omega + bu) = c^TAx + c^Tg\omega + c^Tbu.$$

Solving out ω from the equation, we obtain

$$\omega = \left(c^Tg\right)^{-1}\left(\dot{y} - c^TAx - c^Tbu\right) = -\left(c^Tg\right)^{-1}c^TAx - \left(c^Tg\right)^{-1}c^Tbu + \left(c^Tg\right)^{-1}\dot{y}.$$

Substituting above equation into Eq. (5.4.5) yields

$$\begin{gathered}\dot{x} = \left(A - g\left(c^Tg\right)^{-1}c^TA\right)x + g\left(c^Tg\right)^{-1}\dot{y} + \left(I - g\left(c^Tg\right)^{-1}c^T\right)bu,\\ y = c^Tx.\end{gathered} \tag{5.4.7}$$

The form of (5.4.7) is the same as Eq. (5.4.2) of Lemma 5.4.1, and it is a kind of certain system without set-valued mapping.

Theorem 5.4.1 If the linear system (c^T, A, g) is both controllable and observable and the system is minimum phase with relative degree one, then there always exists an asymptotically linear observer for the Luré differential inclusion system Inc. (5.4.5).

Proof By Lemma 5.4.1, we only need prove that $\left(c^{\mathrm{T}}, A - g\left(c^{\mathrm{T}}g\right)^{-1}c^{\mathrm{T}}A\right)$ is detectable. By Eq. (5.4.6), we can obtain that

$$A - g\left(c^Tg\right)^{-1}c^TA = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ -a_2 + \frac{g_2}{g_1}a_1 & -\frac{g_2}{g_1} & 1 & \cdots & 0 \\ -a_3 + \frac{g_3}{g_1}a_1 & -\frac{g_3}{g_1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ -a_n + \frac{g_n}{g_1}a_1 & -\frac{g_n}{g_1} & 0 & \cdots & 0 \end{bmatrix},$$

where the first row of the matrix is zero. Thus, the unobservable dynamic matrix is

$$\begin{bmatrix} -\frac{g_2}{g_1} & 1 & \cdots & 0 \\ -\frac{g_3}{g_1} & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ -\frac{g_n}{g_1} & 0 & \cdots & 0 \end{bmatrix}.$$

It is an $(n-1)\times(n-1)$ matrix, and its corresponding eigenpolynomial is $s^{n-1} + (g_2/g_1)\,s^{n-2} + \cdots + (g_{n-1}/g_1)\,s + (g_n/g_1)$. Since g is a Hurwitz vector, the polynomial is stable, i.e., $\left(c^T, A - g\left(c^Tg\right)^{-1}c^TA\right)$ is detectable. □

For the sake of simplicity, we denote that $M = I - g\left(c^T g\right)^{-1} c^T$, then Eq. (5.4.7) can be written as

$$\dot{x} = MAx + g\left(c^T g\right)^{-1} \dot{y} + Mbu.$$

By Lemma 5.4.1, the asymptotical linear observer is designed as

$$\begin{aligned} \dot{\zeta} &= \left(MA - lc^T\right)\zeta + MAg\left(c^T g\right)^{-1} y + Mbu, \\ \widehat{x} &= \zeta + g\left(c^T g\right)^{-1} y, \end{aligned} \tag{5.4.8}$$

where $l \in \mathbb{R}^n$ and is substitute of L mentioned in Lemma 5.4.1, and $\left(MA - lc^T\right)$ is a Hurwitz matrix. By the denotation of Lemma 5.4.1, we have $CN = 1$.

It is interesting to compare Theorem 5.4.1 with Theorem 5.3.2. In Theorem 5.3.2, if G is replaced by g, and H is by c^{T}, and $\left(c^T, A - lc^T, g\right)$ is positive real, then (c^T, A, g) is minimum phase with $c^T g \neq 0$. It is indeed that the conditions of Theorem 5.3.2 imply those of Theorem 5.4.1. Thus, the asymptotical linear observer exists under the conditions of Theorem 5.3.2. Therefore, Theorem 5.4.1 presents a stronger conclusion than that established in Sect. 5.3.

By Eq. (5.4.3), Eq. (5.4.8) is derived from

$$\dot{\widehat{x}} = \left(MA - lc^T\right)\widehat{x} + g\left(c^T g\right)^{-1} \dot{y} + ly + bu.$$

The transfer function from y to $\widehat{x}$ is

$$\frac{\widehat{x}(s)}{y(s)} = \left(sI - MA + lc^T\right)^{-1} \left(g\left(c^T g\right)^{-1} s + l\right). \tag{5.4.9}$$

System (5.4.8) is a realization of Eq. (5.4.9). It can be seen the right side of Eq. (5.4.9) is a proper rational matrix, but not strictly proper.

The effectiveness of the linear observer is shown in the following example.

Example 5.4.1 Consider the following Luré differential inclusion system

$$\begin{aligned} \dot{x}_1 &= x_2 + \omega, \\ \dot{x}_2 &= -0.6x_1 + 2\omega, \\ \omega &\in \nu(x), \\ y &= x_1, \end{aligned}$$

where the set-valued mapping $\nu(\cdot)$ is

$$\nu(x) = \begin{cases} -1 & x_1 + x_2 < -2, \\ \left[-1, x_1 + x_2 + 1\right], & -2 \le x_1 + x_2 < 0, \\ \left[x_1 + x_2 - 1, 1\right], & 0 \le x_1 + x_2 < 2, \\ 1, & 2 \le x_1 + x_2. \end{cases}$$

Two selections are considered as follows:

$$\omega_1(t) = \begin{cases} -1, & x_1 + x_2 < -2, \\ 0.5\,(x_1 + x_2)\,, & -2 \le x_1 + x_2 < 0, \\ 0.5\,(x_1 + x_2)\,, & 0 \le x_1 + x_2 < 2, \\ 1, & 2 \le x_1 + x_2, \end{cases}$$

and

$$\omega_2(t) = \begin{cases} -1, & x_1 + x_2 < -2, \\ -1, & -2 \le x_1 + x_2 < 0, \\ x_1 + x_2 - 1, & 0 \le x_1 + x_2 < 2, \\ 1, & 2 \le x_1 + x_2. \end{cases}$$

The selections are different obviously. The trajectories of the system for the same initial conditions when $\omega = \omega_1(t)$ and $\omega = \omega_2(t)$ are shown in Fig. 5.12, respectively. (Due to limitation of space, the trajectories of $x_1(t)$ are only given). It is indeed that the observer with set-valued mapping cannot work.

We employ the design method of linear observer in Theorem 5.4.1. It is easy to determine that

$$A = \begin{bmatrix} 0 & 1 \\ -0.6 & 0 \end{bmatrix}, \ b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ c^T = \begin{bmatrix} 1 \ 0 \end{bmatrix}, \ g = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

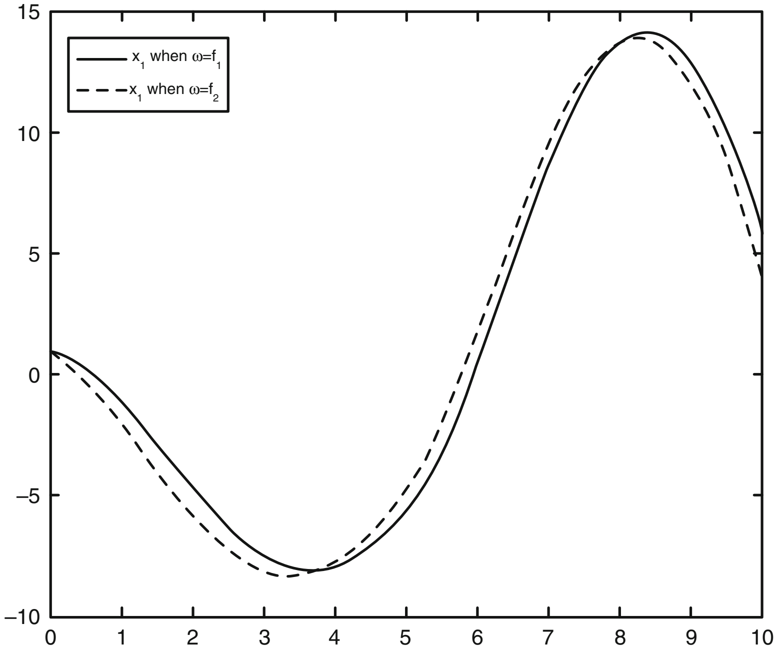


Fig. 5.12 Different trajectories of x_1 under different selections

Let

$$l = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \text{ then } MA - lc^T = \begin{bmatrix} -1 & 0 \\ -3.6 & -2 \end{bmatrix}.$$

The linear observer is designed as

$$\dot{\zeta} = \begin{bmatrix} -1 & 0 \\ -3.6 & -2 \end{bmatrix} \zeta + \begin{bmatrix} 0 \\ -4.6 \end{bmatrix} y,$$
$$\widehat{x} = \zeta + \begin{bmatrix} 1 \\ 2 \end{bmatrix} y\,.$$

The tracking performances are given in Figs. 5.13 and 5.14 for the selections $\omega_1(t)$ and $\omega_2(t)$. From the simulation result, the linear observer works well. □

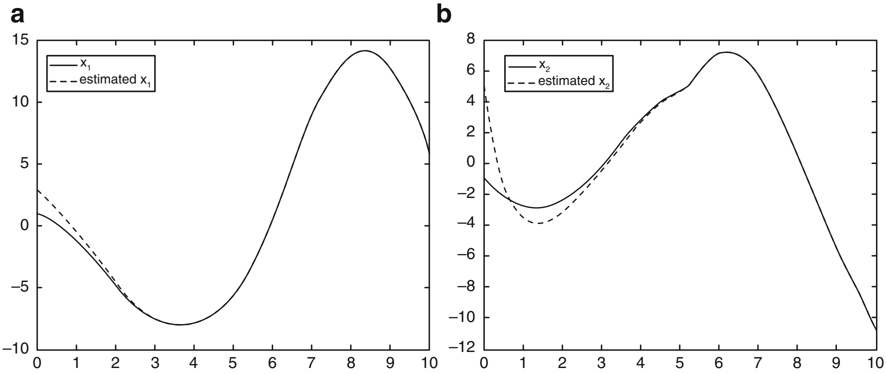


Fig. 5.13 The tracking performances of the linear observer when $\omega = \omega_1(t)$, (**a**) tracking of $x_1(t)$, (**b**) tracking of $x_2(t)$

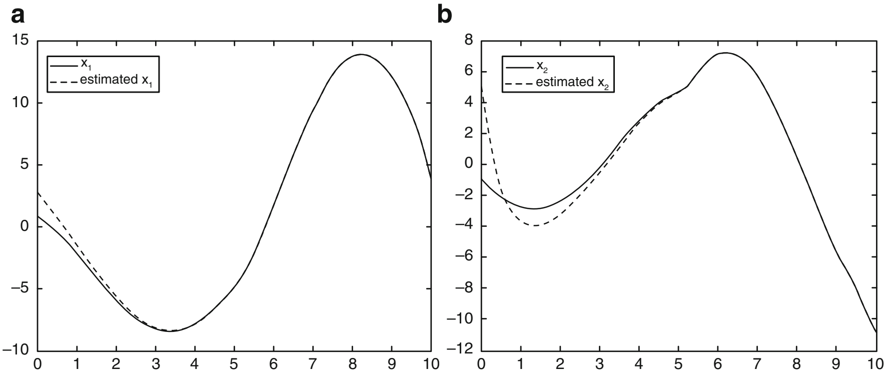


Fig. 5.14 The tracking performances of the linear observer when $\omega = \omega_2(t)$, (**a**) tracking of $x_1(t)$, (**b**) tracking of $x_2(t)$

Now, we turn to discuss the case where the relative degree of the system is larger than one. In order to keep simplicity, the single-variable system with relative degree two is only considered. However, the conclusion can be extended to multivariable case, directly.

