Global Tile Attractor of Second Order Single-Bit Sigma-Delta Modulation

by

SIDONG ZENG

A dissertation submitted to the Graduate Faculty in Electrical Engineering
in partial fulfillment of the requirements for the degree of Doctor of Philosophy,
The City University of New York

2008

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

This manuscript has been read and accepted for the
Graduate Faculty in Electrical Engineering in satisfaction of the
dissertation requirements for the degree of Doctor of Philosophy.

4/21/2008	Truong-Thao Nguyen
Date	Chair of Examining Committee
4/21/2008	Mumtaz K. Kassir
Date	Executive Officer

Professors:

Truong-Thao Nguyen

Kenneth Sobel

Jizhong Xiao

Lucas Parra

Sinan Güntürk

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Global Tile Attractor

of Second Order Single-Bit Sigma-Delta Modulation

by

Sidong Zeng

Adviser: Professor Truong-Thao Nguyen

As a widely used technology of modern A/D data conversion, high resolution Sigma-Delta ($\Sigma\Delta$) modulation keeps simple structure. However, its most fundamental mechanisms are difficult to analyze rigorously without modeling assumptions. This is due to the embedded nonlinear feedback. Following dynamical system approach, the goal of this thesis is to characterize the attractor of second order single-bit $\Sigma\Delta$ modulation. More precisely, we prove that the global attractor is a single tile. This work enables the rigorous analysis of the quantization error spectrum without any modeling assumption. The attractor is known as the finite union of disjointed tiles (up to a 0-measure set) so far under stability conditions.

In this thesis, a framework based on Lyapunov functions is established to prove that the system is globally stable. A family of Lyapunov functions are discovered. Trapping sets are generated systematically in a conceptual and direct way. The dynamical behavior inside a positively invariant set is also studied. As a theorem, if a positively invariant set can be split into two sets by the graph of a function and one of the subsets is positively invariant, such subset is automatically a trapping set. This technique applies to trapping sets generated under the framework based on Lyapunov functions. Smaller trapping sets are obtained such that none of them is possible to contain two disjointed tiles. This implies that the attractor is a tile.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Truong-Thao Nguyen, for his guidance, insightful comments and collaboration in this work. He continually stimulated me to think in high level and greatly assisted me in scientific writing.

I am also grateful to my wife for her constant support and encouragement.

Contents

List of Figures

Chapter 1

Introduction

In modern electronics, digital circuits are omnipresent due to the technological development of very large scale integration circuits (VLSI). Because of the analog nature of the real world, however, analog-to-digital (A/D) converters are needed as front end signal processing. The circuit integration of these A/D converters faces the inherent accuracy limitation of analog circuits. A technique of A/D conversion, called Sigma-Delta ($\Sigma\Delta$) modulation, has become popular due to its robustness to analog circuit imperfections.

1.1 Introduction of $\Sigma\Delta$ modulation

An analog signal, such as speech or image, is a continuous-time signal with continuous amplitude, while a digital signal has discrete-time and discrete amplitude. As show in Figure 1.1(a), through an A/D converter, an analog signal $x(t)$ is first converted to a discrete-time but continuous-amplitude signal $x[n]$, then becomes a digital signal $\hat{x}[n]$. The first process is called *sampling* and the second is called *amplitude quantization* or *quantization*. According to the Shannon's sampling theorem, no information is lost after sampling if the sampling frequency f_s is equal to or larger than the Nyquist rate $2f_0$, where f_0 is the maximum frequency of the analog signal. Assume that $x(t)$ is

bandlimited, it can always be uniquely reconstructed by *sinc* interpolation from its sampled version $x[n]$ if it is taken by Nyquist rate.

However, information is lost permanently after quantization. It works like rounding a real number. The difference between the input and the output of a quantizer is called the *quantization error*. When signal is sampled at Nyquist rate, the only way to reduce the quantization error is to reduce the quantization step size. But this will increase the number of quantization levels. In an electronic system, it increases the circuits complexity.

Modern high resolution A/D conversion is based on oversampling where the sampling frequency is chosen to be greater than the Nyquist rate. Comparing to the A/D converter operated at Nyquist rate (Figure 1.1(a)), in the oversampled A/D converter (Figure 1.1(b)), a lowpass filter (followed by a down sampler) is added to recover information lost in the quantization process. Precisely, it filters out the noise located outside the signal band.

In the oversampled situation, various of circuits are developed to replace the basic amplitude quantizer. Among these designs, $\Sigma\Delta$ modulation has been developed for high resolution oversampled A/D conversion which was first introduced in 1962 [1]. It is widely used in VLSI because of its simplicity and its robustness against circuit imperfections.

Theoretically, although $\Sigma\Delta$ modulation is implemented by mixed analog and digital circuits, it has an equivalent discrete-time description [1], which is shown in Figure 1.2. In the simplest configuration of $\Sigma\Delta$ modulation, F is nothing but a delay unit. As can be seen in the figure, $\Sigma\Delta$ modulation in this case consists of subtracting to

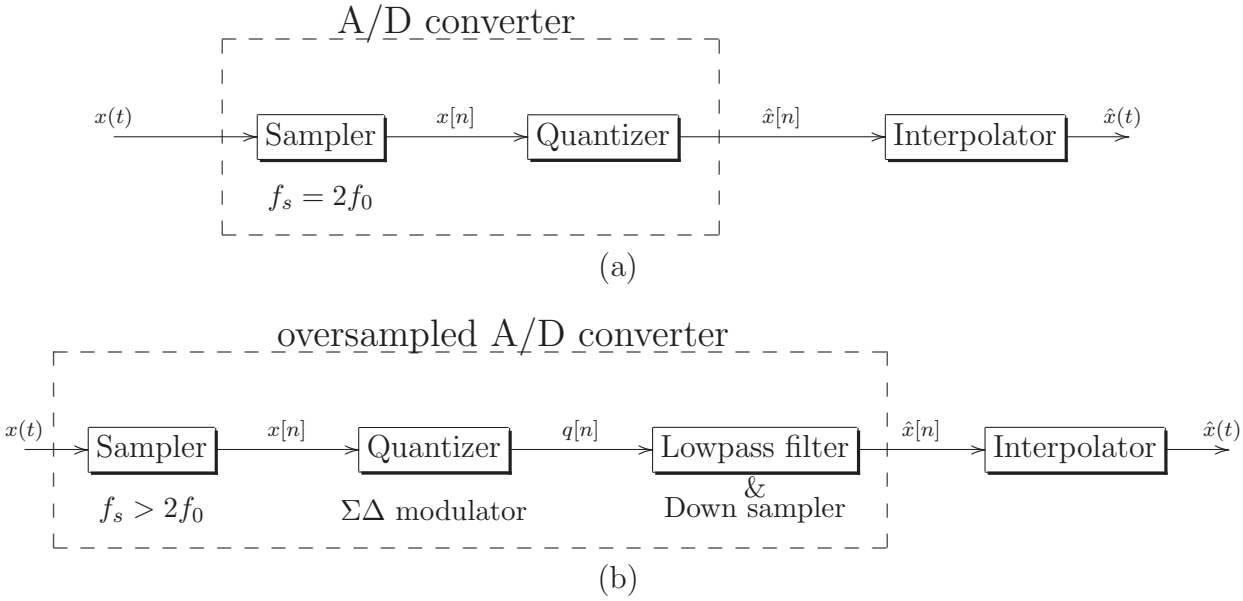


Figure 1.1: The principle of A/D conversion: (a) Nyquist rate; (b) oversampled rate.

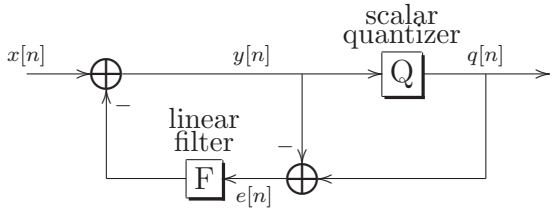


Figure 1.2: Diagram of ΣΔ modulation

the input a delayed version of the error made by the scalar quantizer. This is the known principle of *error diffusion* and allows the use of *coarse quantizers*.

1.2 Classic analysis

Although ΣΔ modulation is easy to understand intuitively, it is difficult to analyze rigorously because a nonlinear operator, the quantizer, is embedded in the feedback loop. No existing analytical tool can fully analyze such nonlinear feedback system. Traditionally, this system is modeled based on the assumption of the quantization error $e[n]$. In the classic linear model [1],[7], the quantizer is treated as an additive

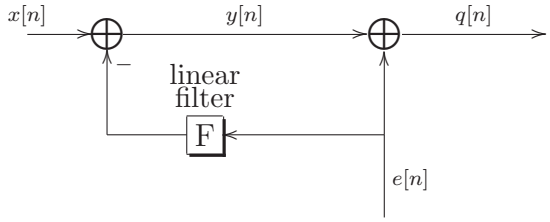


Figure 1.3: Classic linear ΣΔ modulator model assuming $e[n]$ is a white noise.

and independent source of white noise. The linearized system is shown in Figure 1.3. In the z-domain, the system transfer function is given by:

$$Q(z) = X(z) + H(z)E(z), \tag{1.1}$$

where

$$H(z) := 1 - F(z).$$

The modulator output is the sum of its input and a filtered version of the quantization error. The filter $H(z)$ is called the *noise-transfer function* (NTF). When designing a lowpass ΣΔ modulation, we consider a family of NTF in the form

$$H(z) = \frac{(1 - z^{-1})^k}{1 + d_1 z^{-1} + \cdots + d_k z^{-k}}$$

that eliminates DC components of quantization error most. This is a high-pass filter. In first order ΣΔ modulation, F is a pure delay unit. Then $H(z) = 1 - z^{-1}$. The process of filtering the quantization error is called *noise-shaping*. It is illustrated in Figure 1.4(b). With noise-shaping, the quantization error remained in the signal band is reduced dramatically compared to the case of simply using the amplitude quantizer (Figure 1.4(a)).

However, the classic linear model of ΣΔ modulation is proven to be inaccurate [2], especially when coarse quantization is used. The main issue is whether the quanti-

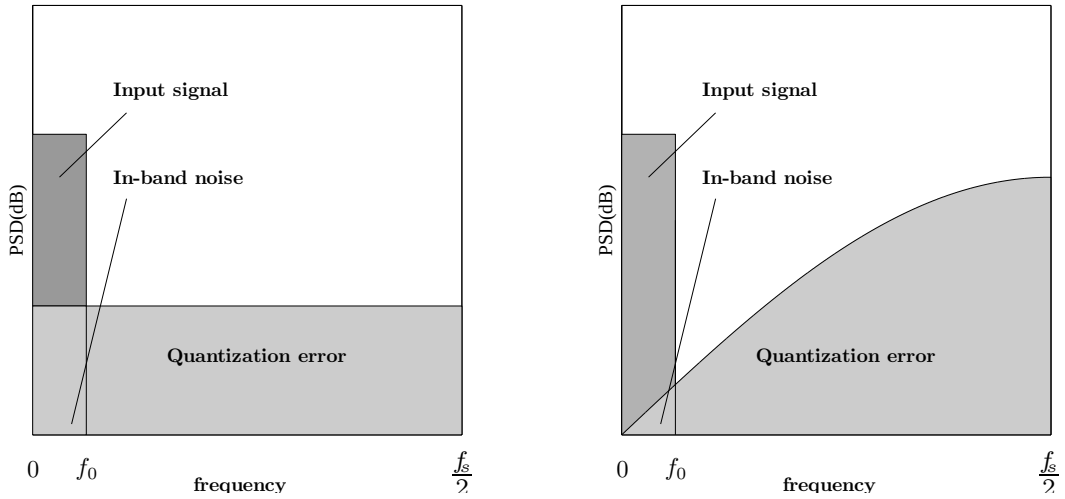


Figure 1.4: Power spectral density of quantized signal assuming the quantization error is white noise. (a) use the amplitude quantizer; (b) use a first order $\Sigma\Delta$ modulator to replace the amplitude quantizer.

zation error is a white noise. We will find it out in the next section with rigorous analysis.

1.3 Dynamical system analysis

In this section, we introduce the reader to the rigorous analysis of $\Sigma\Delta$ modulation with an example, the simplest first order single-bit case. We follow the method in [2], to model a modulator as a nonlinear dynamical system rather than a linearized approximation. The dynamical system is deterministic. We will introduce concepts and terminologies from dynamical system analysis just before we use them in order to analyze the example. Some general results on $\Sigma\Delta$ modulation are presented as well.

From the block diagram in Figure 1.2, the system of first order single-bit $\Sigma\Delta$ modu-

lation is described with the following equations:

$$\begin{array}{l} y[n] = x[n] - e[n-1], \\ e[n] = q[n] - y[n], \\ q[n] = \left\{ \begin{array}{c} \frac{1}{2}, y[n] \geq 0 \\ -\frac{1}{2}, y[n] < 0 \end{array} \right. . \end{array} \tag{1.2}$$

Note that the linear filter F is a pure delay unit and we have implicitly normalized the signal amplitude scale so that the quantization step size is 1. Let us define

$$u[n] := \ -e[n-1]. \tag{1.3}$$

Equations in (1.2) then imply that

$$u[n+1] = u[n] + x[n] - q[n] \tag{1.4}$$

with

$$q := \left\{ \begin{array}{c} \frac{1}{2}, x+u \geq 0 \\ -\frac{1}{2}, x+u < 0 \end{array} \right. . \tag{1.5}$$

One can see that $u[n+1]$ is a function of $u[n]$ for a given input $x[n]$. Explicitly,

$$u[n+1] = \mathcal{M}_{x[n]}(u[n]),$$

where

$$\mathcal{M}_x(u) := u + x - q, \tag{1.6}$$

This is a nonlinear transfer function from $\mathbb{R}$ to $\mathbb{R}$. It is also called a *mapping* from a dynamical system perspective. One can see that the state variable is transferred by different mappings at different instance of n when $x[n]$ is a time varying input. In other words, the dynamical system is fully described by a sequence of mappings.

We simplify our analysis further by assuming that the input $x[n] = x$ is constant. This is the asymptotic situation of a bandlimited signal sampled at an infinitely high rate.