We still consider the system Inc. (5.4.5), where (c^T, A, g) is both controllable and observable with minimum phase, and the relative degree is 2, i.e., $c^T g = 0,\ \ c^T Ag \neq 0$. The canonical form is similar to that of Eq. (5.4.6), but

$$c^T = \begin{bmatrix} 1\ 0 \cdots 0\ 0 \end{bmatrix}, A = \begin{bmatrix} -a_1 & 1 \cdots 0\ 0 \\ -a_2 & 0 \cdots 0\ 0 \\ \vdots & \vdots \ldots \vdots\ \vdots \\ -a_{n-1} & 0 \cdots 0\ 1 \\ -a_n & 0 \cdots 0\ 0 \end{bmatrix}, \ g = \begin{bmatrix} 0 \\ g_1 \\ \vdots \\ g_{n-2} \\ g_{n-1} \end{bmatrix}, \tag{5.4.10}$$

where g is a Hurwitz vector. By the third equation of Inc. (5.4.5), we get that $\dot{y} = c^T Ax + c^T bu$, and $\ddot{y} = c^T A^2 x + c^T Ag\omega + c^T b\dot{u} + c^T Abu$. Solving ω from the last equation yields

$$\omega = \left(c^T Ag\right)^{-1} \left(\ddot{y} - c^T A^2 x - c^T b\dot{u} - c^T Abu\right).$$

Substituting the above expression into the first equation of (5.4.5) results in

$$\dot{x} = Ax + g\left(c^T Ag\right)^{-1} \left(\ddot{y} - c^T A^2 x\right) + bu - g\left(c^T Ag\right)^{-1} b\dot{u} - g\left(c^T Ag\right)^{-1} c^T Abu.$$

Denote that $M = I - g\left(c^T Ag\right)^{-1} c^T A,\ \ n(u) = bu - g\left(c^T Ag\right)^{-1} c^T Abu - g\left(c^T Ag\right)^{-1} b\dot{u}$, the above equation can be simplified as

$$\dot{x} = MAx + g\left(c^T Ag\right)^{-1} \ddot{y} + n(u).$$

The observer is designed as

$$\dot{\widehat{x}} = MA\widehat{x} + g\left(c^T Ag\right)^{-1} \ddot{y} + n(u) + l\left(y - c^T \widehat{x}\right). \tag{5.4.11}$$

Let $e = x - \widehat{x}$, then

$$\dot{e} = \left(MA - lc^T\right) e,$$

If $\left(MA - lc^T\right)$ is a Hurwitz matrix, i.e., (c^T, MA) is detectable, then $e \to 0\ \ (t \to \infty)$.

When the form of (c^T, A, g) is that given in Eq. (5.4.10), it is direct to compute that

$$M = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_1 & 0 & 0 & \cdots & 0 \\ \dfrac{a_1 g_3}{g_2} & -\dfrac{g_3}{g_2} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \dfrac{a_1 g_n}{g_2} & -\dfrac{g_n}{g_2} & 0 & \cdots & 1 \end{bmatrix}, \quad MA = \left[\begin{array}{cc|ccc} -a_1 & 1 & 0 & \cdots & 0 \\ -a_1^2 & a_1 & 0 & \cdots & 0 \\ \hline * & * & -\dfrac{g_3}{g_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ * & * & -\dfrac{g_n}{g_2} & \cdots & 0 \end{array}\right]$$

In MA, the detailed numbers of the left lower block need not be determined, and the right lower block is

$$\begin{bmatrix} -\frac{g_3}{g_2} & 1 \cdots 0 \\ \vdots & \vdots \cdots \vdots \\ -\frac{g_{n-1}}{g_2} & 0 \cdots 1 \\ -\frac{g_n}{g_2} & 0 \cdots 0 \end{bmatrix}.$$

It is a Hurwitz matrix, and it is the unobservable part corresponding to $c^T = \begin{bmatrix} 1 \ 0 \cdots 0 \end{bmatrix}$; thus, (c^T, MA) is detectable.

From Eq. (5.4.11), the transfer function from y to $\widehat{x}$ is

$$\frac{\widehat{x}(s)}{y(s)} = \left(sI - MA + lc^T\right)^{-1} \left(g\left(c^T Ag\right)^{-1} s^2 + l\right).$$

It is not usually a proper rational matrix; thus, there does not exist proper realization for it. Let $\xi = \widehat{x} - g\left(c^T Ag\right)^{-1} \dot{y} + g\left(c^T Ag\right)^{-1} bu$, then it follows from Eq. (5.4.11) that

$$\begin{aligned} \dot{\xi} &= \left(MA - lc^T\right)\xi + MAg\left(c^T Ag\right)^{-1}\dot{y} + ly + Mbu, \\ \widehat{x} &= \xi + g\left(c^T Ag\right)^{-1}\dot{y} - g\left(c^T Ag\right)^{-1} bu. \end{aligned}$$

The above system is a linear observer for the system Inc. (5.4.5) with relative degree two. It should be noted that the input of observer contains $\dot{y}$, i.e., differentiator. In some literature, when y and $\dot{y}$ are both bounded, $\dot{y}$ can be constructed by high gain feedback instead of differentiator. The contents are beyond the scope of this book. The readers are referred to Ly et al. (1994) and Huang et al. (1999).

By the same method, we can studied the case where the relative degree is larger than two, and we also need larger order derivatives of y.

5.4.2 Multivariable Systems

We will extend the result of the former subsection to the multivariable system. System Inc. (5.4.1) is rewritten as:

$$\begin{aligned}&\dot{x} = Ax + G\omega + Bu,\\ &\omega \in \nu(x),\\ &y = Cx,\end{aligned}$$

and holds that $r = q$.

Theorem 5.4.2 If the linear system (C, A, G) is both controllable and observable, and the system is minimum phase with relative degree one, then there always exists asymptotically linear observer for the Luré differential inclusion system Inc. (5.4.1).

Proof By the result of Sect. 5.2, since $r = q$, the relative degree one means that CG is invertible. Taking the derivative of both sides of the third equation of Inc. (5.4.1), we obtain that $\dot{y} = CAx + CG\omega + CBu$. Solving ω results in

$$\omega = (CG)^{-1}\left(\dot{y} - CAx - CBu\right).$$

Substituting the above equation into the first equation of Inc. (5.4.1) yields

$$\begin{aligned}\dot{x} &= Ax + G(CG)^{-1}\left(\dot{y} - CAx - CBu\right) + Bu\\ &= \left[A - G(CG)^{-1}CA\right]x + G(CG)^{-1}\dot{y} + \left[B - G(CG)^{-1}CB\right]u\\ &= MAx + G(CG)^{-1}\dot{y} + MBu,\end{aligned}$$

where $M = I - G(CG)^{-1}C$. By Lemma 5.4.1, if (C, MA) is detectable, then there exists a linear observer. Hence, the key of the proof is to show the observability of (C, MA).

By the discussion of Sect. 5.2, the fact that CG is invertible implies that (C, A, G) has the following forms

$$C = [C_1\ \ 0],\ \ A = \begin{bmatrix} A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix},\ \ G = \begin{bmatrix} I\\ 0\end{bmatrix},$$

where I is the $m \times m$ identity matrix, and C_1 is invertible. Moreover, the fact that (C, A, G) is minimum phase means that A_{22} is a Hurwitz matrix. It is direct to verify that

$$M = I - G(CG)^{-1}C = \begin{bmatrix} 0 & 0\\ 0 & I\end{bmatrix},\quad MA = \begin{bmatrix} 0 & 0\\ 0 & I\end{bmatrix}\begin{bmatrix} A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix} = \begin{bmatrix} 0 & 0\\ A_{21} & A_{22}\end{bmatrix}.$$

A_{22} is the dynamic matrix which is unobservable part of (C, MA). It is a Hurwitz matrix; hence, (C, MA) is detectable. The proof is completed. □

In fact, the assumption of controllability in Theorem 5.4.2 is not essential.

By a similar discussion for single-variable system, the linear observer for the multivariable system with relative degree two can also be designed, and the derivative of y is contained in the observer. It is left to the readers.

To end this section, we will discuss the design of reduced-order observer. If there exists a full-order linear observer for the system, then there also exists a reduced-order linear observer. We now prove the conclusion.

Corollary 5.4.1 If the linear system in Inc. (5.4.1) is both controllable and observable, and the system is minimum phase with relative degree one, then there always exists a linear observer with order $n - r$ for the Luré differential inclusion system Inc. (5.4.1).

Proof The form of Inc. (5.4.1) can be written as

$$\begin{aligned} &\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} G_1 \\ 0 \end{bmatrix} \omega + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u, \\ &y = x_1. \end{aligned}$$

then the second component of state satisfies

$$\dot{x}_2 = A_{21}x_1 + A_{22}x_2 + B_2u = A_{21}y + A_{22}x_2 + B_2u.$$

In view of that the system is minimum phase, since A_{22} is a Hurwitz matrix. The linear observer for x_2 can be designed as

$$\dot{\widehat{x}}_2 = A_{21}y + A_{22}\widehat{x}_2 + B_2u.$$

Let $e_2 = x_2 - \widehat{x}_2$, then $\dot{e}_2 = A_{22}e_2$, i.e., $e_2 \to 0\ (t \to \infty)$. □

In the above proof, the observability is not needed and the condition of minimum phase guarantees that it is detectable. From the proof of Corollary 5.4.1, the design of reduced-order observer is more simple than that of full-order one; it is not necessary to solve the term ω. The reader can try studying the case where the relative degree of the system is larger than one.

Problems

1. In the case where the relative degree is two, if the observer is designed as

$$\dot{\widehat{x}} = MA\widehat{x} + g\left(c^TAg\right)^{-1}\ddot{y} + n(u) + l\left(y - c^T\widehat{x}\right) + k\left(\dot{y} - c^T\dot{\widehat{x}}\right).$$

Please give the condition under which $e = x - \widehat{x}$ converges to zero asymptotically.

2. Consider the Luré differential inclusion System (5.4.1), where $r = q$. If $CG = 0,\ \det CAG \neq 0$, simulating the treatment did for single-variable system, please design a linear observer.
3. Let

$$A = \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix},\ b = \begin{bmatrix} 1 \\ -1 \end{bmatrix},\ c^T = \begin{bmatrix} 1 & 0 \end{bmatrix},\ g = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Please give the detail form of linear observer and reduced-order linear observer, respectively.

4. Extend Corollary 4.3.1 to multivariable system.

5.5 Adaptive Luenberger Observers

The last two sections extend the results of observer design to the Luré differential inclusion system with uncertain parameters; two adaptive observers, i.e., Luenberger observer and linear observer, will be presented for this system. In the theory of adaptive control, the plant considered is always subjected to uncertainty, which can be described by an unknown parameter or a slow variation parameter (may be a vector); the aim is to present a parameter estimation model and a control law with estimated parameters to achieve the required performance. The control is usually called adaptive law. In the adaptive control design, it is usual that the estimated parameters are not required to converge to the value of real parameters, and we only require the control law can achieve the required performance.

Consider the following Luré differential inclusion system with uncertain parameters

$$\begin{aligned} &\dot{x} = Ax + Bu - G\omega + f(x,u)\,\theta, \\ &\omega \in \nu(Hx), \\ &y = Cx, \end{aligned} \tag{5.5.1}$$

where $x \in \mathbb{R}^n,\ u \in \mathbb{R}^m,\ y \in \mathbb{R}^r$ are the state, input and output of the system, respectively, and $\omega \in \mathbb{R}^q$ is the output of set-valued mapping, and A, B, G, C, H are the determined matrices with compatible dimensions. The set-valued mapping $\nu(x)$ is required to guarantee that the system Inc. (5.5.1) is well-posed. $\theta \in \mathbb{R}^p$ is an uncertain parameter vector, and it is always supposed to be a constant vector or bounded and slow variable, i.e., $\|\theta\| \leq \gamma_1$, where $\gamma_1 > 0$ is a known constant; $f(x,u)$ is a given $n \times p$ function matrix; it is assumed that $f(x,u)$ is Lipschitzian with respect to x and is uniform to u, i.e., there exists $\gamma_2 > 0$ which is independent of u such that

$$\|f(x_1,u) - f(x_2,u)\| \leq \gamma_2 \|x_1 - x_2\|.$$

5.5.1 Adaptive Luenberger Observers

If there exists an asymptotical observer for the system Inc. (5.5.1), it implies that a state observer exists for the case of $\theta = 0$. Thus, by Theorem 5.3.1, the following requirements are essential for adaptive Luenberger observer.

- There exists an observer gain matrix L such that the linear system $(H,\ A - LC,\ G)$ is both controllable and observable and is also strictly positive real.
- $\nu(\cdot)$ is monotone.
- The solution of Inc. (5.3.4) exists, i.e., it is well posed.

By Theorem 5.3.2, the first one requires $HG > 0$, and H is of full row rank, and G is of full row rank. In what follows, we will study under which conditions the adaptive observer can be designed.

The norm is always used in this section; thus, let us recall some preliminary mentioned in Sect. 1.1. Two any norms are equivalent in a finite dimension space (Theorem 1.1.1). Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a linear mapping, then $\|Tx\|_2 \leq \|T\|_d \|x\|_1$ for any $x \in \mathbb{R}^n$, where $\|x\|_1$ is some kind of norm in $\mathbb{R}^n$, $\|Tx\|_2$ is some other norm in $\mathbb{R}^n$; $\|T\|_{\mathrm{r}(2,1)}$ is denoted the operator norm induced by $\|\cdot\|_1$ and $\|\cdot\|_2$. If $\|x\|_1$ and $\|Tx\|_2$ are taken as Euclidian norms, then $\|T\|_{\mathrm{r}(2,1)} = \sqrt{\lambda_{\max}\left(T^{\mathrm{T}}T\right)}$; this kind of norm is called spectral norm and denoted by $\|T\|_{\mathrm{P}}$. If $\|x\|_1$ or $\|Tx\|_2$ is taken as other kind of norm, then the induced norm $\|T\|_{\mathrm{P}}$ is equivalent to the former one. Let $\|T\|_c$ be other norm of $T \in \mathbb{R}^{n\times n}$,[5] it may be different from the induced norm $\|T\|_{\mathrm{r}(2,1)}$, but $\|Tx\|_2 \leq \|T\|_c \|x\|_1$ for every $x \in \mathbb{R}^n$, then $\|T\|_c$ is called the compatible norm for $\|\cdot\|_1$ and $\|\cdot\|_2$. If $\|x\|_1$ and $\|Tx\|_2$ are taken as Euclidian norms, then the Frobenius norm

$$\|T\|_{\mathrm{F}} = \sqrt{\sum_{i,j=1}^{n} t_{ij}^2}$$

is a compatible norm of the Euclidian norm.

Except we specially emphasize, the norm of vector x is taken as Euclidian norm, and the norm of operator T is taken as spectral norm or Frobenius norm.

Motivating by Sect. 5.3, consider the system described by Inc. (5.5.2)

$$\begin{aligned} \dot{\widehat{x}} &= (A - LC)\widehat{x} + Bu - G\widehat{\omega} + f\left(\widehat{x}, u\right)\widehat{\theta} + Ly, \\ \widehat{\omega} &\in \nu\left(H\widehat{x}\right), \\ \dot{\widehat{\theta}} &= h\left(\widehat{x}, u\right)\left(y - C\widehat{x}\right), \end{aligned} \tag{5.5.2}$$

where $h\left(\widehat{x}, u\right)$ is a $p \times r$ matrix of functions to be designed later. The last equation in Inc. (5.5.2) is the adaptive law. This section will discuss the conditions, under

[5] When we define the norm of matrix, in order to be consistent with the norm of induced operator, the condition $\|AB\| \leq \|A\|\,\|B\|$ that is always additional considered, but it is not necessary for a vector.

which (5.5.2) is an asymptotical observer for (5.5.1). In (5.5.2), $\widehat{\theta}$ is a time-varying parameter and is also called adaptive parameter, which is used to estimate the unknown parameter θ.

By Theorem 5.3.1, for the system Inc. (5.3.1) without $f(x,u)\theta$, the condition of existence of Luenberger observer is that there exists L such that $(H,\ A-LC,\ G)$ is strictly positive real, i.e., there exist positive definite matrices P and Q, such that

$$P\,(A-LC)+(A-LC)^{T}P=-Q, \quad \text{and} \quad G^{T}P=H. \tag{5.5.3}$$

The sufficient and necessary condition of existence of solution for Eq. (5.5.3) is presented in Lemma 5.3.2 and Theorem 5.3.2.

Let $\lambda_{\max}(P)$ and $\lambda_{\min}(Q)$ be the maximum and minimum eigenvalues of P and Q, respectively. Then we have the following theorem.

Theorem 5.5.1 If the following conditions hold.

(1) There exists an L such that $(H,\ A-LC,\ G)$ is controllable and observable and is also strictly positive real.
(2) $\nu(\cdot)$ is monotone.
(3) The system Inc. (5.3.4) is well posed.
(4) $\lambda_{\min}(Q) > 2\gamma_1\gamma_2\lambda_{\max}(P)$, where γ_1 and γ_2 are two constants defined at the beginning of this section.
(5) There exists a function matrix $h(x,u)$ such that $h\,(x,u)\,C=[Pf\,(x,u)]^{\mathrm{T}}$,

where P and Q satisfy Eq. (5.5.3). Then Inc. (5.5.2) is an adaptive asymptotical observer of Inc. (5.5.1).

Proof Subtracting Inc. (5.5.2) from Inc. (5.5.1) yields

$$\begin{aligned} &\dot{e}=(A-LC)\,e-G\,(\omega-\widehat{\omega})+f\,(x,u)\,\theta-f\,(\widehat{x},u)\,\widehat{\theta},\\ &\omega\in\nu(Hx),\\ &\widehat{\omega}\in\nu\left(H\widehat{x}\right), \end{aligned} \tag{5.5.4}$$

where $e=x-\widehat{x}$ is the observing error.

Let the following function be Lyapunov function candidate

$$V\left(e,\tilde{\theta}\right)=e^{T}Pe+\tilde{\theta}^{T}\tilde{\theta},$$

where $\tilde{\theta}=\theta-\widehat{\theta}$. Taking the derivative $V\left(e,\tilde{\theta}\right)$ along the trajectory of the system Inc. (5.5.4), we obtain

$$\begin{aligned} \dot{V}=\ &e^{T}(A-LC)^{T}Pe+e^{T}P\,(A-LC)\,e\\ &-(\omega-\widehat{\omega})^{T}G^{T}Pe-e^{T}PG\,(\omega-\widehat{\omega})\\ &+\theta^{T}f^{T}\,(x,u)\,Pe+e^{T}Pf\,(x,u)\,\theta\\ &-\widehat{\theta}^{T}f^{T}\,(\widehat{x},u)\,Pe-e^{T}Pf\,(\widehat{x},u)\,\widehat{\theta}\\ &-2\tilde{\theta}^{T}\dot{\widehat{\theta}}, \end{aligned}$$

where we have applied that $\dot{\theta} = 0$. We now analyze $\dot{V}$ term by term.

Firstly, by the positive realness lemma, we have

$$e^T (A - LC)^T Pe + e^T P (A - LC) e = -e^T Qe < 0, (e \neq 0) .$$

Secondly, since $G^T P = H$, it leads to

$$(\omega - \widehat{\omega})^T G^T Pe = (\omega - \widehat{\omega})^T He = (\omega - \widehat{\omega})^T (Hx - H\widehat{x}) .$$

By the monotonicity of $\nu(\cdot)$, we have

$$-(\omega - \widehat{\omega})^T G^T Pe \leq 0.$$

At last, we consider

$$\begin{aligned} & e^T Pf (x, u) \theta - e^T Pf (\widehat{x}, u) \widehat{\theta} \\ & \quad = e^T Pf (x, u) \theta - e^T Pf (\widehat{x}, u) \theta + e^T Pf (\widehat{x}, u) \theta - e^T Pf (\widehat{x}, u) \widehat{\theta} \\ & \quad = e^T P [f (x, u) - f (\widehat{x}, u)] \theta + e^T Pf (\widehat{x}, u) \tilde{\theta}. \end{aligned}$$

The first term holds that

$$\begin{aligned} & e^T P [f (x, u) - f (\widehat{x}, u)] \theta \\ & \quad \leq \left\| e^T P [f (x, u) - f (\widehat{x}, u)] \theta \right\| \\ & \quad \leq \gamma_1 \left\| e^T P [f (x, u) - f (\widehat{x}, u)] \right\| \\ & \quad \leq \gamma_1 \left\| e^T \sqrt{P} \right\| \left\| \sqrt{P} [f (x, u) - f (\widehat{x}, u)] \right\| \\ & \quad \leq \gamma_1 \gamma_2 \left\| e^T \sqrt{P} \right\| \left\| \sqrt{P} \right\| \| e \| \\ & \quad \leq \gamma_1 \gamma_2 \lambda_{\max}(P) e^T e, \end{aligned}$$

where we have used the fact that

$$\left\| \sqrt{P} [f (x, u) - f (\widehat{x}, u)] \right\| \leq \left\| \sqrt{P} \right\| \| f (x, u) - f (\widehat{x}, u) \| \leq \gamma_2 \left\| \sqrt{P} \right\| \| e \| ,$$

where $\left\| \sqrt{P} \right\|$ is the spectrum norm of $\sqrt{P}$.

For the second term

$$e^T Pf (x, u) \tilde{\theta} = \tilde{\theta}^T f^T (x, u) Pe = \tilde{\theta}^T h (x, u) Ce = \tilde{\theta}^T h (x, u) (y - C\widehat{x}) = \tilde{\theta}^T \dot{\widehat{\theta}};$$

Thus, $\dot{V} \leq -e^T [Q - 2\gamma_1 \gamma_2 \lambda_{\max}(P) I] e < 0$.

Denote that $Q - 2\gamma_1 \gamma_2 \lambda_{\max}(P) I = N$, where N is a positive definite matrix. For any $t > 0$, the following holds

$$\int_0^t e^T N e \mathrm{d}\tau \leq -\int_0^t \dot{V} d\tau = V(0) - V(t) < V(0) < \infty.$$

The well-posedness of the system guarantees that e is absolutely continuous; by Barbalet Lemma[6] and the fact that N is a positive definite matrix, we conclude that $e \to 0 \quad (t \to \infty)$. □

We have the following remarks to Theorem 5.5.1.

Remark 1 Condition 4 of Theorem 5.5.1 is derived from the estimation of $e^T P \left[f(x,u) - Pf(\widehat{x},u) \right] \theta$. By different estimation, we can obtain different conditions. For example,

$$\begin{aligned}\left\| e^T P \left[f(x,u) - f(\widehat{x},u) \right] \theta \right\| &\le \gamma_1 \left\| e^T P \left[f(x,u) - f(\widehat{x},u) \right] \right\| \\ &\le \tfrac{1}{2}\gamma_1 \left\| e^T P \right\|^2 + \tfrac{1}{2}\gamma_1 \left\| f(x,u) - f(\widehat{x},u) \right\|^2 \\ &\le \frac{e^T \gamma_1 P^2 e}{2} + \frac{\gamma_1 \gamma_2^2 e^T e}{2}.\end{aligned}$$

Then Condition 4 of Theorem 5.5.1 can be replaced by

$$\lambda_{\min}(Q) > \gamma_1 \lambda_{\max}\left(P^2\right) + \gamma_1 \gamma_2^2.$$

The Condition 4 can also be changed by

$$P(A - LC) + (A - LC)^T P - \gamma_1 P^2 - \gamma_1 \gamma_2^2 I < 0,$$

or its matrix form

$$\begin{bmatrix} P(A - LC) + (A - LC)^T P - \gamma_1 \gamma_2^2 I & P \\ P & -\frac{1}{\gamma_1} I \end{bmatrix} < 0.$$

It is in fact a linear matrix inequality. □

Remark 2 Condition 5 $h(x,u)\, C = \left[Pf(x,u) \right]^T$ means that $Pf(x,u) = C^T h^T(x,u)$. By Cramer Theorem, the necessary condition of existence of the solutions for the equation is that $\mathrm{rank} C^T = \mathrm{rank}\left[C^T \; Pf(x,u) \right]$, i.e., $\mathrm{span} C^T \supset \mathrm{span} Pf(x,u)$. Thus, $\mathrm{rank} f(x,u) \le r$ for any x, u, it also requires that $p \le r$.