Although we only study one mapping at a time, we still need to analyze all mappings for all possible constant inputs. Usually, the *configuration* of a system is a set of system parameters, while the system inputs are excluded. In this thesis, however, we generalize the meaning of a *configuration* from the system to the mapping since a mapping in (1.6) depends on both the system parameters and the input x. In other words, a configuration defines a mapping. We sometimes use *configurations* instead of *mappings* to indicate that the difference between mappings is the coefficients not the structure. We rewrite (1.6) in the form of a piecewise mapping

$$\mathrm{M}(u) = \begin{cases} M_0(u), u \in \Omega_0 \\ M_1(u), u \in \Omega_1 \end{cases}, \tag{1.7}$$

$$\text{with } \mathrm{M}_i(u) := u + x_i, \tag{1.8}$$

where

$$x_0 := x + \tfrac{1}{2} \quad \text{and} \quad x_1 := x - \tfrac{1}{2}.$$

$$\Omega_0 := \big\{u \in \mathbb{R} : u < -x\big\} \quad \text{and} \quad \Omega_1 := \big\{u \in \mathbb{R} : u \geq -x\big\}.$$

In this case, a configuration is specified by a given x.

In general, the system mapping of $\Sigma\Delta$ modulation appears to be piecewise affine when the state variables are chosen to be the integrator outputs. The mapping in (1.7) of first order single-bit $\Sigma\Delta$ modulation is an example. The state space spanned by the state variables is called the *phase space*. It is usually a vector space. We call a vector in a phase space a *point*. A mapping transfers a point to another. The collection of points is called a *set* or a *region*. For any set S within the phase space, we call the set $S_f := \mathrm{M}(S)$ the *forward image* of S, and call the set S_b such that $\mathrm{M}(S_b) = S$ the *backward image*. A point $\mathbf{u}$ in the phase space is called a *fixed point* if $\mathrm{M}\mathbf{u} = \mathbf{u}$.

When M is replaced by M^n for any $n \in \mathbb{N}$, S_f and S_b are called the n^{th} order forward and backward image, respectively. The definition of forward and backward image applies to any set S even if it contains only a single point. In that case, since the mapping is piecewise-affine, the forward image must be a single point. However, the backward image may be a set of multiple points. Note that the system is a discrete time system. So the forward images of any given point at all orders are a sequence of points. This sequence is called the *trajectory* of the given point. For a given mapping and an initial point, the forward images, backward images and trajectory of the point are determined. Although these definitions are simple, they play important roles of formalizing the dynamical system analysis.

Once the mapping is explicitly defined, a question from a dynamical system perspective arises: is this system stable? Since *stability* may have different meanings in different areas, we formalize it in this thesis as following. For a given mapping M and a set S in the phase space, if there exists a bounded set B such that for all $\mathbf{u} \in S$, there exists $n_0 \in \mathbb{N}$, $\forall n > n_0$, $\mathrm{M}^n\mathbf{u} \in B$, we call S the *stable region* of M. If the stable region is the whole phase space, we conclude that the mapping M is globally stable. So an approach of proving stability for any given mapping is to find a stable region. But before we analyze the example, we introduce some related definitions. Note that all definitions are based on a given mapping. If a set happens to be the forward image of itself, the set is called an *invariant set*. We explicitly recall the definition of the forward image and have S to be an invariant set if $\mathrm{M}(S) = S$. The simplest invariant set is a set which contains only a fixed point. It is clear that the trajectory of any point chosen from an invariant set is a subset of that invariant set. When the trajec-

tory of any point in a superset of an invariant set has nonempty intersection with that invariant set, the invariant set is called an *attractor*[1] of the superset. This implies that any attractor is an invariant set. The *global attractor* is an attractor which has nonempty intersection with any trajectory in the phase space. In other words, any point will be mapped into the global attractor with iterations. The stability of the mapping can be proven by finding bounded invariant sets or attractors. However, for many mappings, the attractor, even an invariant set is difficult to derive analytically. We then consider some sets with weaker conditions. A *positively invariant set* is a set such that its forward image is a subset of it. By comparing this with the definition of stability, one can easily see that a positively invariant set is a stable region if it is bounded. A small positively invariant set is also used to estimate the invariant set bounds. For a positively invariant set, if points in a superset of it will be mapped into it, it is called a *trapping set*. Now we have the following propositions to analyze the first order case.

Proposition 1.3.1 *Assume that* $|x| > \frac{1}{2}$. *The stable region of the mapping in* (1.7) *is an empty set.*

This is proven in Appendix-A.1.

Assume $x = \frac{1}{2}$, we have $x_1 = 0$ and $x_0 = 1$. Then for any $u \in \Omega_1$, $u = \mathrm{M}(u)$ is a fixed point. For any $u \in \Omega_0$, $\mathrm{M}(u) = u + 1$. And the trajectory of u is the sequence $\{u, u+1, u+2, \ldots, u+n, u+n, \ldots\}$, where n is the smallest integer such that $u + n \in \Omega_1$. In other words, the state variable u increases by 1 at each step with

[1] An attractor is not defined as the smallest unit that cannot be decomposed into two or more attractors with distinct region of attraction. So the meaning of the term here is slightly different from that in some dynamical system theories.

iterations until it falls into Ω_1 and becomes a fixed point. A similar result is obtained when $x = -\frac{1}{2}$. In both cases, the dynamics is clearly simple. Therefore, we are mostly interested in the case where $x \in (-\frac{1}{2}, \frac{1}{2})$.

Proposition 1.3.2 *Assume that* $x \in (-\frac{1}{2}, \frac{1}{2})$ *and* $a \leq -\frac{1}{2} < \frac{1}{2} \leq b$,

$$\mathrm{M}([a,b)) = \left[\min(a + x_0, -\tfrac{1}{2}), \max(b + x_1, +\tfrac{1}{2})\right). \tag{1.9}$$

This is proven in Appendix-A.2.

Proposition 1.3.3 *Assume that* $x \in (-\frac{1}{2}, \frac{1}{2})$, *an interval* $[a,b) \in \mathbb{R}$ *is positively invariant if and only if* $a \leq -\frac{1}{2} < \frac{1}{2} \leq b$.

This is proven in Appendix-A.3.

Proposition 1.3.4 *Assume* $x \in (-\frac{1}{2}, \frac{1}{2})$ *and* $a \leq -\frac{1}{2} < \frac{1}{2} \leq b$. *There exists* $n_0 \in \mathbb{N}$ *such that* $\forall n \geq n_0$,

$$\mathrm{M}^n\left([a,b)\right) = [-\tfrac{1}{2}, \tfrac{1}{2}). \tag{1.10}$$

This is proven in Appendix-A.4.

Under certain conditions, Proposition 1.3.2 gives the explicit equation how an interval is transformed to another. Proposition 1.3.3 then shows a sufficient and necessary condition of an interval to be a positively invariant set. Proposition 1.3.4 finally gives a global attractor $[-\frac{1}{2}, \frac{1}{2})$. Hence it proves the stability of the mapping.

At this moment, let us check whether the quantization error is a white noise. We proceed with $u[n]$ in order to plot the amplitude spectrum of the quantization error

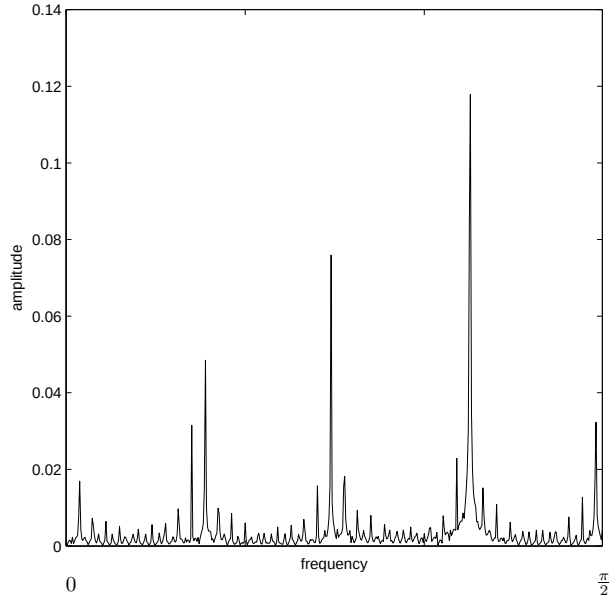


Figure 1.5: Numerical example of spectrum of $u[n]$. We take $u[0] = 0$ and $x = .12345$, and use (1.7) to generate $u[n]$. After discarding the first $10k$ points, we use the next 1024 points to perform FFT and plot the amplitude spectrum.

because $u[n]$ has the identical amplitude spectrum as $e[n]$ according to (1.3). Using the mapping in (1.7), we obtain a sequence of $u[n]$ and show its amplitude spectrum in Figure 1.5. Since the spectrum is not flat, the quantization error is not a white noise. Hence the linearized model is inaccurate.

Let us continue to derive more rigorous results. We are motivated by the following results

(i) $\mathrm{M}_0(u) = \mathrm{M}_1(u) + 1$ from (1.8),

(ii) the attractor $[-\frac{1}{2}, \frac{1}{2})$ is an interval of length 1.

So we define the modulo function mod as the unique 1-periodic function such that it is the identity function in $[-\frac{1}{2}, \frac{1}{2})$. Explicitly

$$\mathrm{mod}(v) := u, \text{ where } u \text{ is the unique value of } [-\tfrac{1}{2}, \tfrac{1}{2}) \text{ such that, } u - v \in \mathbb{Z}.$$

Then we have

$$\mathrm{mod}(\mathrm{M}_0(u)) = \mathrm{mod}(\mathrm{M}_1(u)).$$

It follows from (1.7) that

$$\mathrm{mod}(u) = u, \tag{1.11}$$

$$\mathrm{M}(u) = \mathrm{mod}(\mathrm{M}_0(u)) = \mathrm{mod}(\mathrm{M}_1(u)). \tag{1.12}$$

One does not loose generality by considering an initial state $u[0] \in [-\frac{1}{2}, \frac{1}{2})$, which is in the global attractor. Define a trajectory which is the sequence:

$$u[n] := \mathrm{M}^n(u[0]), \text{ where } n \in \mathbb{N}.$$

According to Proposition 1.3.3, for any $n \in \mathbb{N}$, $u[n] \in [-\frac{1}{2}, \frac{1}{2})$. It follows from (1.11) that

$$u[n] = \mathrm{mod}(\mathrm{M}_0^n(u)). \tag{1.13}$$

Let us define

$$v[n] := \mathrm{M}_0^n(u).$$

Then

$$v[n+1] = v[n] + x_0 = v[n] + x + \tfrac{1}{2},$$

$$u[n] = \mathrm{mod}(v[n]).$$

We define

$$E_{\Sigma\Delta}(z) := H(z)E(z).$$

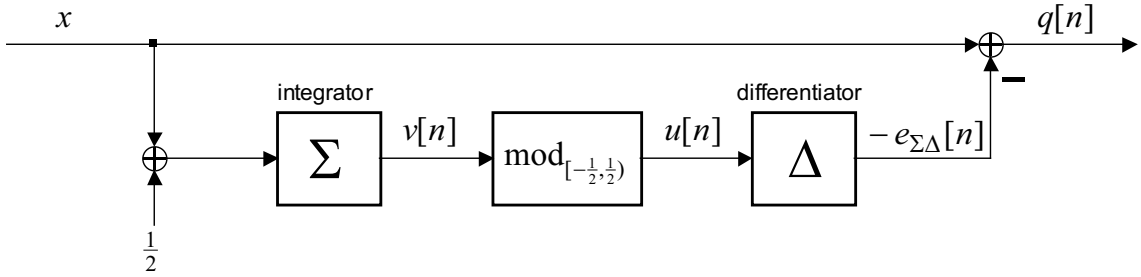


Figure 1.6: First order $\Sigma\Delta$ modulator equivalent block diagram.

where $H(z)$ is the same as in (1.1). Then we have

$$q[n] = x + e_{\Sigma\Delta}[n],$$

$$e_{\Sigma\Delta}[n] = e[n] - e[n-1].$$

It follows from (1.3) that $e_{\Sigma\Delta}[n] = u[n] - u[n+1]$. These equations form an equivalent diagram as shown in Figure 1.6. In the diagram, the state variable $u[n]$ is obtained from $v[n]$ through a nonlinear but memoryless modulo operator, where $v[n]$ is nothing but the output of a linear system, which is the discrete-time integrator. The system is decomposed and the analysis is simplified.

1.4 Literature review

The rigorous analysis of $\Sigma\Delta$ modulation was first introduced by Gray in [2]. He clear showed that a more accurate analysis of $\Sigma\Delta$ modulation is achieved by approaching the modulator in nonlinear dynamical system way. Note that this is the first paper, but only paper where Gray recognizes the dynamical system aspect of $\Sigma\Delta$ modulation. Some results of $\Sigma\Delta$ modulation based on nonlinear methods are summarized in [17]. Those results are from the work of Gray and others. Given the date of 1995 of this paper, we know that $\Sigma\Delta$ modulation has been studied as a dynamical system for quite a long time.

We are mainly interested in two issues about $\Sigma\Delta$ modulation by approaching it in dynamical system way. One is the stability. The other is the structure of attractors, invariant sets and positively invariant sets. This information enables us to analyze the quantization error accurately.

The stability of first order $\Sigma\Delta$ modulation has been sorely studied since [2] was published. Many mappings not limited to single-bit $\Sigma\Delta$ modulation and constant inputs have been studied. In [8], for example, first order $\Sigma\Delta$ modulation with multilevel quantization is rigorously analyzed even if circuit nonidealities such as integrator leak, integrator gain mismatch, and comparator offset are considered. There are also remarkable results of stability in second order $\Sigma\Delta$ modulation. In [11], the stability of second order $\Sigma\Delta$ modulation with constant input is studied through *limit cycles*. The n^{th} order forward image of any given point defines a trajectory. A limit cycle is defined as that trajectory when n tends to infinity. With the help of limit cycles, the upper bound of the quantizer input, which is a state variable and has the most volatile value, is derived as the function of the input. This implies the system stability. From the definition of a limited cycle, one can see that it is actually an invariant set. The concept of *positively invariant set* was introduced to $\Sigma\Delta$ modulation in [18] and [16]. An analytical set which is bounded by two parabolic segments is proven to be positively invariant for general second order $\Sigma\Delta$ modulator with constant inputs. An algorithm was developed to find a positively invariant convex set numerically in [16]. This algorithm-based technique for finding positively invariant set is generalized from the second order case to higher order $\Sigma\Delta$ modulation [20]. This technique can be applied to verify $\Sigma\Delta$ modulator designs. Similar positively invariant sets which are

bounded by parabolic curves are obtained in [9],[15], and [25]. It has been proven that global stability does not hold for arbitrary time varying inputs since unstable examples were presented in [9] and [25]. Dynamical system with bandlimited single inputs is studied and the stable condition is derived [9]. With the help of positively invariant set, the bounds of the state variables are estimated in [9] and [16]. Contrary to lower order cases, there are few results in the third or higher order $\Sigma\Delta$ modulation due to the difficulty. The bounded of state variables are derived based on those positively invariant sets. Wang developed a *geometric model*, a dynamical system, to analyze third order $\Sigma\Delta$ modulation [12]. He applied the dynamical system theory to study the stability for the basic architecture of third-order single-path $\Sigma\Delta$ modulation and constant inputs [13]. The stable conditions are obtained by applying the *bifurcation theory*. Results are confirmed numerically, because the bifurcation theory normally applies to continuous mapping but the system of $\Sigma\Delta$ modulation is discrete.

One can see that the theoretical concepts of positively invariant sets and trapping sets are exploited, but the refine analysis available until now has been mostly based on explicit algebraic inspections of the discrete dynamics. This approach often yields complicated equations or discrete recursive reasonings, from which it is difficult to extract high-level and general guidelines on the state behaviors. A more physical and analytic approach to the state behavior of $\Sigma\Delta$ modulators is to introduce Lyapunov functions. This idea was first suggested in $\Sigma\Delta$ modulation in [24]. The classic technique of Lyapunov functions is used to prove the stability of the solutions to a differential equation, without solving it explicitly [5]. In physical systems, Lyapunov functions are usually extracted from the system's potential energy, which typically

decreases with time and converges to a certain bounded region in the stable case. The notion of potential energy does not naturally exist in $\Sigma\Delta$ modulation, but one can always propose synthetically generated Lyapunov functions of the state space and evaluate their respective abilities to be minimized by the iterations of the $\Sigma\Delta$ state mapping. However it is difficult to find proper Lyapunov functions which can generate positively invariant sets as refine as that based on explicit algebraic derivation.

The structure of an attractor has also been studied. But there are little results for higher orders $\Sigma\Delta$ modulation. Besides the work of Gray in [2], first order $\Sigma\Delta$ modulation is analyzed with techniques from dynamical system theory in [10]. A conclusion is that almost all trajectories converge asymptotically to a fixed point or a periodic orbit. Similar dynamical equations to second order $\Sigma\Delta$ modulation from the digital phase-locked loops (DPLL) is studied in [22] and [23]. Attractors have been derived. The structure of a attractor is simply bounded by two continuous functions. The attractor is called a *belt* in the papers. Although the mappings obtained from DPLL and $\Sigma\Delta$ modulation are similar, they are in different configurations. According to experiments, the attractor in the configuration of $\Sigma\Delta$ modulation is usually not a belt. Followed by an outstanding discovery, the *tiling phenomenon*, many results are derived. We will show the details in the following section.

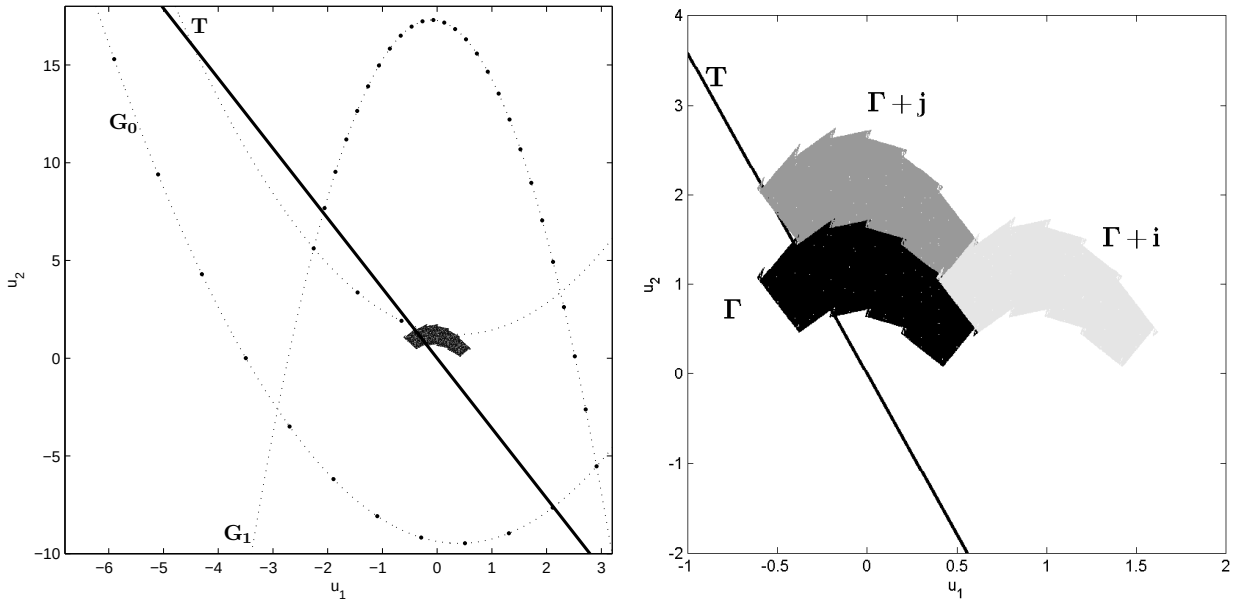
1.5 Tiling phenomenon and nonlinear feedback loop resolution

Contrary to the first order case (Section 1.3), little attractors in second order $\Sigma\Delta$ modulation has been derived explicitly. Meanwhile, rigorous analysis by Gray [2] is

generalized without deriving attractors explicitly [28]. The most important property of the attractor in the first order case is that the attractor is an interval of length 1 which allows the definition of the modulo function. And with the help of the modulo function, nonlinear feedback loop is decomposed into a linear feedback loop followed by a nonlinear but memoryless operation as shown in (1.13). An outstanding phenomenon was discovered in various second order $\Sigma\Delta$ modulation configurations. Every attractor is a single tile. Two examples are shown in Figure 1.7 and Figure 1.8. This enables us to predict properties of the attractor in the second order case without explicit derivation. This is called the *tile phenomenon*. The formal definition of a tile will be given later. But like the interval $[-\frac{1}{2}, \frac{1}{2})$, which is a tile in $\mathbb{R}$, a tile in the $2D$-space is a set such that (i) it has no intersection with its integer-vector shifted version; (ii) the union of all its integer-vector shifted versions cover the entire space. The unit square is an example of tile. Other examples are shown in Figure 1.7(b), 1.8, and 1.9. One can see how a tile is *tiling* the space.

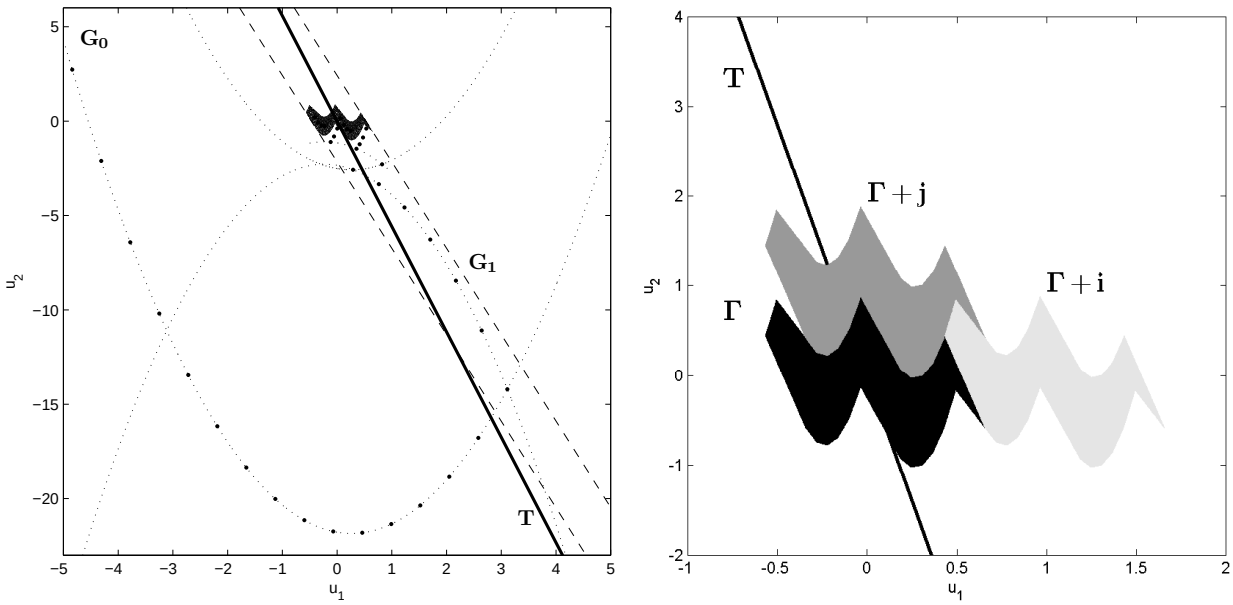
In second order $\Sigma\Delta$ modulation, the state space is the $2D$ vector space $\mathbb{R}^2$. As mentioned before, the difficulty of analyzing $\Sigma\Delta$ modulation is due to the nonlinear operation in a feedback loop. Assume that the attractor is a single tile Γ, the nonlinear feedback loop may be decomposed to a linear feedback loop followed by a $2D$-modulo operation. This idea is generalized from the first order case as shown in equation (1.11) and (1.13). Similar to the first order case, a $2D$ tile Γ defines a nonlinear modulo operation over $\mathbb{R}^2$. For any $\mathbf{u} \in \mathbb{R}^2$, the modulo operation is defined as

$$\mathrm{mod}_\Gamma(\mathbf{u}) := \mathbf{v}, \tag{1.14}$$



(a) (b)

Figure 1.7: Example 1 of dynamics behavior: (a) The attraction of dynamics (b) The tile attractor Γ (marked by black dots) and its vertical and horizontal unit-shifted versions.



(a) (b)

Figure 1.8: Example 2 of Dynamics behavior: (a) The attraction of dynamics (b) The tile attractor Γ (marked by black dots) and its vertical and horizontal unit-shifted versions.

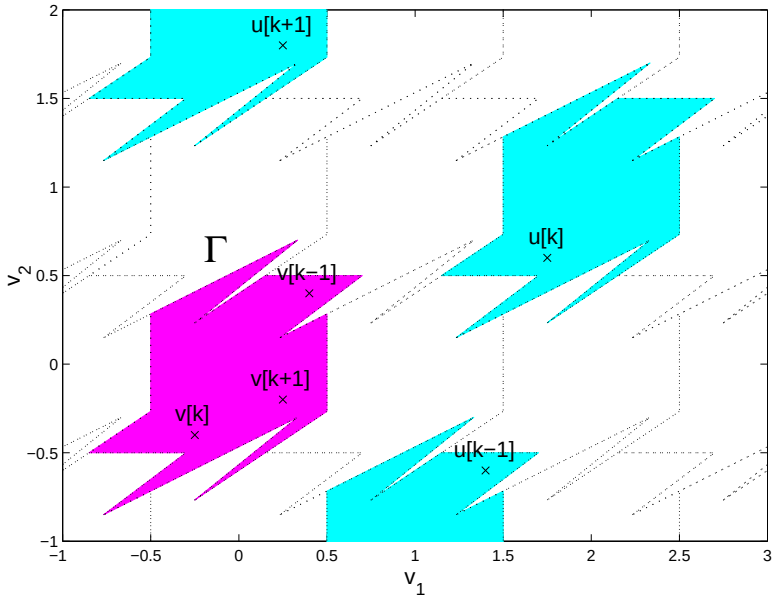


Figure 1.9: Illustration of modulo operation, $\mathbf{v}[n] := \mathrm{mod}_\Gamma(\mathbf{u}[n])$.

where $\mathbf{v} \in \Gamma$ is the unique point such that $\mathbf{u} - \mathbf{v} \in \mathbb{Z}^2$. For any $\mathbf{v} \in \Gamma$, we have

$$\mathrm{mod}_\Gamma(\mathbf{v}) = \mathbf{v}.$$

An example of modulo operation in $\mathbb{R}^2$ is shown in Figure 1.9. Points $\mathbf{v}[n] = \mathrm{mod}_\Gamma(\mathbf{u}[n])$ are located in the tile Γ. A similar definition can be generalized to higher dimensional space.

As shown in [27], $\Sigma\Delta$ modulation with constant input can be described by a piecewise-affine mapping M. And similar to (1.13), the dynamics can be derived as:

$$\forall \mathbf{u} \in \Gamma,\ \forall n \in \mathbb{N},\ \mathrm{M}^n\mathbf{u} = \mathrm{mod}_\Gamma(\mathbb{L}^n\mathbf{u}). \tag{1.15}$$

where $\mathbb{L}$ is an affine-mapping. So with the knowledge of an invariant tile Γ of the mapping M, the whole system can be equally described as a linear system followed by a memoryless nonlinear operation. A general diagram derived in [27] is shown in Figure 1.10.

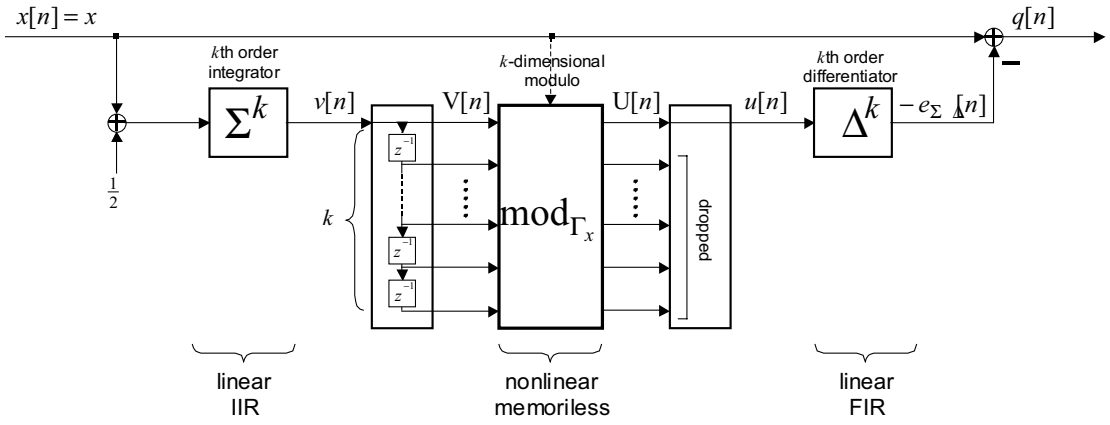


Figure 1.10: Equivalent $\Sigma\Delta$ modulation block diagram of k^{th} order $\Sigma\Delta$ modulation.

Since the tile attractor enables a new direction of rigorous analysis, we have the following questions:

(i) Is M guaranteed to yield a single invariant tile ?

(ii) If so, can we derive the tile ?