Furthermore, C can be written as $C = \left[I \; 0 \right]$. Accordingly, $P = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix}$ and $f(x,u) = \begin{bmatrix} f_1(x,u) \\ f_2(x,u) \end{bmatrix}$. Then, $h(x,u)\, C = \left[Pf(x,u) \right]^T$ becomes

$$\left[h(x,u) \; 0 \right] = \left[f_1(x,u) P_{11} + f_2(x,u) P_{12}^T \quad f_1(x,u) P_{21} + f_2(x,u) P_{22} \right].$$

[6] Barbalet Lemma: If $f(t)$ is uniformly continuous, and the integral $\int_0^\infty f(t)dt$ exists, then $f(t) \to 0 \quad (t \to \infty)$. Please see Rochafellar (1970).

We have $f_2(x,u) = -f_1(x,u) P_{21} P_{22}^{-1}$. If the equality holds, then $h(x,u)$ always exists. □

Remark 3 In the theory of adaptive control system design, the gain matrix $f(x,u)$ before the uncertain parameter θ is always assumed to be a constant, or it is equal to B or G (see the books of adaptive control system), and Condition 5 is a natural result of positive realness. □

5.5.2 *Reduced-Order Adaptive Observers*

This subsection is to prove that there also exists a reduced-order observer under the conditions of full-order adaptive observer.

Theorem 5.5.2 If the following conditions hold

(1) There exists L such that $(H,\ A - LC,\ G)$ is both controllable and observable and is also strictly positive real.
(2) $\nu(\cdot)$ is monotone.
(3) The system Inc. (5.3.4) is well-posed.
(4) There exists a $h(x,u)$, such that $h(x,u)\,C = [Pf(x,u)]^T$.

Then, there exists an $(n-r)$-order reduce-order observer for Inc. (5.5.1).

Proof The differential inclusion System (5.5.1) can be written as

$$\begin{aligned}
&\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u - \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \omega + \begin{bmatrix} f_1(x,u) \\ f_2(x,u) \end{bmatrix} \theta, \\
&\quad \omega \in \nu(Hx), \\
&\quad y = x_1.
\end{aligned}$$

By Condition 1 of Theorem 5.5.1, there exists $P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$ such that

$$\begin{aligned}
&P(A - LC) + (A - LC)^T P \\
&\quad = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} A_{11} - L_1 & A_{12} \\ A_{21} - L_2 & A_{22} \end{bmatrix} + \begin{bmatrix} A_{11}^{\mathrm{T}} - L_1^{\mathrm{T}} & A_{21}^{\mathrm{T}} - L_2^{\mathrm{T}} \\ A_{12}^{\mathrm{T}} & A_{22}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \\
&\quad = -\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix},
\end{aligned}$$

where $Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$, $L = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}$. Then

$$P_{21} A_{12} + P_{22} A_{22} + A_{12}^T P_{12} + A_{22}^T P_{22} = -Q_{22},$$

i. e.,

$$P_{22}\left(P_{22}^{-1}P_{21}A_{12}+A_{22}\right)+\left(A_{12}^{\mathrm{T}}P_{12}P_{22}^{-1}+A_{22}^{T}\right)P_{22}=-Q_{22}.$$

It implies that $P_{22}^{-1}P_{21}A_{12}+A_{22}$ is a Hurwitz matrix.

Secondly, since

$$\left[\,G_1^{\mathrm{T}}\ G_2^{\mathrm{T}}\,\right]\begin{bmatrix}P_{11} & P_{12}\\ P_{21} & P_{22}\end{bmatrix}=\left[\,H_1\ H_2\,\right],$$

then $G_1^TP_{12}+G_2^TP_{22}=H_2$, i.e., $\left(G_1^TP_{12}P_{22}^{-1}+G_2^T\right)P_{22}=H_2$, which means that $\left(H_2,\ A_{22}+P_{22}^{-1}P_{12}A_{12}, G_2+P_{22}^{-1}P_{12}G_1\right)$ is strictly positive real.

By Condition 3, $h\left(x,u\right)C=\left[Pf\left(x,u\right)\right]^T$, we have

$$\left[f_1^T\left(x,u\right)f_2^T\left(x,u\right)\right]\begin{bmatrix}P_{11} & P_{12}\\ P_{21} & P_{22}\end{bmatrix}=\left[\,h\left(x,u\right)\ 0\,\right],$$

i.e.,$f_1^{\mathrm{T}}\left(x,u\right)P_{12}+f_2^T\left(x,u\right)P_{22}=0,\ \ f_1^T\left(x,u\right)P_{12}P_{22}^{-1}+f_2^T\left(x,u\right)=0.$

Choose $K=P_{22}^{-1}P_{12}$, $z_2=x_2+Kx_1=x_2+Ky$, then

$$\begin{aligned}\dot{z}_2&=\dot{x}_2+K\dot{x}_1\\&=\left(A_{12}+KA_{11}\right)x_1+\left(A_{22}+KA_{12}\right)x_2+\left(B_2+KB_1\right)u-\left(G_2+KG_1\right)\omega\\&\quad+\left(f_2\left(x,u\right)+Kf_1\left(x,u\right)\right)\theta.\end{aligned}$$

Because $f_2\left(x,u\right)+Kf_1\left(x,u\right)=0$, we can construct the reduced observer as follows

$$\begin{aligned}\dot{\widehat{z}}_2=&\left(A_{22}+KA_{12}\right)\widehat{z}_2+\left(B_2+KB_1\right)u-\left(G_2+KG_1\right)\widehat{\omega}+\left[\left(A_{21}+KA_{11}\right)\right.\\&\left.+\left(A_{22}+KA_{22}\right)K\right]y.\end{aligned}$$

Denote that $e_2=z_2-\widehat{z}_2$, then

$$\begin{aligned}&\dot{e}_2=\left(A_{22}+KA_{12}\right)e+\left(G_2+KG_1\right)\left(\omega-\widehat{\omega}\right),\\&\left(\omega-\widehat{\omega}\right)\in\nu(Hx)-\nu\left(H\widehat{x}\right).\end{aligned}$$

By Theorem 5.3.1, we then conclude that $e\to 0\quad\left(t\to\infty\right)$. □

It is necessary to mention the advantages of the reduced-order observer. Firstly, the condition 4 of Theorem 5.3.1 is not needed; secondly, the adaptive law does not need to be designed in the reduced-order observer.

From the proof of Theorem 5.5.2, if the equation $f_1^T\left(x,u\right)P_{12}P_{22}^{-1}+f_2^T\left(x,u\right)=0$ holds, then the condition $h\left(x,u\right)C=\left[Pf\left(x,u\right)\right]^T$ holds, i.e., Condition 5 can be replaced by $f_2^T(x,u) = -f_1^T(x,u)P_{12}P_{22}^{-1}$. Thus, the condition $h(x,u)C = [Pf(x,u)]^T$ is equivalent to the equation $PB = C^T$ which is necessary for the positive realness.

5.5.3 An Example of Adaptive Observer

A numerical example is given to show the effectiveness of the adaptive observer and the advantage of the reduce-order observer. Especially, the example illustrates the adaptive law need not be designed in the reduced-order observer.

Example 5.5.1 (Huang et al. 2011) Consider the following system

$$\begin{aligned}
\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} &= \begin{bmatrix} -10 & -3 & -1 \\ 6 & -5 & 4 \\ 1 & 0 & -9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \omega + \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 0.6\sin x_3 \\ 0 \\ 0 \end{bmatrix} \theta, \\
y &= x_1.
\end{aligned}$$

The set-valued mapping is defined as

$$v(x_1+3x_2+2x_3) = \begin{cases} x_1 + 3x_2 + 2x_3 + 3\,\mathrm{sgn}(x_1 + 3x_2 + 2x_3) & x_1 + 3x_2 + 2x_3 \neq 0, \\ [-3, 3] & x_1 + 3x_2 + 2x_3 = 0. \end{cases}$$

Let $|\theta| \leq 2$, i.e., $\gamma_1 = 2$; it can also be computed that $\gamma_2 = 0.6$. Choose that $l = \begin{bmatrix} -6 & 2 & 1 \end{bmatrix}^T$, then

$$A - lc^T = \begin{bmatrix} -4 & -3 & -1 \\ 4 & -5 & 4 \\ 0 & 0 & -9 \end{bmatrix}.$$

It is easy to determine that

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 0.5 \end{bmatrix}, Q = \begin{bmatrix} 8 & -1 & -1 \\ -1 & 10 & 3 \\ -1 & 3 & 5 \end{bmatrix},$$

and $\lambda_{\min}(Q) = 3.5401 > 3.1236 = 2\gamma_1\gamma_2\lambda_{\max}(P)$.

Solving $[Pf(x,u)]^T = \begin{bmatrix} 0.6\sin x_3 & 0 & 0 \end{bmatrix} = \begin{bmatrix} h(x) & 0 & 0 \end{bmatrix}$ yields $h(x) = 0.6\sin x_3$. Thus, the observer can be designed as

$$\begin{bmatrix} \dot{\widehat{x}}_1 \\ \dot{\widehat{x}}_2 \\ \dot{\widehat{x}}_3 \end{bmatrix} = \begin{bmatrix} -4 & -3 & -1 \\ 4 & -5 & 4 \\ 0 & 0 & -9 \end{bmatrix} \begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{x}_3 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \widehat{\omega} + \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 0.6\sin\widehat{x}_3 \\ 0 \\ 0 \end{bmatrix} \widehat{\theta} + \begin{bmatrix} -6 \\ 2 \\ 1 \end{bmatrix} y,$$

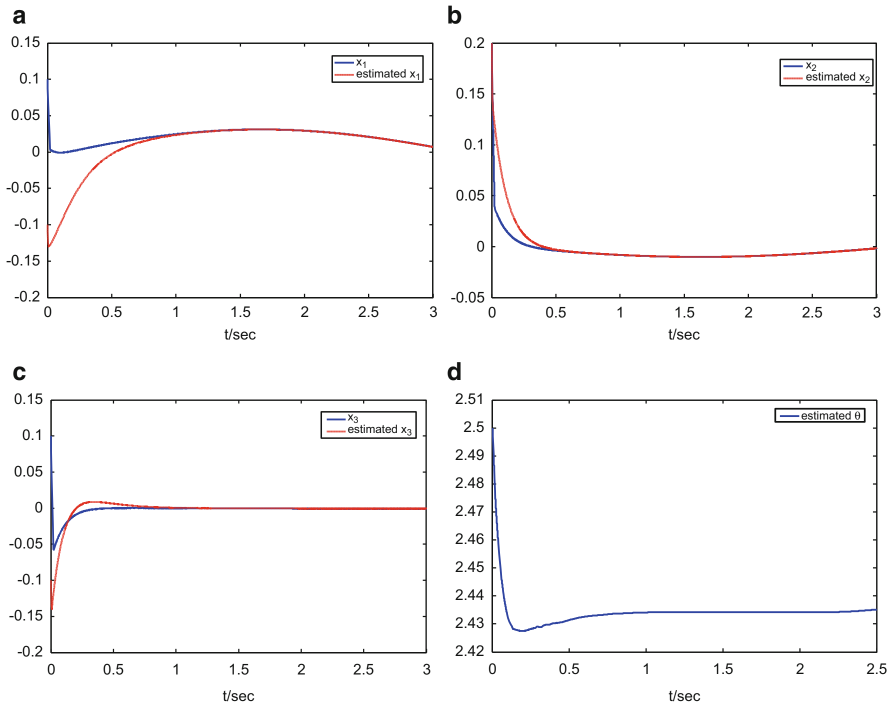


Fig. 5.15 The simulation result of Example 5.5.1. (**a**), (**b**) and (**c**) are the tracking performance of the states x_1, x_2, x_3, respectively, (**d**) is the estimation of $\widehat{\theta}$

$$\begin{aligned}\widehat{\omega} &\in \nu\left(\widehat{x}_1 + 3\widehat{x}_2 + 2\widehat{x}_3\right)\\ &= \begin{cases} \widehat{x}_1 + 3\widehat{x}_2 + 2\widehat{x}_3 + 3\operatorname{sgn}\left(\widehat{x}_1 + 3\widehat{x}_2 + 2\widehat{x}_3\right), & \widehat{x}_1 + 3\widehat{x}_2 + 2\widehat{x}_3 \neq 0, \\ [-3, 3], & \widehat{x}_1 + 3\widehat{x}_2 + 2\widehat{x}_3 = 0. \end{cases}\end{aligned}$$

The adaption law is

$$\dot{\widehat{\theta}} = 0.6 \sin \widehat{x}_3 \left(y - x_1\right).$$

In the simulation, θ is a constant 2.