In this thesis, we will concentrate on the first question only. Our approach is based on an advanced theorem in [28], which proves that the attractor of the dynamical system is the finite union of disjointed tiles up to a measure zero set. This is the most advanced proven result on the tile phenomenon until now. For a piecewise-affine mapping, the threshold is used to partition the phase space into different regions. In each region the mapping is defined by a piece of affine mapping. As shown in Figure 1.7 and 1.8, for example, the threshold is the graph of a linear real function which is marked as T. For a long time, only single-tile attractors were observed for the mappings of second order single-bit $\Sigma\Delta$ modulation such that the threshold is a continuous real function. However, we show in Figure 1.11 an experimental counter-example which has two-tile invariant set with DC input and the thresholding function

is as simple as a continuous piecewise-affine function. So in this thesis, we will focus on linear thresholding functions, which is also the case of $\Sigma\Delta$ modulators used in applications. In fact, for mappings of second order single-bit $\Sigma\Delta$ modulation with linear thresholding functions, only single tile attractor has been observed but has not been rigorously proven yet. The goal of this thesis is to prove the single tile attractor in this situation.

1.6 Outline of this thesis

In this thesis, we concentrate on the analysis of second order single-bit $\Sigma\Delta$ modulation from a dynamical system approach. In Chapter 2, we derive the mathematical equations which describe the system. Instead of using the general block diagram shown in Figure 1.2, we proceed our derivation from a concrete block diagram of second order $\Sigma\Delta$ modulation. We then prove that it is equivalent to the general block diagram by showing transfer functions in the z-domain. This procedure clearly connects a mapping to an implementation. As the result of constant inputs case, the obtained mapping is indeed a piecewise-affine mapping from $\mathbb{R}^2$ to $\mathbb{R}^2$.

In Chapter 3, We follow the idea in [24] to establish a framework based on Lyapunov functions. General guidelines of finding Lyapunov functions for piecewise-affine mappings are presented. The procedure for obtaining positively invariant sets and trapping sets through these functions are derived. We then discover a family of Lyapunov functions and prove global stability of the system accordingly. We under this new framework obtain not only the existing results in more conceptual and direct way, but also more general results at the same time.

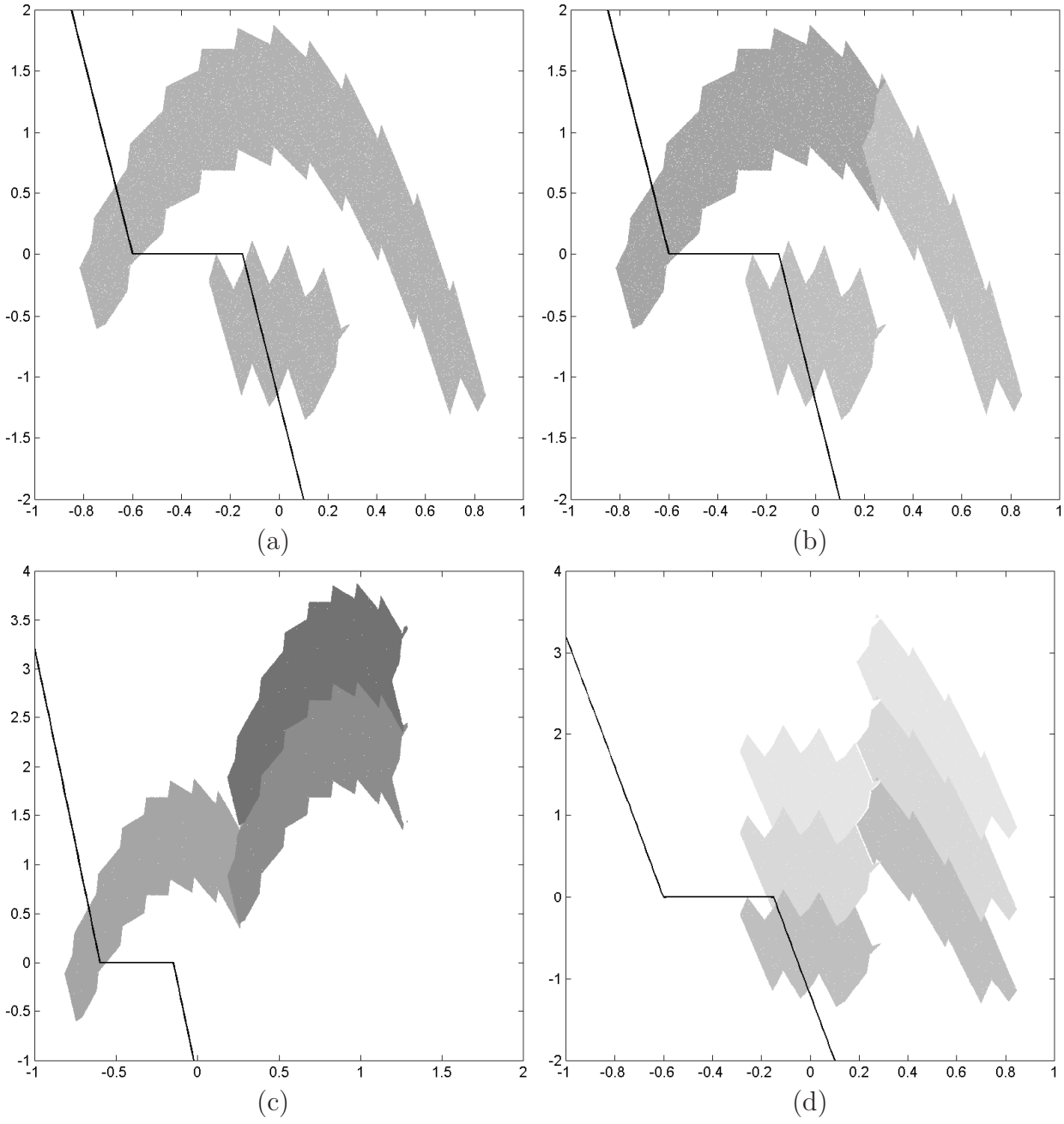


Figure 1.11: Two-tile invariant set with nonlinear but continuous piecewise-linear thresholding function: (a) The invariant set of mapping, with input $x = .35314$. The thresholding function has two pieces of slope -8 and the segment in the middle is connecting point (-.6,0) and (-.15,0); (b) By using two gray tones, we show that the set is the union of two disjointed tiles (the two-tile partition is not unique); (c) The dark gray set in (b) and its shifted version; (d) The light gray set in (b) and its shifted version.

However, it requires more information to analyze the quantization error accurately beyond the stability. Our goal in this thesis is to prove the attractor is a single tile. In Chapter 4, we formalize a method by using the set theory for checking whether a set contains more than one tile. With the help of a theorem in [28], we can immediately recognize a tile attractor inside a positively invariant set which can not contain more than one tile. By applying this method directly to the global trapping set obtained from the previous chapter, we can prove that the global attractor is a single tile under certain configurations. However, it only covers a limited range of configurations that we are interested in. We need to pursue our goal with new techniques.

Chapter 5 is devoted to study the details of the dynamical behavior inside a positively invariant set. The motivation is to develop a new technique with which we can find inside smaller trapping sets. We proceed by studying the process of a set being transformed into the attractor. In particular, we start from a set containing only two points $\mathbf{u}$ and $\mathbf{v}$. The corresponding trajectories are $\{\mathrm{M}^n\mathbf{u}\}_{n\in\mathbb{N}}$ and $\{\mathrm{M}^n\mathbf{v}\}_{n\in\mathbb{N}}$, respectively. And the difference of them $\{\mathrm{M}^n\mathbf{u} - \mathrm{M}^n\mathbf{v}\}_{n\in\mathbb{N}}$ is also a vector sequence. We discover that this sequence is impossible to be a mixture of integer vectors and non-integer vectors. It implies that integer vector difference is always preserved. Inspired by the tile phenomenon, we predict that the dynamical behavior of two points with integer vector difference is crucial. We then discover a theorem which enables us to find trapping sets from existing ones. According to this theorem, assume that a positively invariant set can be split into two sets by a graph of a real function and one of the subsets is positively invariant, the positively invariant subset up to a measure zero set is automatically a trapping set.

In Chapter 6, we combine techniques developed in Chapter 3, 5 and 4 to prove that the global attractor is a single tile. We obtain a global trapping set under the framework established with Lyapunov functions. Using the new tool in Chapter 5, we derive a smaller trapping set. The final conclusion is obtained by proving that such trapping set can not contain more than one tile.

We study global stability of timevarying inputs case in Chapter 7. Since global stability does not hold for arbitrary inputs, an input which is the sum of finite numbers of sinusoids is considered. A unique mapping defined by the DC component of the input and a sequence of translations defined by the AC components of the input can be used for modeling the dynamical system. At each instance, the state variables are transferred with that mapping followed by a translation. We under the framework with Lyapunov functions derive a sequence of trapping sets, thus prove that the system is globally stable. We also show the existence that an attractor is a sequence of tiles.

Chapter 2

The second order single-bit $\Sigma\Delta$ modulation

We derive dynamical mappings in this chapter for latter use. Mappings are obtained based on a general diagram of second order single-bit $\Sigma\Delta$ modulation. All mappings are formalized into piecewise-affine mappings over the $2D$ space $\mathbb{R}^2$. As mentioned in Section 1.3, a configuration, which includes explicit system parameters and the input, fully defines a mapping.

2.1 General equations

The general block diagram of a second order $\Sigma\Delta$ modulator in the cascade-integrator feed-forward (CIFF) structure is shown in Figure 2.1 as like in [3]. Using the *delta-sigma matlab toolbox* [4] designed by R. Schreier, one can obtain typical values of the coefficients a_1 and a_2 for certain optimization, which is $(a_1,\ a_2) = (0.7749,\ 0.2164)$.

In the z-domain, the transfer functions of the system in Figure 2.1 is derived to be

$$\begin{array}{l} U_1(z) = \frac{z^{-1}}{1-z^{-1}}(X(z) - Q(z)) \\ U_2(z) = \frac{z^{-1}}{1-z^{-1}}U_1(z) \\ Y(z) = X(z) + a_1U_1(z) + a_2U_2(z) \end{array} .$$

These equations imply that

$$Y(z) = X(z) + \frac{a_1(z^{-1} - z^{-2}) + a_2z^{-2}}{(1 - z^{-1})^2}(X(z) - Q(z)).$$

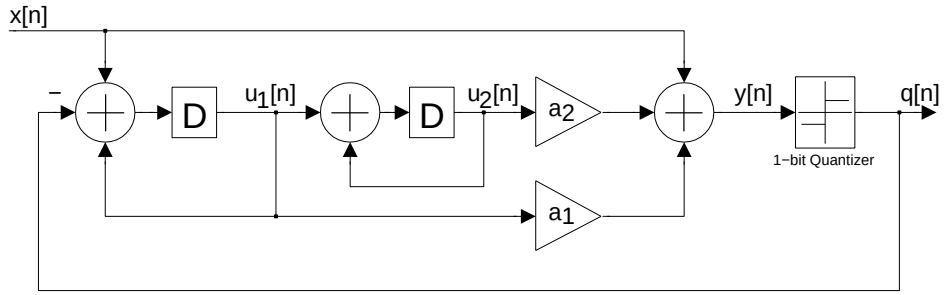


Figure 2.1: Second order $\Sigma\Delta$ modulator in CIFF structure.

It follows from the quantization error $E(z) = Q(z) - Y(z)$ that

$$Q(z) - E(z) = X(z) + \frac{a_1(z^{-1} - z^{-2}) + a_2 z^{-2}}{(1 - z^{-1})^2}(X(z) - Q(z)),$$

which leads to

$$Q(z) = X(z) + H(z)E(z)), \tag{2.1}$$

where

$$H(z) := \frac{(1 - z^{-1})^2}{1 + (a_1 - 2)z^{-1} + (1 - a_1 + a_2)z^{-2}}.$$

Since all zeros of $H(z)$ are located at $z = 1$ and holes are not, $H(z)$ is a highpass filter. So this system is equivalent to the general form shown in Figure 1.2.

From the block diagram, we derive the following state transfer equations

$$\begin{cases} u_1[n+1] &= u_1[n] + x[n] - q[n] \\ u_2[n+1] &= u_1[n] + u_2[n] \end{cases} \tag{2.2}$$

with

$$q[n] = \begin{cases} \frac{1}{2}, & y[n] \geq 0 \\ -\frac{1}{2}, & y[n] < 0 \end{cases}, \tag{2.3}$$

$$y[n] = a_1 u_1[n] + a_2 u_2[n] + x[n]. \tag{2.4}$$

We have implicitly normalized the signal amplitude scale so that the quantization step size is 1. In practise, $x[n] \in (-\frac{1}{2}, \frac{1}{2})$ is generally assumed.

2.2 Dynamics equations

Let us define the two-dimensional state vector

$$\mathbf{u} := (u_1, u_2)^\top,$$

with the following projection notation

$$(\mathbf{u})_1 := u_1, \text{ and } (\mathbf{u})_2 := u_2.$$

In this thesis, we simplify notation by replacing $(\mathbf{u})_1$ and $(\mathbf{u})_2$ with u_1 and u_2 , respectively, unless there is ambiguity. Such notation will be used for other vectors as well.

We then define $\mathbf{u}[n] := (u_1[n], u_2[n])^\top$. Equation (2.2) is rewritten in the vector form

$$\mathbf{u}[n+1] = \mathrm{L}\,\mathbf{u}[n] + (x[n] - q[n])\,\mathbf{i} \tag{2.5}$$

where

$$\mathrm{L} := \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{i} := (1, 0)^\top. \tag{2.6}$$

Meanwhile, from (2.3) and (2.4)

$$q[n] = \begin{cases} \frac{1}{2}, & t(\mathbf{u}[n]) \geq c_{x[n]} \\ -\frac{1}{2}, & t(\mathbf{u}[n]) < c_{x[n]} \end{cases}. \tag{2.7}$$

where

$$t(\mathbf{u}) := u_2 + s\,u_1, \quad \text{where } s := \frac{a_1}{a_2}, \tag{2.8}$$

$$c_x := -\frac{x}{a_2}. \tag{2.9}$$

We then partition $\mathbb{R}^2$ into two parts according to the function t,

$$\Omega_0 := \{\mathbf{u} : \quad t(\mathbf{u}) < 0\} \quad \text{and} \quad \Omega_1 := \{\mathbf{u} : \quad t(\mathbf{u}) \geq 0\} \tag{2.10}$$

The quantizer output is rewritten as

$$q[n] = \begin{cases} \frac{1}{2}, & \mathbf{u}[n] \in \Omega_1 + c_{x[n]}\mathbf{j} \\ -\frac{1}{2}, & \mathbf{u}[n] \in \Omega_0 + c_{x[n]}\mathbf{j} \end{cases} \tag{2.11}$$

By substituting (2.11) into (2.5), $\mathbf{u}[n+1]$ can be seen as a sole function of $\mathbf{u}[n]$ with a given input $x[n]$, explicitly,

$$\mathbf{u}[n+1] = \mathcal{M}_{x[n]}\,\mathbf{u}[n] \tag{2.12}$$

where

$$\mathcal{M}_x\,\mathbf{u} := \begin{cases} \mathrm{M}_1\mathbf{u}, & \mathbf{u} \in \Omega_1 + c_x\,\mathbf{j} \\ \mathrm{M}_0\mathbf{u}, & \mathbf{u} \in \Omega_0 + c_x\,\mathbf{j} \end{cases}, \tag{2.13}$$

$$\text{with} \quad \mathrm{M}_i\mathbf{u} := \mathrm{L}\,\mathbf{u} + x_i\mathbf{i}, \text{ where } x_0 := x + \tfrac{1}{2},\ x_1 := x - \tfrac{1}{2} \tag{2.14}$$

and

$$\mathbf{j} := (0,1)^\top. \tag{2.15}$$

Note that the mapping M_i is always implicitly associated with an input x. This x is defined according to the context of M_i. For example, in (2.13), the x is the same as the subscript of $\mathcal{M}_x$.