From Fig. 5.15, it is obvious that the observer works well, and the estimated state will converge to the real state fast. It should be noted that $\widehat{\theta}$ does not converge to the real value 2.

By Theorem 5.5.2, since $P_{12} = 0$, $K = 0$, $\widehat{x}_2 = \widehat{z}_2$. A reduced-order observer is designed as follows

$$\begin{bmatrix} \dot{\widehat{x}}_2 \\ \dot{\widehat{x}}_3 \end{bmatrix} = \begin{bmatrix} -5 & 4 \\ 0 & -9 \end{bmatrix} \begin{bmatrix} \widehat{x}_2 \\ \widehat{x}_3 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix} \widehat{\omega} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 6 \\ 1 \end{bmatrix} y,$$

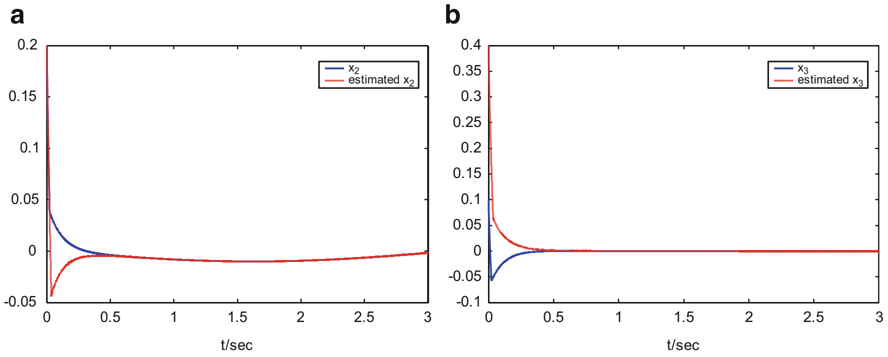


Fig. 5.16 The performance of reduced-order observer in Example 5.5.1. (**a**) and (**b**) are the tracking performance of the states x_2, x_3, respectively

$$\begin{aligned}\widehat{\omega} &\in \nu\left(y + 3\widehat{x}_2 + 2\widehat{x}_3\right)\\ &= \begin{cases} y + 3\widehat{x}_2 + 2\widehat{x}_3 + 3\operatorname{sgn}\left(y + 3\widehat{x}_2 + 2\widehat{x}_3\right), & y + 3\widehat{x}_2 + 2\widehat{x}_3 \neq 0,\\ [-3, 3], & y + 3\widehat{x}_2 + 2\widehat{x}_3 = 0.\end{cases}\end{aligned}$$

The estimation of parameter is not contained in reduced-order observer. The performance of reduced-order observer is shown in Fig. 5.16, and the convergence result is also well.

Problems

1. If $A \in \mathbb{R}^{n\times n}$, $\|A\|$ is any matrix norm, then there exists a kind of norm $\|x\|$ in $\mathbb{R}^n$, such that $\|A\|$ is induced norm by $\|x\|$.
2. Consider the system

$$\begin{aligned}\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} &= \begin{bmatrix}2 & -5\\ 3 & -7\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} - \begin{bmatrix}1\\ 0\end{bmatrix}\omega + \begin{bmatrix}1\\ -1\end{bmatrix}u + \begin{bmatrix}1\\ 0\end{bmatrix}\theta\\ \nu(x) &= \begin{cases} x_1 + \operatorname{sgn}(x_1), & x_1 \neq 0,\\ [-1, 1], & x_1 = 0,\end{cases}\\ y &= x_1.\end{aligned}$$

Please design an adaptive observer and a reduced-order observer, respectively.

3. Give a condition which is different from that in Theorem 5.5.1 or Remark 1 to estimate $e^T P\left[f(x, u) - f(\widehat{x}, u)\right]\theta$.
4. Consider the existence condition of adaptive observer for the system

$$\begin{aligned}\dot{x} &= Ax + B(x)u - G\omega + f(x, u)\,\theta,\\ \omega &\in \nu(Hx),\\ y &= Cx,\end{aligned}$$

where $B(x)$ satisfies Lipschitz condition, i.e., there exists γ_3 such that $\|B(x_1) - B(x_2)\| \leq \gamma_3 \|x_1 - x_2\|$. Then, consider the case where the input $B(x)u$ is extended to $B(x, u)$.

5. If the observer is designed as

$$\begin{aligned}
&\dot{\widehat{x}} = (A - LC)\widehat{x} + Bu - G\widehat{\omega} + f(\widehat{x}, u)\,\widehat{\theta} + Ly,\\
&\widehat{\omega} \in \nu\left(H\widehat{x} + C\widehat{x} - y\right),\\
&\dot{\widehat{\theta}} = h(\widehat{x}, u)\,(y - C\widehat{x}),
\end{aligned}$$

please give the existence conditions of the observer.

6. Rewrite condition 4 in Theorem 5.5.1 to linear matrix inequality.
7. Please discuss the condition $h(x, u)\,C = [Pf(x, u)]^T$ in Theorem 5.5.1, where $f(x, u)$ is an $n \times q$ function matrix. The necessary condition for $h(x, u)\,C = [Pf(x, u)]^T$ is that $q \leq v$. If $f(x, u) = Ef_1(x, u)$, $E \in \mathbb{R}^{n\times q}$, let $h(x, u) = f_1^T(x, u)\,F$, then the above equation becomes $FC = E^T P$. The condition 1 and condition 5 of Theorem 5.5.1 are that there exist $F \in \mathbb{R}^{q\times n}$ and $L \in \mathbb{R}^{n\times r}$, such that

$$\left(\begin{bmatrix} H \\ FC \end{bmatrix}, (A - LC), \begin{bmatrix} G & E \end{bmatrix}\right)$$

is strictly positive real. Please give the necessary condition, under which $\left(\begin{bmatrix} H \\ FC \end{bmatrix}, (A - LC), \begin{bmatrix} G & E \end{bmatrix}\right)$ is strictly positive real.

5.6 Adaptive Linear Observers

The adaptive linear observer design problem for the Luré differential inclusion system will be studied in this section. Consider the system Inc. (5.6.1). It is somewhat different from the system Inc. (5.5.1) by that the nonlinear term $f(x, u)$ is replaced by $f(y, u)$.

$$\begin{aligned}
&\dot{x} = Ax + Bu - G\omega - f(y, u)\,\theta,\\
&\omega \in \nu(x),\\
&y = Cx,
\end{aligned} \tag{5.6.1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^r$ are the state, input and output of the system, respectively, and $\omega \in \mathbb{R}^r$ is the output of set-valued mapping. A, B, G, C are the determined matrices with compatible dimensions. The condition of set-valued $\nu(x)$ is just to guarantee that the system Inc. (5.6.1) is well-posed, i.e., it may be not monotone. $\theta \in \mathbb{R}^p$ is a uncertain parameter vector, and it is always supposed to be a constant or a slow varying and bounded variable, i.e., $\|\theta\| \le \gamma_1$, with $\gamma_1 > 0; f(y, u)$ is a given function matrix; it is assumed that $f(y, u)$ is piecewise continuous.

The Problem 1 in this section is an extension of Lemma 5.4.1 to adaptive system. However, it is not feasible to extend the result to adaptive linear observer for the Luré differential inclusion system. This is because that the condition $h(x, u)\, C = f^T(x, u)\, P$ will not be satisfied. In the system Inc. (5.6.1), if we want to solve ω, the equation should be replaced by $h(x, u)\, C = f^T(x, u)\, M^T P$, where the matrix M is defined in §5.4. Since M is not of full rank, the equation may not be solvable except some special case. Thus, we will resort to some new method for the adaptive linear observer. In order to guarantee the convergence, a new definition of persistent excitation is needed.

5.6.1 Persistent Excitation

In the identification theory of control system, the persistently exciting signal is a very important, since the variation of the signal is persistent. The persistent variation can excite the inner features of the modeled system as much as possible. The persistent excitation is also used as the condition for variable convergence. Some basic properties of persistent excitation will be introduced in this section. The book does not be planned to present a detailed discussion on the persistent excitation which can be referred to other books about system identification or adaptive control. From viewpoint of application, we only deal with the relationship among the steady output of system, linear independence of functions, and the persistent excitations.

Definition 5.6.1 Let $\varphi : [0, \infty) \to \mathbb{R}^n$ be a piecewise continuous vector mapping. φ is called n dimension persistent excitation in $[0, \infty)$ if there exist positive constants $0 < k_1 \le k_2 < \infty$ and $M > 0$, such that for any $t \in [0, \infty)$,

$$k_1 I \le \int_t^{t+M} \varphi(\tau)\, \varphi^T(\tau)\, \mathrm{d}\tau \le k_2 I, \tag{5.6.2}$$

where I is the $n \times n$ identity matrix, and parameter k_1 is called the excitation level of φ. □

The integral defined in Inequality (5.6.2) can be considered the Lebesgue integration.

In the theory of control system design, we are concerned whether or not the output of a linear system is a persistent excitation. Consider the following example.

Example 5.6.1 Consider a linear system, its transfer function is

$$G(s) = \begin{bmatrix} 1 \\ \frac{1}{s+1} \end{bmatrix}.$$

When the input is the unit step signal $u(t) = 1 \ (t \geq 0)$, i.e., its Laplace transformation is $u(s) = 1/s$, then the output is

$$y(t) = \begin{bmatrix} 1 \\ 1 - e^{-t} \end{bmatrix}.$$

We now compute

$$y(t)y^T(t) = \begin{bmatrix} 1 & 1 - e^{-t} \\ 1 - e^{-t} & 1 - 2e^{-t} + e^{-2t} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -e^{-t} \\ -e^{-t} & -2e^{-t} + e^{-2t} \end{bmatrix}.$$

The first term is the steady term at the last expression of the above equation, and the second term is transient term which converges to 0 as $t \to \infty$. Its integration can be computed that

$$\begin{aligned} &\begin{bmatrix} \int\limits_t^{t+M} 1 d\tau & \int\limits_t^{t+M} \left(1 - e^{-t}\right) d\tau \\ \int\limits_t^{t+M} \left(1 - e^{-t}\right) d\tau & \int\limits_t^{t+M} \left(1 - 2e^{-t} + e^{-2t}\right) d\tau \end{bmatrix} \\ &= \begin{bmatrix} \int\limits_t^{t+M} 1 d\tau & \int\limits_t^{t+M} 1 d\tau \\ \int\limits_t^{t+M} 1 d\tau & \int\limits_t^{t+M} 1 d\tau \end{bmatrix} + \begin{bmatrix} 0 & -\int\limits_t^{t+M} e^{-t} d\tau \\ -\int\limits_t^{t+M} e^{-t} d\tau & \int\limits_t^{t+M} \left(-2e^{-t} + e^{-2t}\right) d\tau \end{bmatrix} \\ &= \begin{bmatrix} M & M \\ M & M \end{bmatrix} + \begin{bmatrix} 0 & e^{-(t+M)} - e^{-t} \\ e^{-(t+M)} - e^{-t} & \frac{e^{-2t} - e^{-2(t+M)}}{2} \end{bmatrix}. \end{aligned}$$

Since the second term converges to 0 as $t \to \infty$, it can be neglected in checking Inequality (5.6.2). The reason is that the exponential decay term will vanish when t goes to infinite. The first term is not positive definite for any M; thus, the output $y(t)$ is not a persistent excitation.

If the input signal is $u(t) = \sin t \ \ (t \geq 0)$, i.e., $u(s) = 1/\left(s^2 + 1\right)$, then

$$y(t) = \begin{bmatrix} \sin t \\ \frac{e^{-t} + \sqrt{2} \sin\left(t - \frac{\pi}{4}\right)}{2} \end{bmatrix}.$$

By the above analysis, deleting exponential decay term does not affect its judgment of persistent excitation. Thus, we only need to verify Inequality (5.6.2) for the following matrix

$$\begin{bmatrix} \sin t \\ \dfrac{\sin\left(t-\frac{\pi}{4}\right)}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sin t & \dfrac{\sin\left(t-\frac{\pi}{4}\right)}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \sin^2 t & \frac{1}{\sqrt{2}}\sin t \sin\left(t-\frac{\pi}{4}\right) \\ \frac{1}{\sqrt{2}}\sin t \sin\left(t-\frac{\pi}{4}\right) & \frac{1}{2}\sin^2\left(t-\frac{\pi}{4}\right) \end{bmatrix}.$$

Let $M = 2\pi$, then

$$\begin{bmatrix} \int\limits_t^{t+2\pi} \sin^2\tau d\tau & \frac{1}{\sqrt{2}}\int\limits_t^{t+2\pi} \sin\tau\sin\left(\tau-\frac{\pi}{4}\right)d\tau \\ \frac{1}{\sqrt{2}}\int\limits_t^{t+2\pi} \sin\tau\sin\left(\tau-\frac{\pi}{4}\right)d\tau & \frac{1}{2}\int\limits_t^{t+2\tau} \sin^2\left(\tau-\frac{\pi}{4}\right)d\tau \end{bmatrix},$$

$$= \begin{bmatrix} \int\limits_0^{2\pi} \sin^2\tau d\tau & \frac{1}{\sqrt{2}}\int\limits_0^{2\pi} \sin\tau\sin\left(\tau-\frac{\pi}{4}\right)d\tau \\ \frac{1}{\sqrt{2}}\int\limits_0^{2\pi} \sin\tau\sin\left(\tau-\frac{\pi}{4}\right)d\tau & \frac{1}{2}\int\limits_0^{2\tau} \sin^2\left(\tau-\frac{\pi}{4}\right)d\tau \end{bmatrix}$$

$$= \pi \begin{bmatrix} 1 & \dfrac{1}{2} \\ \dfrac{1}{2} & \dfrac{1}{2} \end{bmatrix}.$$

It is direct to conclude that the matrix derived is positive definite. In Inequality (5.6.2), k_1 and k_2 can be chosen as $\dfrac{3-\sqrt{5}}{4}\pi$ and $\dfrac{3+\sqrt{5}}{4}\pi$, respectively; thus, the output $y(t)$ is a persistent excitation signal. □

Obviously, if the transfer function is changed as

$$G(s) = \begin{bmatrix} \dfrac{1}{s+1} \\ \dfrac{1}{s+1} \end{bmatrix}.$$

No matter what the input is, the components of the output are the same, and Inequality (5.6.2) does not hold. Thus, the persistent excitation property of $y(t)$ depends on both the input and the transfer function of a linear control system.

From Example 5.6.1, we can see that it not easy to verify the persistent excitation property of an output signal by using Inequality (5.6.2). Thus, it is necessary to look for other criteria. Spectrum measure is often used to describe the sufficient and necessary condition of the persistent excitations. But it is still quite complicated. Instead of these theoretic measures, we only give an applied sufficient condition for the excitations by using linear independence of functions.

From Inequality (5.6.2), for a constant $a \in \mathbb{R}^n$, it holds that

$$a^T \left(\int_t^{t+M} \varphi(\tau) \varphi^T(\tau) d\tau \right) a = \int_t^{t+M} a^T \varphi(\tau) \varphi^T(\tau) a d\tau = \int_t^{t+M} \left(a^T \varphi(\tau) \right)^2 d\tau \geq 0.$$

For any $M > 0$, the integration $\int_t^{t+M} \varphi(\tau) \varphi^T(\tau) d\tau$ is an $n \times n$ semipositive definite matrix in the real number field. For promoting the study further, we need the following definition.

Definition 5.6.2 Let $\varphi_1(t), \varphi_2(t), \ldots, \varphi_n(t)$ be n functions whose images are $\mathbb{R}^p$ vectors for every t. The n vector-valued functions are linear dependent in the closed real interval $[t_1, t_2]$ if there exist real numbers $a_1, a_2, \ldots, a_n$ which are not all zero such that

$$a_1 \varphi_1(t) + a_2 \varphi_2(t) + \cdots + a_n \varphi_n(t) \equiv 0, \quad t \in [t_1, t_2].$$

Otherwise, they are linear independent in the closed real interval $[t_1, t_2]$. For simplicity, it is called by linear independent. □

Obviously, linear independence in a closed real interval $[t_1, t_2]$ means that linear independence in every subset which includes the interval $[t_1, t_2]$.

The following theorem gives a criterion for determining whether the n vector functions are linear dependence, the proof is left to readers.

Let $\varphi_1(t), \varphi_2(t), \ldots, \varphi_n(t)$ be n functions with range $\mathbb{R}^p$. These functions are assumed to exist at least $(n-1)$ derivative in the interval $[t_1, t_2]$. Denote that

$$\Phi(t) = \begin{bmatrix} \varphi_1^T(t) \\ \varphi_2^T(t) \\ \vdots \\ \varphi_n^T(t) \end{bmatrix} \tag{5.6.3}$$

then $\Phi(t)$ is an $n \times p$ function matrix defined on $[t_1, t_2]$. Denote $\Phi^{(i)}(t)$ for the matrix which is the differential of the ith order of $\Phi(t)$.

Theorem 5.6.1 $\varphi_1(t), \varphi_2(t), \ldots, \varphi_n(t)$ are linear independent on the interval $[t_1, t_2]$ if and only if there exists a $t_0 \in [t_1, t_2]$, such that

$$\operatorname{rank} \left[\Phi(t_0) \quad \Phi^{(1)}(t_0) \quad \cdots \quad \Phi^{(n-1)}(t_0) \right] = n.$$

Not that $\left[\Phi(t_0) \quad \Phi^{(1)}(t_0) \quad \cdots \quad \Phi^{(n-1)}(t_0) \right]$ is an $n \times np$ real matrix.

Example 5.6.2 Let $\varphi_1(t) = 1,\ \varphi_2(t) = t,\ \varphi_3(t) = t^2$ be three scale functions, but they are linear independent on any interval of $[0, \infty)$.

It is because

$$\Phi(t) = \begin{bmatrix} 1 \\ t \\ t^2 \end{bmatrix}, \quad \begin{bmatrix} \Phi\left(t_0\right) & \Phi^{(1)}\left(t_0\right) & \Phi^{(2)}\left(t_0\right) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ t_0 & 1 & 0 \\ t_0^2 & 2t_0 & 1 \end{bmatrix},$$

By Theorem 5.6.1, $\varphi_1(t), \varphi_2(t), \varphi_3(t)$ are linear independent. □

The above example shows that for the linear independency the function is different from the vector. We have proved that three scare functions are linear independent in the real domain.

The following theorem is about the linear independency of functions. We need a definition.

Definition 5.6.3 Let $\varphi_1(t), \varphi_2(t), \ldots, \varphi_n(t)$ be n functions whose images are all in $\mathbb{R}^p$, and the matrix $\Phi(t)$ is defined by Eq.(5.6.3), then

$$W\left(t_1, t_2\right) = \int_{t_1}^{t_2} \Phi\left(\tau\right) \Phi^T\left(\tau\right) d\tau$$

is the Gram matrix of $\varphi_1(t), \varphi_2(t), \ldots, \varphi_n(t)$ on the interval $[t_1, t_2]$. □

From the definition, for any $t_1 < t_2$, Gram matrix $W(t_1, t_2)$ is always an $n \times n$ semipositive definite.

Theorem 5.6.2 If $\varphi_1(t), \varphi_2(t), \ldots, \varphi_n(t)$ are continuous on the interval $[t_1, t_2]$, then they are linear independent on the interval $[t_1, t_2]$ if and only if their Gram matrix $W(t_1, t_2)$ is invertible, or positive definite.

Proof Necessity. Assume that the Gram matrix $W(t_1, t_2)$ is not invertible, then there exists a nonzero real vector $a \in \mathbb{R}^n$, such that $W\left(t_1, t_2\right) a = 0$. Then we have the following equation

$$a^T W\left(t_1, t_2\right) a = \int_{t_1}^{t_2} a^T \Phi\left(\tau\right) \Phi^T\left(\tau\right) a d\tau = \int_{t_1}^{t_2} \left(a^T \Phi\left(\tau\right)\right)^2 d\tau = 0,$$

$a^T\Phi(t) \equiv 0,\ t \in [t_1, t_2]$, i.e., $\varphi_1(t), \varphi_2(t), \ldots, \varphi_n(t)$ are linear dependent on the interval $[t_1, t_2]$.

Sufficiency. If $\varphi_1(t), \varphi_2(t), \ldots, \varphi_n(t)$ are linear dependent on the interval $[t_1, t_2]$, then there exist real number $a_1, a_2, \ldots, a_n$ which are not all zero, such that

$$a_1\varphi_1(t) + a_2\varphi_2(t) + \cdots + a_n\varphi_n(t) \equiv 0,\ \ t \in [t_1, t_2]\,.$$

Denote $a = [a_1 \ \ a_2 \ \ \cdots a_n]^T$, then $\alpha \neq 0$. Consider that $\Phi^T(t)a = 0$, we obtain that

$$W(t_1, t_2)\, a = \int_{t_1}^{t_2} \Phi(\tau)\, \Phi^T(\tau)\, a d\tau = 0.$$

Thus, $W(t_1, t_2)$ is not invertible. □

Let us return to the persistent excitation. The output of the system $y(t)$ can be decomposed to steady term $y_s(t)$ and transient term $y_t(t)$, i.e., $y(t) = y_s(t) + y_t(s)$, where the transient term is infinitesimal when $t \to \infty$, i.e., $\lim\limits_{t\to\infty} y_t(t) = 0$. In the Example 5.6.1, when the input is a unit step signal, the output $y(t)$ can be written as

$$y(t) = \begin{bmatrix} 1 \\ 1 - e^{-t} \end{bmatrix}, \ \ y_s(t) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \ \ y_t(t) = \begin{bmatrix} 0 \\ -e^{-t} \end{bmatrix}.$$

By the frequency response in control theory, if transfer function of a stable single-input and single-output system is $G(s)$, when it is inputted by $A \sin \omega t$, its steady output is $y_s(t) = A\, |G(j\omega)| \sin(\omega t + \phi)$, where $\phi = \angle G(j\omega)$, i.e., the steady output is a sinusoidal signal with the same frequency. The result can be extended to the vector case.[7] Continue to consider Example 5.6.1, when the input is $\sin t$,

$$G(j) = \begin{bmatrix} 1 \\ \frac{1}{1+j} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{\sqrt{2}}\left(\cos\left(-\frac{\pi}{4}\right) + j \sin\left(-\frac{\pi}{4}\right)\right) \end{bmatrix},$$

then the steady output is

$$y_s(t) = \begin{bmatrix} \sin t \\ \frac{1}{\sqrt{2}} \sin\left(t - \frac{\pi}{4}\right) \end{bmatrix}.$$

The steady term $y_s(t)$ of output $y(t)$ can be written as the following component form:

$$y_s(t) = \begin{bmatrix} y_{1s}(t) \\ y_{2s}(t) \\ \vdots \\ y_{rs}(t) \end{bmatrix}.$$

[7] The meaning of $A\,|G(j\omega)| \sin(\omega t + \phi)$ is that if $G(s) = \left[\, g_1(s) \cdots g_n(s) \,\right]^T$, then $A\,|G(\mathrm{j}\omega)| \sin(\omega t + \phi) = A\left[\, |g_1(\mathrm{j}\omega)| \sin(\omega t + \angle g_1(\mathrm{j}\omega)) \cdots |g_n(\mathrm{j}\omega)| \sin(\omega t + \angle g_n(j\omega)) \,\right]^T$.

If $y_{1s}(t),\ y_{2s}(t),\ \ldots, y_{ms}(t)$ are linear independent on the interval $[0,\ \infty)$, then its Gram matrix $W(0,\ t) = \int\limits_0^t y_s(\tau)\, y_s^T(\tau)\, d\tau$ is always invertible. We are ready to state the following theorem.

Theorem 5.6.3 If the output $y(t)$ has at least $(n-1)$th order continuously derivative, and its steady term satisfies

(1) $y_{1s}(t),\ y_{2s}(t),\ \ldots, y_{ms}(t)$ are linear independent on the interval $[0,\ \infty)$.
(2) $y_{1s}(t),\ y_{2s}(t),\ \ldots, y_{ms}(t)$ are bounded on the interval $[0,\ \infty)$.

Then $y(t)$ is a persistent excitation.