2.3 Basic properties of the mapping

Equation (2.14) implies that for all $x \in \mathbb{R}$,

$$\mathrm{M}_0 = \mathrm{M}_1 + \mathbf{i}. \tag{2.16}$$

$$\mathrm{M}_i^{-1}\mathbf{u} = \mathbf{L}^{-1}(\mathbf{u} - x_i\mathbf{i}) \tag{2.17}$$

Proposition 2.3.1 *For any $d \in \mathbb{R}$*

$$\mathrm{M}_i(\mathbf{u} + d\mathbf{j}) = \mathrm{M}_i\mathbf{u} + d\mathbf{j}, \tag{2.18}$$

$$\mathrm{M}_i^{-1}(\mathbf{u} + d\mathbf{j}) = \mathrm{M}_i^{-1}\mathbf{u} + d\mathbf{j}, \tag{2.19}$$

This is proven in Appendix-B.1.

2.4 Dynamics equations of DC inputs

Since $\Sigma\Delta$ modulation normally operates on oversampling, DC inputs are always considered such that $x[n] = x$. Then the dynamics is simply defined by one mapping:

$$\mathbf{u}[n+1] = \mathcal{M}_x\,\mathbf{u}[n].$$

The output $q[n]$ is

$$q[n] = \begin{cases} \frac{1}{2}, & t(\mathbf{u}[n]) \geq c_x \\ -\frac{1}{2}, & t(\mathbf{u}[n]) < c_x \end{cases}. \tag{2.20}$$

The following proposition shows the equivalence of some dynamics.

Proposition 2.4.1 *For any $c \in \mathbb{R}$, consider two mappings*

$$\mathrm{M}\,\mathbf{u} := \begin{cases} \mathrm{M}_0\mathbf{u}, & u \in \Omega_0 \\ \mathrm{M}_1\mathbf{u}, & u \in \Omega_1 \end{cases} \qquad \textit{and} \qquad \mathrm{M}'\,\mathbf{u} := \begin{cases} \mathrm{M}_0\mathbf{u}, & u \in \Omega_0 + c\,\mathbf{j} \\ \mathrm{M}_1\mathbf{u}, & u \in \Omega_1 + c\,\mathbf{j} \end{cases}.$$

Take any $\mathbf{u}, \mathbf{u}' \in \mathbb{R}^2$ such that $\mathbf{u}' = \mathbf{u} + c\,\mathbf{j}$, then

$$\mathrm{M}'\,\mathbf{u}' = \mathrm{M}\,\mathbf{u} + c\,\mathbf{j}.$$

Proof: It is clear that $\mathbf{u} \in \Omega_0$ if and only if $\mathbf{u}' = \mathbf{u} + c\,\mathbf{j} \in \Omega_0 + c\,\mathbf{j}$. It follows from (2.18) that $\mathrm{M}'(\mathbf{u}') = \mathrm{M}'(\mathbf{u} + c\,\mathbf{j}) = \mathrm{M}\mathbf{u} + c\,\mathbf{j}$ for all $\mathbf{u} \in \mathbb{R}^2$.

■

It implies the following proposition with recursion.

Proposition 2.4.2 *For any $\mathbf{u}, \mathbf{u}' \in \mathbb{R}^2$ such that $\mathbf{u}' = \mathbf{u} + c\,\mathbf{j}$, where $c \in \mathbb{R}$,*

$$\mathrm{M}'^n\mathbf{u}' = \mathrm{M}^n\mathbf{u} + c\,\mathbf{j},\ \forall n \in \mathbb{N},$$

where M and M' are defined as in Proposition 2.4.1.

This implies that the dynamics of the mappings M and M′ are exactly the same up to a fixed vertical shift resulting from the different partitions. So this difference can be basically ignored. Furthermore, equation (2.13) implies that the only effect of c_x is a vertical shift of the partition $\{\Omega_0, \Omega_1\}$. Therefore, from now on, we concentrate on the mapping

$$\mathrm{M}\,\mathbf{u} := \begin{cases} \mathrm{M}_1\mathbf{u}, & \mathbf{u} \in \Omega_1 \\ \mathrm{M}_0\mathbf{u}, & \mathbf{u} \in \Omega_0 \end{cases} . \tag{2.21}$$

Since, we have by Proposition 2.4.2, for all $n \in \mathbb{N}$,

$$\mathcal{M}_x^n\,\mathbf{u} = \mathrm{M}^n(\mathbf{u} - c_x\,\mathbf{j}) + c_x\,\mathbf{j}. \tag{2.22}$$

Proposition 2.4.3 *Assume that* $|x| \geq \frac{1}{2}$. *The stable region of* M *defined in* (2.21) *is a measure*[1] *zero set.*

This is proven in Appendix-B.2.

Therefore, $x \in (-\frac{1}{2}, \frac{1}{2})$ is the necessary condition for the system to have a nonzero-measure stable region. Moreover, because the system operation is symmetrical with respect to the sign of the input, one does not loose generality to assume:

Condition 2.4.4

$$x \in [0, \tfrac{1}{2}).$$

The configurations are described by parameter s and input x.

[1]It is the *Lebesgue measure*.

Chapter 3

Lyapunov function and global stability

The mixed discrete-time and discrete-amplitude nature of the feedback has prevented the use of the typical theories of stability in physical systems. The theoretical concepts of positively invariant sets and trapping sets of the state space are exploited, but the refine analysis available until now has been mostly based on explicit algebraic inspections of the discrete dynamics. This approach often yields complicated equations or discrete recursive reasonings, from which it is difficult to extract high-level and general guidelines on the state behaviors. In thesis, we attempt a more physical and analytic approach to the state behavior of $\Sigma\Delta$ modulators, with the introduction of Lyapunov functions.

3.1 Introduction

The introduction of Lyapunov functions was first suggested in $\Sigma\Delta$ modulation in [24]. We show in Figure 3.1(a,b,c) proposed examples of such functions for second order $\Sigma\Delta$ modulators. Using (2.21), the three figures 3.1(a,b,c) represent the same trajectory of state points $\mathbf{u}[n]$, but with third dimension values $h(\mathbf{u}[n])$ corresponding to three different Lyapunov functions $h(\mathbf{u})$. If the chosen function is lower bounded, one will always identify a region in which the function can no longer be decreased by the system

mapping. This is fundamentally due to the discrete and finite displacements of the state points $\mathbf{u}[n]$. This region is denoted by Υ_h and is illustrated in Figures 3.1(a',b',c') (appearing as a polygon) in the three cases of Figures 3.1(a,b,c), respectively. We show in this thesis that the mere determination of Υ_h for any given $h(\mathbf{u})$ leads to the identification of a trapping set of the system states. This set is denoted by Λ_h in Figures 3.1(a',b',c'). How close this set is to the actual attractor (represented by the gray dots in the figures) depends on the choice of $h(\mathbf{u})$.

This approach has several advantages. It is firstly non-recursive and concentrates all the algebraic difficulties into the analysis of a single memoryless function of the space. This replaces the discrete reasonings, difficult to formalize, by standard function analysis. In the same trend, this opens the door to more mathematical tools such as set theory and topology. Secondly, it enjoys the free choice of a tractable function $h(\mathbf{u})$, which is not constrained to tightly reproduce the discrete dynamics of the system, and on which can be attached convenient mathematical properties, such as continuity or convexity for example. It is true that a "too convenient" function $h(\mathbf{u})$ may simply yield dull results. However, with a basic piecewise quadratic function $h(\mathbf{u})$, we already reproduce and even improve or generalize under a unifying approach quantitative results previously obtained on the trapping sets of second order $\Sigma\Delta$ modulators. In particular, trapping properties and tight quantitative bounds are obtained in a single shot. But the third advantage of our proposed method is its very conceptual nature, giving it high potential for future generalizations. We give in this thesis general guidelines that are applicable to a whole class of Lyapunov functions, which can incorporate more refine mechanisms than a piecewise quadratic function

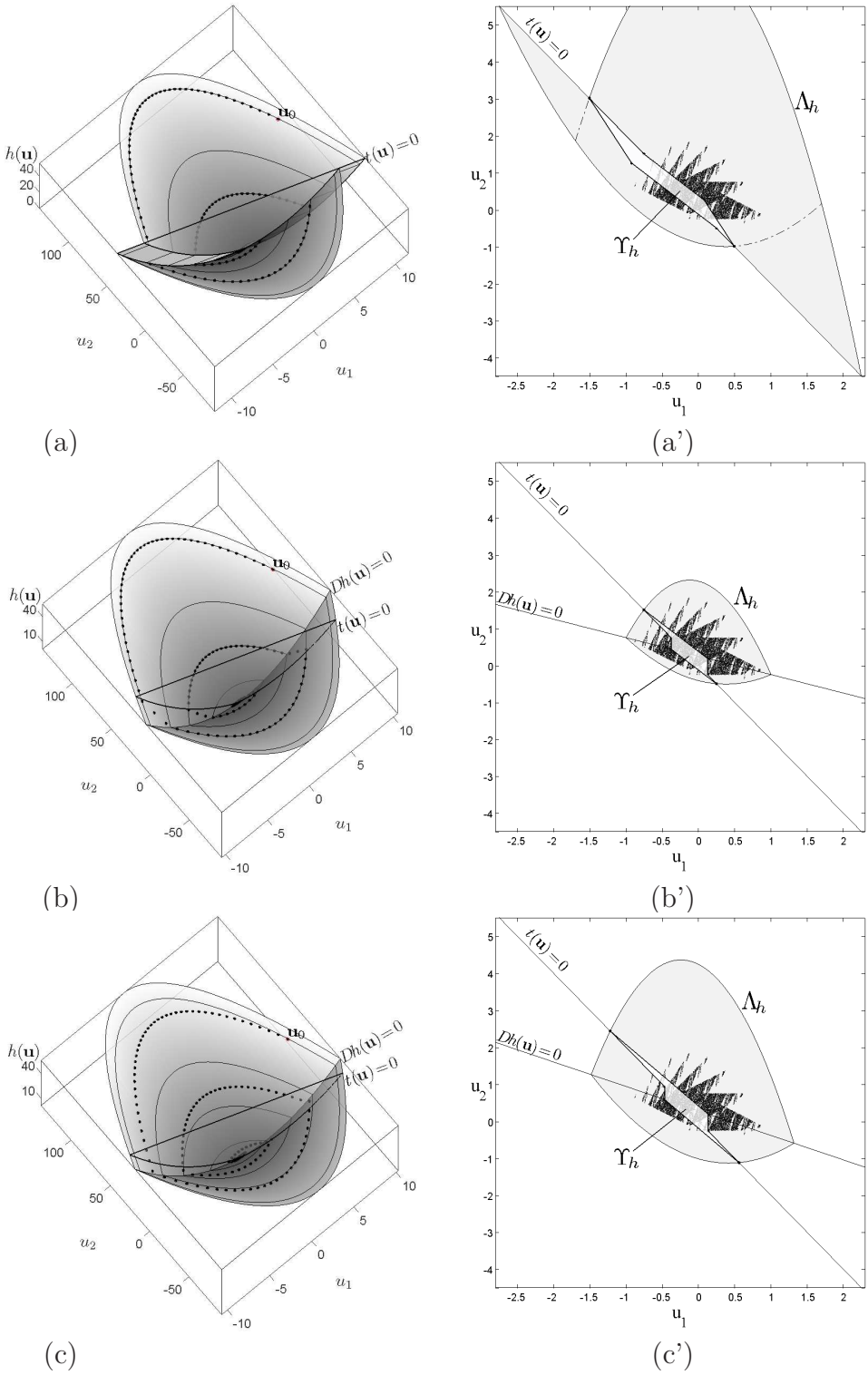


Figure 3.1: Surface of Lyapunov function $h(\mathbf{u})$ and the smallest trapping set Λ_h resulting from the Lyapunov functions: (a) case of (3.9) with $H_i = \Omega_i$ and $\Delta_i h^i(\mathbf{u}) = 0$ for both $i = 0, 1$; (b) case of (3.20) with $\Delta_i h^i(\mathbf{u}) = 0$; (c) case of (3.20) with $\Delta_i h^i(\mathbf{u}) = eps < 0$; (a',b',c') show the smallest trapping set Λ_h resulting from the Lyapunov functions $h(\mathbf{u})$ of (a,b,c), respectively.

and can be explored in the future. Furthermore, we will show in this thesis the effectiveness of the Lyapunov function method in the time-varying case. Like in [9], we show the existence of a trapping set for inputs that are finite sums of sinusoids with dc components that remain in the stable region of constant inputs. Our particular contribution here is to make the proof conceptual and substantially more concise with our new framework of analysis, results for more general configurations obtained simultaneously. Finally, another potential use of the Lyapunov functions will be the analytical derivation of stability results at the higher-order $\Sigma\Delta$ modulators.

3.2 Lyapunov function approach

3.2.1 Trapping sets

We consider a dynamical system of equation $\mathbf{u}[n+1] = \mathrm{M}\mathbf{u}[n]$, where M is a given in (2.21). But in this subsection, all results must apply to any transformation of $\mathbb{R}^2$ without any particular assumption. We define

Definition 3.2.1 *A* trapping set *of* M *in* S_0*, any set* $S \subset S_0$ *such that*

(i) $\mathrm{M}(S) \subset S$*,*

(ii) for any $\mathbf{u} \in S_0$*, there exists* $n \geq 0$ *such that* $\mathrm{M}^n\mathbf{u} \in S$*.*

When $S_0 = \mathbb{R}^2$, we call trapping set S the global trapping set.