Proof Denote that

$$W_s(t_1, t_2) = \int\limits_{t_1}^{t_2} y_s(\tau)\, y_s^T(\tau)\, d\tau$$

Condition 1 implies that $W_s(t_1, t_2)$ is an invertible matrix by Theorem 5.6.2. Replacing t_2 by $t_1 + M$, let us consider the matrix $W_s(t_1, t_1 + M)$. Assume that there do not exist M and k_2 such that $W(t_1, t_1 + M) \le k_2 I$, then for any fixed M, there exist $t_{1i},\ i = 1, 2, \ldots,$ such that $W(t_{1i},\ t_{1i} + M) > iI$. Hence, for every i, there exists $\alpha \in \mathbb{R}^n$, $|\alpha| = 1$, such that $\alpha_i^T W_s(t_{1i}, t_{1i} + M)\, \alpha_i > i$. Then the following inequality holds

$$\begin{aligned} \alpha_i^T W_s(t_1, t_2)\, \alpha_i &= \int\limits_{t_i}^{t_i+M} \alpha_i^T y_s(\tau)\, y_s^T(\tau)\, \alpha_i d\tau, \\ &= \int\limits_{t_i}^{t_i+M} \left\|\alpha_i^T y_s(\tau)\right\|^2 d\tau \\ &= \left\|\alpha_i^T y_s(\bar{t}_i)\right\|^2 M \\ &\le \|y_s(\bar{t}_i)\|^2 M, \end{aligned}$$

where $\bar{t}_i \in [t_i, t_i + M]$. The last equality is derived by the integral mean value theorem. Thus, $\|y_s(\bar{t}_i)\| > M^{-1}\sqrt{i}$. In view of the fact that M is fixed, it contradicts to Condition 2.

Then we prove that there exist M and k_1 such that $W(t_1, t_1 + M) \ge k_1 I$. For any component $y_{js}(t), j = 1, 2, \ldots, m$, $y_s(t)$ is not equal to zero on an open interval. Because the steady term does not converge to zero when $t \to \infty$, for any given $\varepsilon > 0$, there exists an $M > 0$, such that $W(t_1,\ t_1 + M) > \varepsilon I$ for any t_1. Thus, $y(t)$ is a persistent excitation. □

Theorem 5.6.3 presents a practical criterion for persistent excitation. We note that Theorem 5.6.3 is also useful for multiinput and multioutput systems but not only for single ones

Using Theorem 5.6.3, we continue to consider Example 5.6.1. When the input is $\sin t$, the steady output is

$$y_s(t) = \begin{bmatrix} \sin t \\ \frac{1}{\sqrt{2}} \sin\left(t - \frac{\pi}{4}\right) \end{bmatrix}.$$

Since

$$\left[y_s(t)\ \dot{y}_s(t) \right] = \begin{bmatrix} \sin t & \cos t \\ \frac{1}{\sqrt{2}} \sin\left(t - \frac{\pi}{4}\right) & \frac{1}{\sqrt{2}} \cos\left(t - \frac{\pi}{4}\right) \end{bmatrix},$$

$\det\left[y_s(t)\ \dot{y}_s(t) \right] \neq 0$, the components of $y_s(t)$ are linear independent on the real domain. The components are all bounded; thus, $y(t)$ is a persistent excitation.

5.6.2 Linear Adaptive Observers

A linear adaptive observer for the Luré differential inclusion system is designed by using persistent excitation signal in this subsection. In the design of adaptive observer, it is common to employ persistent excitation (see, for example, Rochafellar 1970). A lemma is introduced firstly.

Lemma 5.6.1 If φ(t) is a persistent excitation signal in $\mathbb{R}^n$ and $\gamma(t): \ \mathbb{R}(\geq 0) \to \mathbb{R}$ satisfies $|\gamma(t)| \leq \beta e^{-\alpha t}$ where $\alpha, \beta > 0$, then the following differential equation

$$\dot{\theta}(t) = -\varphi(t)\varphi^T(t)\theta(t) + \varphi(t)\lambda(t)$$

is globally asymptotically stable.

Proof The solution of the differential equation is

$$\theta(t) = e^{-\int\limits_0^t \varphi(\tau)\varphi^T(\tau)d\tau}\theta(0) + \int\limits_0^t e^{-\int\limits_s^t \varphi(\tau)\varphi^T(\tau)d\tau}\varphi(s)\lambda(s)ds.$$

Hence,

$$\theta(t) \leq e^{-k_1 t}\left\|\theta(0)\right\| + \int\limits_0^t e^{-k_1(t-s)}\left\|\varphi(s)\lambda(s)\right\| ds.$$

Since $\varphi(t)$ is a persistent excitation, $\|\varphi(t)\|$ is bounded. Assume that $\|\varphi(t)\| \leq A$, and select an $\alpha \neq k_1$, then

$$\theta(t) \leq e^{-k_1 t}\|\theta(0)\| + A\int_0^t e^{-k_1(t-s)}e^{-\alpha s}ds \leq e^{-k_1 t}\|\theta(0)\| + \frac{A}{k_1-\alpha}\left(e^{-\alpha t} - e^{-k_1 t}\right).$$

Thus, we have completed the lemma. □

We now consider the adaptive observer design for the Luré system Inc. (5.6.1). Continue to use the symbols defined in Sect. 5.4, particularly $M = I - G(CG)^{-1}C$. When the relative degree of (C, A, G) is one, M exists and is not invertible. By the results of Sect. 5.4, if (C, A, G) is both controllable and observable, and minimum phase, then (MA, C) is detectable; therefore, there exists an L such that $MA - LC$ is a Hurwitz matrix.

Theorem 5.6.4 Consider the Luré system Inc. (5.6.1). If the following conditions are all satisfied.

(1) The system (C, A, G) is both controllable and observable, and the system is minimum phase with relative degree one.
(2) $\xi(t)$ is the solution of the following differential equation, and $\xi(t)$ is a persistent excitation,

$$\xi = (MA - LC)\,\xi - Mf\,(y, u)\,, \tag{5.6.4}$$

where L is the gain matrix such that $MA - LC$ is a Hurwitz matrix.

Then,

$$\begin{aligned} \dot{\eta} &= (MA - LC)\,\eta + MBu - Mf\,(y, u)\,\widehat{\theta} + \xi\dot{\widehat{\theta}} + MAG(CG)^{-1}y, \\ \widehat{x} &= \eta + G(CG)^{-1}y, \end{aligned} \tag{5.6.5}$$

is an asymptotically adaptive linear observer for Inc. (5.6.1), and the adaption law is

$$\dot{\widehat{\theta}} = \xi^T C^T\left(y - C\widehat{x}\right). \tag{5.6.6}$$

Proof By the proof of Theorem 5.4.2, we can obtain from Inc. (5.6.1) that

$$\omega = (CG)^{-1}\,(CAx + cBu) - Cf\,(y, u)\,\theta - \dot{y}$$

Thus, the first equation of Inc. (5.6.1) is

$$\dot{x} = MAx + MBu - Mf\,(y, u)\,\theta + G(CG)^{-1}\dot{y}. \tag{5.6.7}$$

Taking derivation of the second equation of Inc. (5.6.5), we have

$$\begin{aligned}\dot{\widehat{x}} &= (MA-LC)\left(\widehat{x}-G(CG)^{-1}y\right)+MBu-Mf\,(y,u)\,\widehat{\theta}+\xi\dot{\widehat{\theta}}\\ &\quad+MAG(CG)^{-1}y+G(CG)^{-1}\dot{y}\\ &=(MA-LC)\widehat{x}+Ly+MBu-Mf\,(y,u)\,\widehat{\theta}+\xi\dot{\widehat{\theta}}+G(CG)^{-1}\dot{y}.\end{aligned}\tag{5.6.8}$$

Denote that $e=x-\widehat{x}$, subtracting Eq. (5.6.8) from Eq.(5.6.7) yields

$$\dot{e}=(MA-LC)\,e-Mf\,(y,u)\,\tilde{\theta}-\xi\dot{\widehat{\theta}},$$

where $\tilde{\theta}=\theta-\widehat{\theta}$. Since θ is treated as a constant, hence, $\dot{\tilde{\theta}}=-\dot{\widehat{\theta}}$ in the above equation.

We now analyze the convergence of e. Define that $\zeta=e-\xi\tilde{\theta}$, then

$$\begin{aligned}\dot{\zeta}&=\dot{e}-\dot{\xi}\tilde{\theta}+\xi\dot{\widehat{\theta}}\\ &=(MA-LC)\,e-Mf\,(y,u)\,\tilde{\theta}-\dot{\xi}\tilde{\theta}\\ &=(MA-LC)\,\zeta+(MA-LC)\,\xi\tilde{\theta}-Mf\,(y,u)\,\tilde{\theta}-\dot{\xi}\tilde{\theta}\\ &=(MA-LC)\,\zeta+\left((MA-LC)\,\xi-Mf\,(y,u)-\dot{\xi}\right)\tilde{\theta}\ .\end{aligned}$$

By Eq. (5.6.4), we obtain that

$$\dot{\zeta}=(MA-LC)\,\zeta.$$

Since $MA-LC$ is a Hurwitz matrix, there exist $\alpha,\ \beta>0$ such that $|\zeta(t)|<\beta e^{-\alpha t}$ where β may depend on the initial condition $\zeta(0)$.

According to the adaption law Eq. (5.6.6),

$$\dot{\tilde{\theta}}=-\xi^T C^T Ce=-\xi^T C^T C\xi\tilde{\theta}-\xi^T C^T C\zeta.\tag{5.6.9}$$

In view of Condition 2 and the fact that $|\zeta(t)|<e^{-\alpha t}$, by Lemma 5.6.1, we have $\tilde{\theta}\to 0\ \ (t\to\infty)$. Since ξ is persistent exciting, and it is bounded; thus $e=\zeta+\xi\tilde{\theta}\to 0\ \ (t\to\infty)$. □

The following remarks are given for Theorem 5.6.4.

Remark 1 Differing from Sect. 5.4, Lyapunov function is not employed to design the adaptive law. The adaptive law Eq. (5.6.6) is designed to make the auxiliary variable ζ stable. It points out that the $\widehat{\theta}$ can also converge to the real value of θ under the adaptive law. □

Remark 2 In Eq. (5.6.4), $MA - LC$ is a Hurwitz matrix, whether the steady term is persistently exciting depends on $Mf(y,u)$. Hence, it depends on $f(y,u)$. □

Remark 3 Eq. (5.6.9) is a time-varying nonhomogeneous linear equation, and the persistent excitation condition guarantees that Eq. (5.6.9) is stable. If there exist other conditions such that the equation is stable, then the persistent excitation condition can be replaced by that one. Especially, in some specific example, there may exist much more practical conditions. □

Example 5.6.2 Consider the Luré differential inclusion system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -0.5 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \omega + \begin{bmatrix} 0 \\ u^2 + y \end{bmatrix} \theta.$$
$$y = x_1,$$

The set-valued mapping is

$$\nu(x_1, x_2) = \begin{cases} x_1 + x_2 + 6, & x_1 + x_2 < -6; \\ [-x_1 - x_2 - 6,\ x_1 + x_2 + 6], & -6 \le x_1 + x_2 < 0; \\ [x_1 + x_2 - 6, -x_1 - x_2 + 6], & 0 \le x_1 + x_2 < 6; \\ -x_1 - x_2 + 6, & 6 \le x_1 + x_2. \end{cases}$$

The graph of set-valued mapping $\nu(x_1, x_2)$ is shown in Fig. 5.17; any curve in the diamond belongs to the selections of set-valued mapping. From Fig. 5.17, we can see that $\nu(x_1, x_2)$ is not monotone; thus, the solution of the Luré differential inclusion system is not unique. But for every piecewise continuous function, the solution is unique; thus, it is well posed.

It is easy to verify that $c^{\mathrm{T}}g = 1$, the zero of the system is -1; thus, the condition 1 of Theorem 5.6.4 holds. Let $l = \begin{bmatrix} 0.5 & 1 \end{bmatrix}^{\mathrm{T}}$. Then

$$M = \begin{bmatrix} 0 & 0 \\ -1 & 1 \end{bmatrix}, \quad MA - lc^T = \begin{bmatrix} -0.5 & 0 \\ -1.5 & -1 \end{bmatrix}.$$

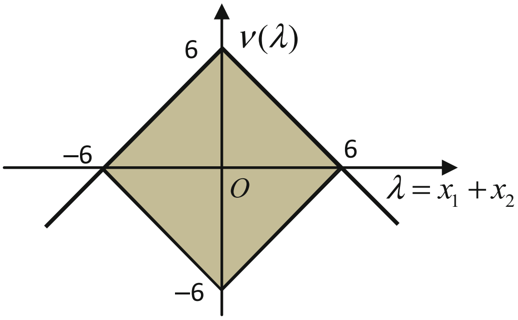


Fig. 5.17 The set-valued mapping in Example 5.6.1

$MA - lc^{\mathrm{T}}$ is a Hurwitz matrix. Using these data, Eq. (5.6.4) becomes

$$\dot{\xi}(t) = \begin{bmatrix} -0.5 & 0 \\ -1.5 & -1 \end{bmatrix} \xi(t) + \begin{bmatrix} 0 \\ u^2 + y \end{bmatrix}.$$

The linear adaptive observer is

$$\begin{aligned} \dot{\eta} &= \begin{bmatrix} -0.5 & 0 \\ -1.5 & -1 \end{bmatrix} \eta + \begin{bmatrix} 0 \\ -1 \end{bmatrix} u - \begin{bmatrix} 0 \\ u^2 + y \end{bmatrix} \widehat{\theta} + \xi \dot{\widehat{\theta}} + \begin{bmatrix} 0 \\ -1.5 \end{bmatrix} y, \\ \widehat{x} &= \eta + \begin{bmatrix} 1 \\ 1 \end{bmatrix} y, \end{aligned}$$

and the adaption law is

$$\dot{\widehat{\theta}} = \begin{bmatrix} \xi_1 & \xi_2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} (y - \widehat{x}_1) = \xi_1 \left(y - \widehat{x}_1 \right).$$

In order to make $\xi(t)$ to be a persistent excitation, $u(t)$ is chosen as $3 \sin t$.

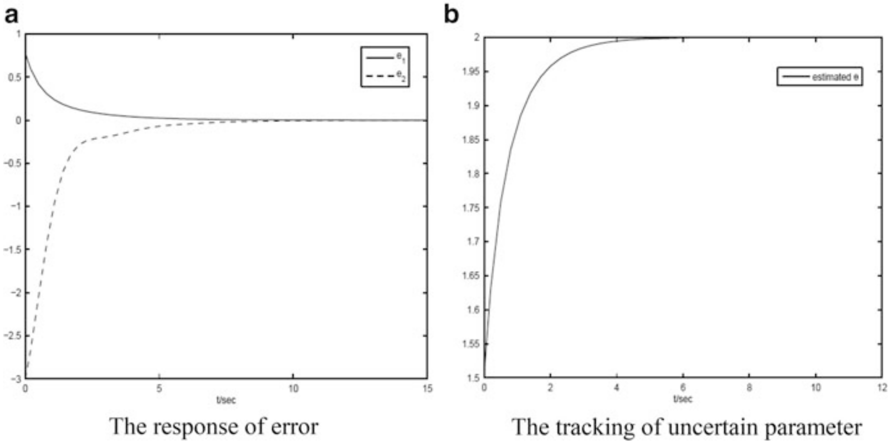
We choose two selections of $v(x_1, x_2)$ as follows:

$$\omega_1 = \begin{cases} \lambda + 6, & \lambda < 0, \\ -\lambda + 6, & 0 \leq \lambda; \end{cases} \qquad \omega_2 = \begin{cases} \lambda + 6, & \lambda < -6, \\ -\lambda - 6, & -6 \leq \lambda < 0, \\ \lambda - 6, & 0 \leq \lambda < 6, \\ -\lambda + 6, & 6 \leq \lambda; \end{cases}$$

where $\lambda = x_1 + x_2$. ω_1 and ω_2 are the boundaries of the set-valued mapping (Fig. 5.17). Obviously, the trajectories are different with different selections. It is also illustrated that the asymptotical convergence can be not achieved by Luenberger observer.

The responses of errors are shown in Figs. 5.18 and 5.19, respectively, when $v = \omega_1$ and $v = \omega_2$, where the uncertain parameter θ is chosen as 2. In Fig. 5.18, the initial state is $x(0) = \begin{bmatrix} 2 & 1 \end{bmatrix}^T$, the initial value of auxiliary variable is $\xi(0) = \begin{bmatrix} -0.8 & 2 \end{bmatrix}^T$, the initial value of observer is $\eta(0) = \begin{bmatrix} 1 & 2 \end{bmatrix}^T$, and the initial value of adaption law is $\widehat{\theta}(0) = 1.5$. In Fig. 5.19, the initial value of the state is $x(0) = \begin{bmatrix} 0 & 1 \end{bmatrix}^T$, the initial value of auxiliary variable is $\xi(0) = \begin{bmatrix} -1.6 & 1 \end{bmatrix}^T$, the initial value of observer is $\eta(0) = \begin{bmatrix} -1 & -2 \end{bmatrix}^T$, and the initial value of adaptive law is $\widehat{\theta}(0) = 3$.

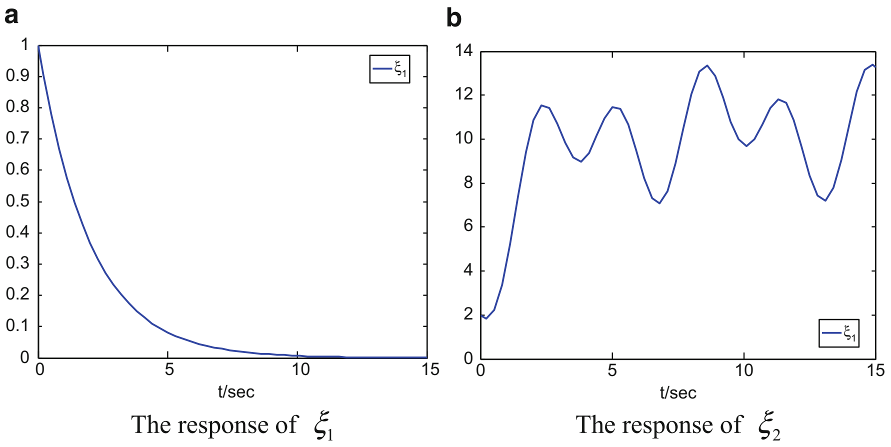
From Figs. 5.18 and 5.20, the observer works well and the parameter is also convergent. The response of auxiliary variable ξ is given in Fig. 5.19, where ξ_1 is not affected by $f(y, u)$. It is then monotone, and its steady component is 0. The steady component of ξ_2 is vibrating and bounded, and it is a persistent excitation. □



The response of error

The tracking of uncertain parameter

Fig. 5.18 The performance of adaptive observer when $\nu = \omega_1$. (**a**) The response of error. (**b**) The tracking of uncertain parameter



The response of ξ_1

The response of ξ_2

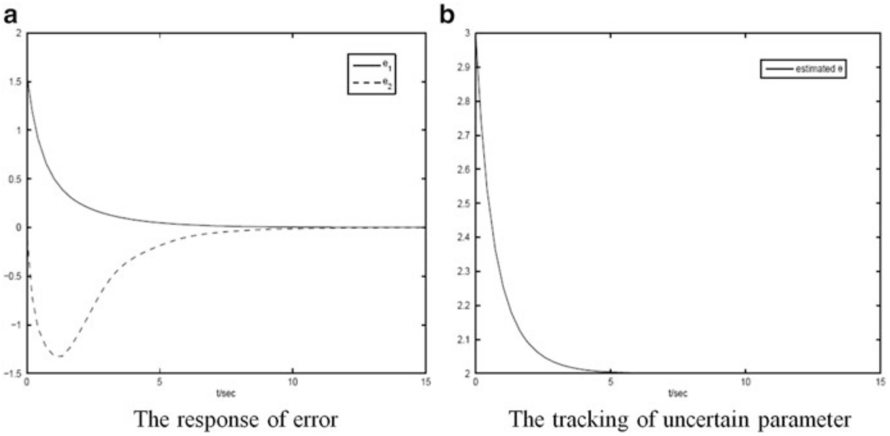
Fig. 5.19 The response of ξ when $\nu = \omega_1$. (**a**) The response of ξ_1 (**b**) The response of ξ_2

Problems

1. Consider system

$$\dot{x} = Mx + Ny + Bu + f(x, u)\,\theta,$$
$$\dot{y} = Cx,$$

where $\|\theta\| \leq \gamma_1$ and $f(x, u)$ satisfies Lipschitzian condition, and Lipschitzian constant is γ_2. Besides, the following conditions hold.



The response of error

The tracking of uncertain parameter

Fig. 5.20 The performance of adaptive observer when $\nu = \omega_2$. (**a**) The response of error. (**b**) The tracking of uncertain parameter

(1) There exist positive definite matrices P, Q and matrix L with compatible dimensions, such that

$$P(M - LC) + (M - LC)^T P = -Q$$

(2) In the above equation, it holds that $\lambda_{\min}(Q) > 2\gamma_1\gamma_2\lambda_{\max}(P)$,
(3) There exists $h(x, u)$ such that $h(x, u)\, C = f^T(x, u)\, P$.

Then an adaptive observer for the system can be designed.

2. In Inc. (5.6.1), the solution of the equation $h(x, u)\, C = f^T(x, u)\, M^T P$ may not exist. Can you give a condition for the existence?
3. Prove that if φ(t) is a persistently excitation, then φ(t) is bounded on the interval $[0, \infty)$.
4. Please prove Theorem 5.6.1.
5. Please prove that $k_1 I \le \int\limits_t^{t+M} \varphi(\tau)\varphi^T(\tau)\, d\tau \le k_2 I$ if and only if for any unit vector α, the following inequalities hold

$$k_1 \le \int\limits_t^{t+M} \left[\alpha^{\mathrm{T}}\varphi(\tau)\right]^2 d\tau \le k_2.$$

6. Let

$$y(t) = \begin{bmatrix} a \\ b \sin \omega_0 t \end{bmatrix}.$$

Give conditions for a, b, ω_0 under which $y(t)$ is a persistent excitation.

7. Consider the following system

$$\begin{aligned}
\dot{x} &= \begin{bmatrix} -1 & 0.5 & 0 \\ 0 & 1 & 1 \\ 0.5 & 0.6 & 0 \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u + \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \omega + \begin{bmatrix} 0 \\ u^2 \\ 0 \end{bmatrix} \theta, \\
y &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \\
\omega &\in v(\lambda),
\end{aligned}$$

where the set-valued mapping $v(\lambda)$ is the same as that in Example 5.6.1, but $\lambda = x_1 + x_2 + x_3$. Please design an adaptive linear observer for the system where $u(t)$ is chosen as $u(t) = 2\sin t + 2$, and then verify the effectiveness by simulation.

References

Anderson BDO (1967) A system theory criterion for positive real matrices [J]. SIAM J Control 5(2):172–182

Chen CT (1984) Linear systems theory and design [M]. Rinehart and Winston, New York

de Bruim JCA, Doris A, Van de Wouw et al (2009) Control of mechnical motion systems with non-collocation of actuation and friction: a Popov criterion approach for input-to-state stability and set-valued nonlinearities [J]. Automatica 45:405–415

Han Z (1993) (A,B) Characteristic subspaces of linear systems and decentralized control of large scale systems [M]. Science Press, Beijing

Huang C-H, Ioannon PA, Maroulas J et al (1999) Design of strictly positive real systems using constant output feedback [J]. IEEE Trans Autoa Control 44(3):569–573

Huang J, Han Z, Cai X et al (2011) Adaptive full-order and reduced-order observer for the Luré differential inclusion system [J]. Commun Nonlinear Sci Numer Simul 16(7):2869–2879

Ly JH, Safonov MG, Ahmad F (1994) Positive real Parrott Theorem with application to LMI control synthesis [C]. Proc ACC Baltimore, MD, pp 50–52

Rochafellar RT (1970) Convex analysis [M]. Princeton University Press, Princeton

Smirnov GV (2002) Introduction to the theory of differential inclusions [M]. AMS, Providence

Sontag ED (1989) Smooth stabilization implies coprime factorization [J]. IEEE Trans AC 34(4):435–442

Index

Z. Han et al., *Theory of Control Systems Described by Differential Inclusions*,
Springer Tracts in Mechanical Engineering, DOI 10.1007/978-3-662-49245-1