As consequence of the two properties of trapping set, for any initial condition $\mathbf{u}[0] \in S_0$, there exists $n_0 \geq 0$ such that $\mathbf{u}[n]$ remains in S for all $n \geq n_0$. This implies the idea of trapping in the strong sense that any point in the set S_0 gets trapped into the set S within a *finite* number of iterations. When a set S satisfies (i), it is said to

be *positively invariant* [16, 20]. Meanwhile, we will call any set S that satisfies (ii) an *attracting set*[1]. Our main purpose is to find global trapping sets that are as small as possible.

With the introduction of a Lyapunov function $h(\mathbf{u})$, there are *immediate* ways to construct positively invariant sets and trapping sets, at least conceptually. For any real function $h(\mathbf{u})$ of $\mathbb{R}^2$, define

$$\Delta h(\mathbf{u}) := h(\mathrm{M}\mathbf{u}) - h(\mathbf{u}). \tag{3.1}$$

Within the case $\Delta h(\mathbf{u}) \leq 0$ everywhere, the set

$$\Lambda_h(\ell) := \big\{\mathbf{u} \in \mathbb{R}^2 : h(\mathbf{u}) \leq \ell\big\} \tag{3.2}$$

is automatically positively invariant for any $\ell \in \mathbb{R}$. Of course, with any function $h(\mathbf{u})$ picked at random, $\Delta h(\mathbf{u}) \leq 0$ is not always guaranteed. Moreover, finding a function $h(\mathbf{u})$ such that $\Delta h(\mathbf{u}) \leq 0$ globally is ideal but not necessary. To have more freedom on choosing functions, we construct positively invariant sets by defining

$$\Upsilon_h := \big\{\mathbf{u} \in \mathbb{R}^2 : \Delta h(\mathbf{u}) > 0\big\}. \tag{3.3}$$

The set Υ_h is qualitatively the subset of $\mathbb{R}^2$ where $\Delta h(\mathbf{u}) \leq 0$ fails. We have the following general property.

Proposition 3.2.2 *Consider any set $\Lambda_h(\ell)$ that includes $\mathrm{M}(\Upsilon_h)$. Then,*

(i) $\Lambda_h(\ell)$ includes Υ_h,

(ii) $\Lambda_h(\ell)$ is positively invariant.

[1] Although we use the term "attracting", (ii) is not to be confused with the standard notion of *attractor* in dynamical systems, which is different.

This is proven in Appendix-C.1.

Now, $\{\Lambda_h(\ell)\}_\ell$ is an increasing family of sets, i.e.

$$\ell \leq \ell' \qquad \Longrightarrow \qquad \Lambda_h(\ell) \subset \Lambda_h(\ell').$$

Therefore, this family has a smallest set that includes $\mathrm{M}(\Upsilon_h)$ and that we denote by Λ_h. Analytically,

$$\Lambda_h := \Lambda_h(\ell_h)$$

where

$$\ell_h := \sup_{\mathbf{u}\in \mathrm{M}(\Upsilon_h)} h(\mathbf{u}). \tag{3.4}$$

Proposition 3.2.2 is then equivalent to the following statement.

Proposition 3.2.3 *The set $\Lambda_h(\ell)$ is positively invariant for any $\ell \geq \ell_h$.*

This proposition thus provides a whole family of positively invariant sets, with no condition imposed on M and on $h(\mathbf{u})$. One however should not see in this a miracle. At this point, nothing prevents ℓ_h from being $+\infty$ for example, then Λ_{ℓ_h} is unbounded. On the other hand, even if Λ_{ℓ_h} is bounded, one cannot say either that ℓ_h is the smallest ℓ such that $\Lambda_h(\ell)$ is positively invariant because we have not yet proven that the set $\Lambda_h(\ell)$ is not when $\ell < \ell_h$. We will fix this problem later which involves more.

The next question is how to ensure that these positively invariant sets are also attracting sets. With Proposition 3.2.2(i) and (3.4), we know that $\Lambda_h(\ell)$ includes Υ_h for any given $\ell \geq \ell_h$. Then for all $\mathbf{u} \notin \Lambda_h(\ell)$, $\mathbf{u} \notin \Upsilon_h$, which implies $\Delta h(\mathbf{u}) \leq 0$. This is unfortunately too weak to make $\Lambda_h(\ell)$ an attracting set because it is possible

$\Delta h(\mathbf{u}) = 0$ for all $\mathbf{u}$. We need to develop more tools. For any given $\varepsilon \geq 0$, let us define more generally

$$\Upsilon_h(\varepsilon) := \big\{\mathbf{u} \in \mathbb{R}^2 : \Delta h(\mathbf{u}) > -\varepsilon\big\}. \tag{3.5}$$

Proposition 3.2.4 *Assume that the function $h(\mathbf{u})$ is lower bounded. Then, for any $\varepsilon > 0$, $\Upsilon_h(\varepsilon)$ is an attracting set.*

This proposition is proven in Appendix-C.2. It has the following consequence.

Proposition 3.2.5 *Assuming that $h(\mathbf{u})$ is lower bounded, any positively invariant set S that includes $\Upsilon_h(\varepsilon)$ for some $\varepsilon > 0$ is a trapping set.*

As a reminder, positive invariance is a notion intrinsic to the space transformation M. Therefore, how S has been obtained as a positively invariant set is of no concern in this proposition. For example, one can imagine situations where the positive invariance of S has been previously established using another Lyapunov function $h'(\mathbf{u})$. This non-trivial idea will be used later.

3.2.2 Lyapunov functions for piecewise mapping

The previous section gave basic tools to extract positively invariant sets and trapping sets from any given Lyapunov function. The difficulty is now how to optimize the Lyapunov function so that these sets are as small as possible. The particular obstacle is that our considered mapping M from a $\Sigma\Delta$ modulator does not have a single analytical expression, but instead has a discontinuous piecewise definition as in (2.21).

In the first introduction of a Lyapunov function for the stability analysis of a $\Sigma\Delta$ modulator [24], it was proposed to design $h(\mathbf{u})$ in a piecewise manner as well. The

idea was to start with two separate functions $h^0(\mathbf{u})$ and $h^1(\mathbf{u})$ such that

$$h^0(\mathrm{M}_0\mathbf{u}) = h^0(\mathbf{u}) \qquad \text{and} \qquad h^1(\mathrm{M}_1\mathbf{u}) = h^1(\mathbf{u}). \tag{3.6}$$

Defining the function operation

$$\Delta_i f(\mathbf{u}) := f(\mathrm{M}_i\mathbf{u}) - f(\mathbf{u}), \tag{3.7}$$

(3.6) can be rewritten as

$$\Delta_i h^i(\mathbf{u}) = 0 \tag{3.8}$$

for both $i = 0, 1$. Then, a global Lyapunov function was formed by defining

$$h(\mathbf{u}) := \begin{cases} h^0(\mathbf{u}), & \mathbf{u} \in H_0 \\ h^1(\mathbf{u}), & \mathbf{u} \in H_1 \end{cases} \tag{3.9}$$

with $H_0 := \Omega_0$ and $H_1 := \Omega_1$. The function $\Delta h(\mathbf{u})$ is thus guaranteed to be 0 for all points $\mathbf{u}$ such that either $\mathbf{u}$ and $\mathrm{M}_0\mathbf{u}$ belong to Ω_0, or, $\mathbf{u}$ and $\mathrm{M}_1\mathbf{u}$ belong to Ω_1. We show in Figure 3.1(a) an example of such a Lyapunov function. One can observe that $\Delta h(\mathbf{u}[n])$ for a given trajectory $\mathbf{u}[n+1] = \mathrm{M}(\mathbf{u}[n])$ is zero most of the time, and fortunately appears to be negative in the non-controlled events when $\mathbf{u}[n]$ crosses the partition boundary $t(\mathbf{u}) = 0$, at least at the resolution scale of the picture. The second step is then to derive the resulting set Υ_h to find the smallest positively invariant set Λ_h yielded by this Lyapunov function, using the results from Section 3.2.1.

Finding functions $h^i(\mathbf{u})$ that satisfy (3.6) actually appears to be an old problem. Indeed, with any trajectory $\mathbf{u}[n+1] = \mathrm{M}\mathbf{u}[n]$, (3.6) implies that $h^i(\mathbf{u}[n+1]) = h^i(\mathbf{u}[n])$ when $\mathbf{u}[n] \in \Omega_i$. In other words, this implies that $h^i(\mathbf{u}[n])$ remains constant as long as $\mathbf{u}[n]$ remains in Ω_i. Such functions were derived in [9, 15, 16]. By applying (2.14),

one can find $h^i(\mathbf{u})$ such that $\Delta_i h^i(\mathbf{u}) = 0$ in the form of as simple as polynomial function

$$h^i(\mathbf{u}) = \alpha_i\big(u_2 - \tfrac{1}{2x_i}(u_1 - \tfrac{x_i}{2})^2 - \beta_i\big). \tag{3.10}$$

Coming back to the global Lyapunov function of (3.9) for mapping (2.21), one can also guarantee $\Lambda_h(\ell)$ to be bounded for all ℓ by choosing

$$\alpha_0 < 0 < \alpha_1. \tag{3.11}$$

The examples of Figure 3.1(a) and (a') were actually designed in this way.

Next, we push further the idea of [24] by considering the following degrees of freedom:

1. Functions $h^i(\mathbf{u})$ that satisfy $\Delta_i h^i(\mathbf{u})$ may not necessarily be of polynomial type like in (3.10). One could actually seek for functions $h^i(\mathbf{u})$ that satisfy $\Delta_i h^i(\mathbf{u}) = -\varepsilon < 0$, and thus actively participate to the decrease of the global Lyapunov function.

2. The partition $\{H_0, H_1\}$ in (3.9) may not necessarily be equal to the original partition $\{\Omega_0, \Omega_1\}$ of the mapping M.

In the subsequent sections, we detail our specific contributions to these orientations.

3.2.3 Controlling $\Delta_i h^i(\mathbf{u})$

Inspired by (3.10), we search for the solutions of the equation

$$\Delta_i h^i(\mathbf{u}) = -\varepsilon \tag{3.12}$$

that are of the form

$$h^i(\mathbf{u}) = \alpha_i\big(u_2 - g_i(u_1)\big) \tag{3.13}$$

with $\alpha_i \neq 0$. The dependence of $h^i(\mathbf{u})$ with u_1 and u_2 lies in two separate terms, and the function of u_2 being linear.

Proposition 3.2.6 *Assume that $f_i(u_1)$ is a particular solution to the equation*

$$f_i(u_1 + x_i) = f_i(u_1) + u_1. \tag{3.14}$$

Then, the function $h^i(\mathbf{u})$ of (3.13) *satisfies* (3.12) if and only if

$$g_i(u_1) = f_i(u_1 + \tfrac{\varepsilon}{\alpha_i}) + \beta_i(u_1), \tag{3.15}$$

where $\beta_i(u_1)$ is an x_i-periodic function.

This is proven in Appendix-C.3.

One can easily check that

$$p_i(u_1) := \tfrac{1}{2x_i}(u_1 - \tfrac{x_i}{2})^2 \tag{3.16}$$

is a particular solution to (3.14). Then by taking $f_i(u_1) = p_i(u_1)$ and choosing all possible x_i-periodic functions $\beta_i(u_1)$, the function $g_i(\mathbf{u}_1)$ of (3.15) generates all the functions $h^i(\mathbf{u})$ of (3.13) that satisfy (3.12). Although there is a potential to exploit this freedom of solutions, we will restrict ourselves in this thesis to a constant function $\beta_i(u_1) = \beta_i$. Explicitly, $h^i(\mathbf{u})$ has the form

$$h^i(\mathbf{u}) = \alpha_i\big(u_2 - p_i(u_1 + \tfrac{\varepsilon}{\alpha_i}) - \beta_i\big). \tag{3.17}$$

Note that (3.10) is the particular case of $\varepsilon = 0$.

3.2.4 Continuous Lyapunov function

A shortcoming of the Lyapunov function $h(\mathbf{u})$ of (3.9) is its discontinuity between H_0 and H_1. By allowing freedom on the choice of partition $\{H_0, H_1\}$, continuity can be

enforced. One method is to choose

$$H_0 := \big\{\mathbf{u} : Dh(\mathbf{u}) < 0\big\} \qquad \text{and} \qquad H_1 := \big\{\mathbf{u} : Dh(\mathbf{u}) \geq 0\big\} \tag{3.18}$$

where

$$Dh(\mathbf{u}) := h^1(\mathbf{u}) - h^0(\mathbf{u}). \tag{3.19}$$

With such a partition, we can easily see that

$$h(\mathbf{u}) = \max\big(h^0(\mathbf{u}), h^1(\mathbf{u})\big). \tag{3.20}$$

Whenever $h^0(\mathbf{u})$ and $h^1(\mathbf{u})$ are chosen continuous, $h(\mathbf{u})$ is automatically continuous. In fact, the functions $h^i(\mathbf{u})$ of (3.17) carry more analytical properties. Since $p_i(u_1)$ is quadratic of second derivative $-\frac{1}{x_i}$, $h^i(\mathbf{u})$ can even be made convex by choosing α_i of opposite sign to x_i. Since $x_1 < 0 < x_0$, this constrain compromises to (3.11), too. We recall that a function $f(\mathbf{u})$ is convex when

$$\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^2, \quad \forall \theta \in [0,1], \qquad f\big(\theta\mathbf{u} + (1-\theta)\mathbf{v}\big) \;\leq\; \theta f(\mathbf{u}) + (1-\theta) f(\mathbf{v}). \tag{3.21}$$

This notion is indeed stronger than that of continuity. When $h^0(\mathbf{u})$ and $h^1(\mathbf{u})$ are both convex, it is easy to show that $h(\mathbf{u})$ is convex as well. Figures 3.1(b,c) show two examples of convex functions $h(\mathbf{u})$. Besides the general advantages of convex functions such as having no local minima except for a possible global minimum, an advantage of particular interest to us is from the following proposition.

Proposition 3.2.7 *Assume that $t(\mathbf{u})$ is affine, the restrictions of $h(\mathbf{u})$ to Ω_0 and Ω_1 are both convex and Υ_h is bounded. Then, for any $\ell \in \mathbb{R}$,*

$$\Lambda_h(\ell) \textit{ includes } \mathrm{M}(\Upsilon_h) \quad \text{if and only if} \quad \textit{it includes } \Upsilon_h.$$

Moreover,

$$\ell_h = \sup_{\mathbf{u}\in\Upsilon_h} h(\mathbf{u}). \tag{3.22}$$

This is proven in Appendix-C.4.

Thus, to look for positively invariant sets $\Lambda_h(\ell)$ through the criterion of Proposition 3.2.2, one will not need to derive $\mathrm{M}(\Upsilon_h)$ but only Υ_h, which simplifies the task. The sets $\Lambda_h = \Lambda_h(\ell_h)$ of Figure 3.1(a',b',c') were actually derived using (3.22).

3.2.5 Analysis of $\Delta h(\mathbf{u})$

One can easily control $\Delta_0 h^0(\mathbf{u})$ and $\Delta_1 h^1(\mathbf{u})$ as shown in Section 3.2.3. But through this to make $\Upsilon_h(\varepsilon)$ and Υ_h small involves more. Because of the piecewise definitions of M in (2.13) and $h(\mathbf{u})$ in (3.9), the function $\Delta h(\mathbf{u}) = h(\mathrm{M}\mathbf{u}) - h(\mathbf{u})$ can locally coincide with any of the functions $h_j(\mathrm{M}_i\mathbf{u}) - h_k(\mathbf{u})$ with $i, j, k \in \{0, 1\}$. However, when the Lyapunov function is chosen in the form $h(\mathbf{u}) = \max\big(h^0(\mathbf{u}), h^1(\mathbf{u})\big)$ as introduced in (3.20), we show in this section that $\Delta h(\mathbf{u})$ yields some outstanding properties.

To simplify the analysis, note that

$$\Delta h(\mathbf{u}) = \begin{cases} \Delta_0 h(\mathbf{u}), & \mathbf{u} \in \Omega_0 \\ \Delta_1 h(\mathbf{u}), & \mathbf{u} \in \Omega_1 \end{cases}, \tag{3.23}$$

which is obtained by simple application (2.21), (3.1) and (3.7). One intuitively wishes $\Delta_i h(\mathbf{u})$ to be equal to the well controlled function $\Delta_i h^i(\mathbf{u})$ as often as possible in Ω_i, for each $i \in \{0, 1\}$. If not, it is at least hoped that the inequality $\Delta_i h(\mathbf{u}) > \Delta_i h^i(\mathbf{u})$ occurs in Ω_i as rarely as possible. This can be tested thanks to the following proposition.

Proposition 3.2.8

$$\Delta_0 h(\mathbf{u}) > \Delta_0 h^0(\mathbf{u}) \quad \Longleftrightarrow \quad \mathrm{M}_0 Dh(\mathbf{u}) > 0 \quad \textit{and} \quad \Delta_0 Dh(\mathbf{u}) > 0, \tag{3.24}$$

$$\Delta_1 h(\mathbf{u}) > \Delta_1 h^1(\mathbf{u}) \quad \Longleftrightarrow \quad \mathrm{M}_1 Dh(\mathbf{u}) < 0 \quad \textit{and} \quad \Delta_1 Dh(\mathbf{u}) < 0, \tag{3.25}$$

where for any scalar function $f(\mathbf{u})$*,*

$$\mathrm{M}_i f(\mathbf{u}) := f(\mathrm{M}_i \mathbf{u}). \tag{3.26}$$

This is proven in Appendix-C.5.

The major contribution of (3.24) and (3.25) is that the functions $\mathrm{M}_i Dh(\mathbf{u})$ and $\Delta_i Dh(\mathbf{u})$ involved in their right hand sides have one-piece analytical expressions, which will be in practice easy to derive. Next, we give an operational application of Proposition 3.2.8. This requires the additional set definitions

$$S_f^- := \big\{\mathbf{u} \in \mathbb{R}^2 : f(\mathbf{u}) < 0\big\} \qquad \text{and} \qquad S_f^+ := \big\{\mathbf{u} \in \mathbb{R}^2 : f(\mathbf{u}) > 0\big\}, \tag{3.27}$$

for any real function $f(\mathbf{u})$.

Proposition 3.2.9 *Consider some constant* ε *and assume that* $\Delta_i h^i(\mathbf{u}) = -\varepsilon$ *for all* $\mathbf{u} \in \mathbb{R}^2$ *and* $i \in \{0, 1\}$*. Then, the set* $\Upsilon_h(\varepsilon)$ *defined in* (3.5) *is equal to*

$$\Upsilon_h(\varepsilon) = \Gamma_{Dh} \tag{3.28}$$

where for any function $f(\mathbf{u})$*,*

$$\Gamma_f := \Gamma_f^0 \cup \Gamma_f^1, \tag{3.29}$$

$$\Gamma_f^0 := S^+_{\mathrm{M}_0 f} \cap S^+_{\Delta_0 f} \cap S_t^- \qquad \textit{and} \qquad \Gamma_f^1 := S^-_{\mathrm{M}_1 f} \cap S^-_{\Delta_1 f} \cap \overline{S_t^+}. \tag{3.30}$$

This is proven in Appendix-C.6.

Note that the set Γ_{Dh} is defined as soon as the functions $Dh(\mathbf{u})$ and $t(\mathbf{u})$ are given. Proposition 3.2.4 tells us that, in the case where $\Delta_i h^i(\mathbf{u}) = -\varepsilon$, Γ_{Dh} coincides with $\Upsilon_h(\varepsilon)$. So when $\varepsilon = 0$, any set $\Lambda_h(\ell)$ that includes Γ_{Dh} is positively invariant. And when $\varepsilon > 0$, Γ_{Dh} is automatically an attracting set. Studying the structure of the set Γ_f is therefore the key.

3.3 A family of sets $\left\{\Gamma_{Dh_{\delta,\varepsilon}}\right\}_{\varepsilon\geq 0}$

3.3.1 Description

Obtained from (3.17) with $\alpha_i = -x_i$ and $\beta_i = \delta$, $h^i_{\delta,\varepsilon}$ has explicit expression

$$h^i_{\delta,\varepsilon}(\mathbf{u}) = \tfrac{1}{2}(u_1 - \tfrac{\varepsilon}{x_i} - \tfrac{x_i}{2})^2 - x_i(u_2 - \delta). \tag{3.31}$$

We then define $h_{\delta,\varepsilon}(\mathbf{u}) := \max\left(h^0_{\delta,\varepsilon}(\mathbf{u}), h^1_{\delta,\varepsilon}(\mathbf{u})\right)$. Note that $h^i_{\delta,\varepsilon}(\mathbf{u}) = h^i_{0,\varepsilon}(\mathbf{u} - \delta\mathbf{j})$, so

$$h_{\delta,\varepsilon}(\mathbf{u}) = h_{0,\varepsilon}(\mathbf{u} - \delta\mathbf{j}). \tag{3.32}$$

With the choice $\alpha_i = -x_i$, we have forced $h^i(\mathbf{u})$ to have the same curvature in u_1 for $i = 0, 1$, thus making the difference $Dh_{\delta,\varepsilon}(\mathbf{u}) = h^1_{\delta,\varepsilon}(\mathbf{u}) - h^0_{\delta,\varepsilon}(\mathbf{u})$ affine. Using the fact $x_1 < 0 < x_0$ from (2.14), we have explicitly

$$Dh_{\delta,\varepsilon}(\mathbf{u}) = u_2 + a_\varepsilon\, u_1 + b_{\delta,\varepsilon} \tag{3.33}$$

where

$$a_\varepsilon := \tfrac{1}{2} + \bar{\varepsilon}, \qquad b_{\delta,\varepsilon} := (-\tfrac{1}{4} + \bar{\varepsilon}^2)\, x - \delta \tag{3.34}$$

and

$$\bar{\varepsilon} := \frac{\varepsilon}{|x_0 x_1|}.$$

Proposition 3.3.1

$$\forall \mathbf{u} \in \mathbb{R}^2, \qquad h_{\delta,\varepsilon}(\mathbf{u}) = v\big(Dh_{\delta,\varepsilon}(\mathbf{u})\big) + h^{\mathrm{q}}_{\bar{\varepsilon}}(\mathbf{u}), \tag{3.35}$$

where

$$v(d) := \begin{cases} |x_0\, d|, & d \leq 0 \\ |x_1\, d|, & d \geq 0. \end{cases}, \tag{3.36}$$

$$h^{\mathrm{q}}_{\bar{\varepsilon}}(\mathbf{u}) := \tfrac{1}{2}\,(u_1 + 2x\bar{\varepsilon})^2 - \tfrac{1}{2}x_0x_1(\tfrac{1}{2} + \bar{\varepsilon})^2. \tag{3.37}$$

This is proven in Appendix-C.7.

3.3.2 General properties of Γ_f

In this section, we study the properties of Γ_f when $f(\mathbf{u})$ is the general affine function in the form

$$f(\mathbf{u}) = u_2 + a\,u_1 + b. \tag{3.38}$$

As a remainder, $t(\mathbf{u}) = u_2 + s\,u_1$ is defined in (2.8). By applying (3.26), (3.7), and (2.14), one easily finds for each $i \in \{0, 1\}$,

$$\mathrm{M}_i f(\mathbf{u}) = f(\mathrm{M}_i\mathbf{u}) = u_2 + (a{+}1)\,u_1 + (b{+}ax_i), \tag{3.39}$$

$$\Delta_i f(\mathbf{u}) = f(\mathrm{M}_i\mathbf{u}) - f(\mathbf{u}) = u_1 + ax_i. \tag{3.40}$$

Then, the sets Γ^i_f described in (3.30) are the intersections of three *half-planes* of the form S^+_g or S^-_g (up to closure) where $g(\mathbf{u})$ is one of the affine mappings $t(\mathbf{u})$, $\mathrm{M}_i f(\mathbf{u})$ and $\Delta_i f(\mathbf{u})$. The structure of Γ^i_f is then completely derived by linear algebra.

In this context of affine mappings, Γ^i_f is nothing but a polygon, in the widest sense that it can be open (unbounded) or empty. Using the following notation

$$R_f := \big\{\mathbf{u} \in \mathbb{R}^2 : f(\mathbf{u}) = 0\big\},$$

the edges of Γ_f^i are included in the straight lines R_t, $R_{\mathrm{M}_i f}$ and $R_{\Delta_i f}$ whose slopes are $-s$, $-(a+1)$ and ∞ (vertical line) according to (2.8), (3.39) and (3.40).

Proposition 3.3.2 *The set* Γ_f *is bounded* if and only if $s > a+1$.

This is proven in Appendix-C.8.

We assume from now on that $s > a+1$. The vertices of Γ_f as a polygon must be among the points

$$\mathbf{a}^i := R_{\Delta_i f} \cap R_t,\ \mathbf{b}^i := R_t \cap R_{\mathrm{M}_i f},\ \mathbf{c}^i := R_{\Delta_i f} \cap R_{\mathrm{M}_i f}. \tag{3.41}$$

These points are illustrated in Figure 3.2. It is tempting to claim that $\overline{\Gamma_f^i}$ is equal to the closed triangle $\triangle \mathbf{a}^i \mathbf{b}^i \mathbf{c}^i$ of vertices $\mathbf{a}^i$, $\mathbf{b}^i$ and $\mathbf{c}^i$, but this is true only when $\Gamma_f^i \neq \emptyset$.

Proposition 3.3.3 *For each* $i = 0, 1$,

$$\Gamma_f^0 \neq \emptyset \ \Leftrightarrow\ \mathrm{M}_0 f(\mathbf{a}^0) > 0 \ \Leftrightarrow\ \Delta_0 f(\mathbf{b}^0) > 0 \ \Leftrightarrow\ t(\mathbf{c}^0) < 0 \ \Leftrightarrow\ \overline{\Gamma_f^0} = \triangle \mathbf{a}^0 \mathbf{b}^0 \mathbf{c}^0, \tag{3.42}$$

$$\Gamma_f^1 \neq \emptyset \ \Leftrightarrow\ \mathrm{M}_1 f(\mathbf{a}^1) < 0 \ \Leftrightarrow\ \Delta_1 f(\mathbf{b}^1) < 0 \ \Leftrightarrow\ t(\mathbf{c}^1) > 0 \ \Leftrightarrow\ \overline{\Gamma_f^1} = \triangle \mathbf{a}^1 \mathbf{b}^1 \mathbf{c}^1. \tag{3.43}$$

This is proven in Appendix-C.8.

Figure 3.2 illustrates this proposition. Next, there are special properties on $\mathrm{M}_i f(\mathbf{u})$ and $\Delta_i f(\mathbf{u})$ that have not been used and that are the following:

$$\mathrm{M}_0 f(\mathbf{u}) - \Delta_0 f(\mathbf{u}) = \quad f(\mathbf{u}) \quad = \mathrm{M}_1 f(\mathbf{u}) - \Delta_1 f(\mathbf{u}), \tag{3.44}$$

$$\mathrm{M}_0 f(\mathbf{u}) - \mathrm{M}_1 f(\mathbf{u}) = \quad a \quad = \Delta_0 f(\mathbf{u}) - \Delta_1 f(\mathbf{u}). \tag{3.45}$$

The first equation is a trivial consequence of the definitions (3.7) and (3.26). The second equation results from (3.39), (3.40) and (2.14). As a result of (3.41) and

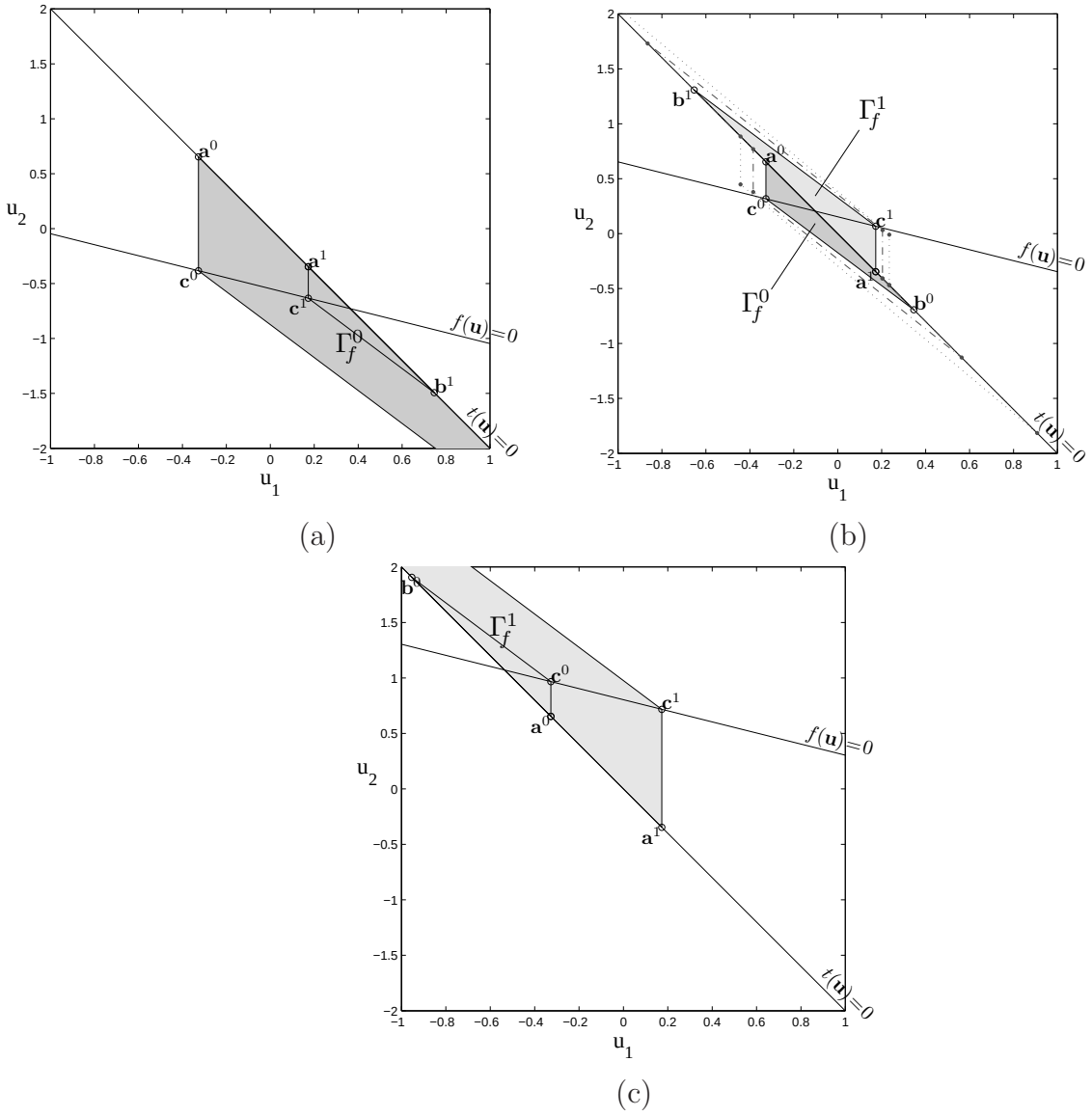


Figure 3.2: Sets Γ_f^0 (dark gray) and Γ_f^1 (light gray) with $s = 2$, $x = 0.153$ and various affine functions $f(\mathbf{u})$: (a) $\Gamma_f^0 \neq \emptyset$ and $\Gamma_f^1 = \emptyset$ ($\delta \leq \delta_\varepsilon^1$); (b) $\Gamma_f^0 \neq \emptyset$ and $\Gamma_f^1 \neq \emptyset$ ($\delta_\varepsilon^1 < \delta < \delta_\varepsilon^0$); (c) $\Gamma_f^0 = \emptyset$ and $\Gamma_f^1 \neq \emptyset$ ($\delta \geq \delta_\varepsilon^0$).

(3.44), one already has

$$\Delta_i f(\mathbf{b}^i) = 0,\ \mathrm{M}_i f(\mathbf{a}^i) = f(\mathbf{a}^i),\ \mathrm{M}_i f(\mathbf{b}^i) = 0,\ \Delta_i f(\mathbf{b}^i) = -f(\mathbf{b}^i). \tag{3.46}$$

So (3.42) and (3.43) imply the following new equivalences

$$\begin{array}{lllll} \Gamma_f^0 \neq \emptyset & \Leftrightarrow & f(\mathbf{a}^0) > 0 & \Leftrightarrow & f(\mathbf{b}^0) < 0 \\ \Gamma_f^1 \neq \emptyset & \Leftrightarrow & f(\mathbf{a}^1) < 0 & \Leftrightarrow & f(\mathbf{b}^1) > 0 \end{array} . \tag{3.47}$$

The following two propositions are consequences of (3.46) and (3.45), respectively.

Proposition 3.3.4 *The set Γ_f is always connected.*

This is proven in Appendix-C.9.

Proposition 3.3.5 *Assume $a > 0$ in* (3.38)*. Then,*

(i) Γ_f is non-empty,

(ii) when $\Gamma_f^i = \emptyset$, $\triangle \mathbf{a}^i \mathbf{b}^i \mathbf{c}^i \subset \triangle \mathbf{a}^{\bar{i}} \mathbf{b}^{\bar{i}} \mathbf{c}^{\bar{i}}$, where $\bar{i} = 1 - i$.

This is proven in Appendix-C.10.

When $a > 0$, we conclude from Propositions 3.3.3 and 3.3.5 that

$$\overline{\Gamma_f} = \triangle \mathbf{a}^0 \mathbf{b}^0 \mathbf{c}^0 \cup \triangle \mathbf{a}^1 \mathbf{b}^1 \mathbf{c}^1. \tag{3.48}$$

We now derive explicit results in terms of the coefficient parameters a, b and s. Let us first find the condition for Γ_f^i to be non-empty, using (3.47). Since the function $\Delta_i f(\mathbf{u})$ is simpler than $\mathrm{M}_i f(\mathbf{u})$, let us derive $\mathbf{a}^i$. We find

$$\mathbf{a}^i = (-a x_i\,,\ s a x_i). \tag{3.49}$$

This leads to

$$f(\mathbf{a}^i) = a(s-a)x_i + b. \tag{3.50}$$

Next, let us derive $\mathbf{b}^i$ and $\mathbf{c}^i$. Since $\mathbf{a}^i, \mathbf{b}^i \in O_t$ and $\mathbf{a}^i, \mathbf{c}^i \in O_{\Delta_i f}$, there exist $\alpha_i, \beta_i \in \mathbb{R}$ such that $\mathbf{b}^i = \mathbf{a}^i + \alpha_i\,(1, -s)$ and $\mathbf{c}^i = \mathbf{a}^i + \beta_i\,\mathbf{j}$. Using (3.39), we obtain $\mathrm{M}_i f(\mathbf{b}^i) = \mathrm{M}_i f(\mathbf{a}^i) + \alpha_i\,(a + 1 - s)$ and $\mathrm{M}_i f(\mathbf{c}^i) = \mathrm{M}_i f(\mathbf{a}^i) + \beta_i$. We solve α_i and β_i by writing $\mathrm{M}_i f(\mathbf{b}^i) = \mathrm{M}_i f(\mathbf{c}^i) = 0$ and $\mathrm{M}_i f(\mathbf{a}^i) = f(\mathbf{a}^i)$. This yields the following formulas:

$$\mathbf{b}^i = \mathbf{a}^i + \tfrac{f(\mathbf{a}^i)}{s-a-1}\,(1, -s) \qquad \text{and} \qquad \mathbf{c}^i = \mathbf{a}^i - f(\mathbf{a}^i)\,\mathbf{j}. \tag{3.51}$$

With (3.49) and (3.50), we know the points $\mathbf{b}^i$ and $\mathbf{c}^i$ explicitly.

3.3.3 Characterization of $\{\Gamma_{Dh_{\delta,\varepsilon}}\}_{\varepsilon \geq 0}$

Let us return to the Lyapunov function $h(\mathbf{u}) = h_{\delta,\varepsilon}(\mathbf{u}) := \max\left(h^0_{\delta,\varepsilon}(\mathbf{u}), h^1_{\delta,\varepsilon}(\mathbf{u})\right)$, where $h^i_{\delta,\varepsilon}(\mathbf{u})$ is described in (3.31) and $\delta \in \mathbb{R}$ and $\varepsilon \geq 0$ are adjustable parameters. By taking $f(\mathbf{u}) = Dh_{\delta,\varepsilon}(\mathbf{u})$ in the previous sections, we know that $\{\Gamma_{Dh_{\delta,\varepsilon}}\}_{\varepsilon \geq 0}$ is bounded when $s > a_\varepsilon + 1$, where $a_\varepsilon := \frac{1}{2} + \bar{\varepsilon}$ as given in (3.34). Assuming that $s > \frac{3}{2}$, this is equivalent to the constraint $\varepsilon \in [0, \varepsilon_s)$, where

$$\varepsilon_s := |x_0 x_1|(s - \tfrac{3}{2}). \tag{3.52}$$

Since $a_\varepsilon > 0$, $\overline{\Gamma_{Dh_{\delta,\varepsilon}}}$ is according to (3.48) the union of two triangles $\triangle \mathbf{a}^0\mathbf{b}^0\mathbf{c}^0$ and $\triangle \mathbf{a}^1\mathbf{b}^1\mathbf{c}^1$, which we now derive explicitly. By applying (3.50) and (3.34), we have

$$Dh_{\delta,\varepsilon}(\mathbf{a}^i) = \delta^i_\varepsilon - \delta \tag{3.53}$$

where

$$\delta^i_\varepsilon := a_\varepsilon(s - a_\varepsilon)x_i + (-\tfrac{1}{4} + \bar{\varepsilon}^2)x. \tag{3.54}$$

From (3.48), (3.49), (3.51) and (3.53), we find

$$\mathbf{a}^i = (-a_\varepsilon x_i\,,\; s a_\varepsilon x_i), \qquad \mathbf{b}^i = \mathbf{a}^i + \tfrac{1}{d_\varepsilon}\,(\delta^i_\varepsilon - \delta)\,(1,-s), \qquad \mathbf{c}^i = \mathbf{a}^i - (\delta^i_\varepsilon - \delta)\,\mathbf{j} \quad (3.55)$$

and $d_\varepsilon := s - a_\varepsilon - 1$.

Proposition 3.3.6 *Assume $s > \frac{3}{2}$. When δ belongs to a bounded interval $D \subset \mathbb{R}$ and ε goes to 0, the vertices of $\overline{\Gamma_{Dh_{\delta,\varepsilon}}}$ converge to those of $\overline{\Gamma_{Dh_{\delta,0}}}$* uniformly *with $\delta \in D$.*

This is proven in Appendix-C.11.

Figure 3.2(b) shows an actual example of comparison between $\Gamma_{Dh_{\delta,\varepsilon}}$ (dashed contour) and $\Gamma_{Dh_{\delta,0}}$ ($\Gamma^0_f \cup \Gamma^1_f$ in the figure) in a real case of $\Sigma\Delta$ modulation ($s = 2$, $x = 0.153$, $d = 0.115$, $\varepsilon = 0.025$ in mixed lines and $\varepsilon = 0.05$ in dotted lines). From (3.47) and (3.53), we have

$$\Gamma^0_{Dh_{\delta,\varepsilon}} \neq \emptyset \quad \Leftrightarrow \quad \delta < \delta^0_\varepsilon \qquad \text{and} \qquad \Gamma^1_{Dh_{\delta,\varepsilon}} \neq \emptyset \quad \Leftrightarrow \quad \delta > \delta^1_\varepsilon. \qquad (3.56)$$

Since $a_\varepsilon > 0$ and $\Gamma_{Dh_{\delta,\varepsilon}} = \Gamma^0_{Dh_{\delta,\varepsilon}} \cup \Gamma^1_{Dh_{\delta,\varepsilon}}$, we cannot have $\Gamma^0_{Dh_{\delta,\varepsilon}} = \emptyset$ and $\Gamma^1_{Dh_{\delta,\varepsilon}} = \emptyset$ simultaneously according to Proposition 3.3.5. So we expect to have $\delta^0_\varepsilon > \delta^1_\varepsilon$. Indeed, (3.54) and (2.14) imply that $\delta^0_\varepsilon - \delta^1_\varepsilon = a_\varepsilon(s - a_\varepsilon) > 0$, since $s > a_\varepsilon + 1 > a_\varepsilon$. Then, by Proposition 3.3.5(ii),

$$\Gamma_{Dh_{\delta,\varepsilon}} = \begin{cases} \triangle\mathbf{a}^0\mathbf{b}^0\mathbf{c}^0 & , \quad \delta \leq \delta^1_\varepsilon \\ \triangle\mathbf{a}^0\mathbf{b}^0\mathbf{c}^0 \cup \triangle\mathbf{a}^1\mathbf{b}^1\mathbf{c}^1 & , \quad \delta^1_\varepsilon < \delta < \delta^0_\varepsilon \\ \triangle\mathbf{a}^1\mathbf{b}^1\mathbf{c}^1 & , \quad \delta \geq \delta^0_\varepsilon \end{cases}. \qquad (3.57)$$

The three above cases are illustrated in Figure 3.2.

3.4 Family of trapping sets

3.4.1 Family of sets $\Lambda_\delta(\ell)$

We saw in Section 3.2 that for any given function $h(\mathbf{u})$, one obtains a smallest set $\Lambda_h = \Lambda_h(\ell_h)$ that is guaranteed to be positively invariant. We wish to optimize

the function $h(\mathbf{u})$ to make Λ_h as small as possible. In Section 3.3, we provided a parameterized family of Lyapunov functions $h_{\delta,\varepsilon}(\mathbf{u})$ with $\delta \in \mathbb{R}$ and $\varepsilon \geq 0$. A procedure is then to optimize δ and ε such that the size of $\Lambda_{h_{\delta,\varepsilon}}$ is minimized. Now, ε may not be a relevant parameter to play with. Indeed, ε was introduced to make $h_{\delta,\varepsilon}(\mathrm{M}^n\mathbf{u})$ decrease faster with n, through the relation $\Delta_i h^i_{\delta,\varepsilon}(\mathbf{u}) = -\varepsilon$. This however tends to make $h_{\delta,\varepsilon}(\mathrm{M}^n\mathbf{u})$ rebound earlier with n, thus increasing the size of $\Upsilon_{h_{\delta,\varepsilon}}$. We did observe experimentally that $\Lambda_{h_{\delta,\varepsilon}}$ increases in size when $\varepsilon > 0$, compared to $\varepsilon = 0$. This can be seen for example in Figures 3.1(b',c'). For this reason, we restrict ourselves in this section to the single-parameter family of Lyapunov functions $\{h_{\delta,0}\}_{\delta\in\mathbb{R}}$. Let us introduce the following simplified notation:

$$h^i_\delta(\mathbf{u}) := h^i_{\delta,0}(\mathbf{u}), \qquad h_\delta(\mathbf{u}) := h_{\delta,0}(\mathbf{u}), \qquad Dh_\delta(\mathbf{u}) := Dh_{\delta,0}(\mathbf{u}),$$
$$\Upsilon_\delta := \Upsilon_{h_{\delta,0}}, \qquad \Lambda_\delta(\ell) := \Lambda_{h_{\delta,0}}(\ell) \qquad \text{and} \qquad \ell_\delta := \ell_{h_{\delta,0}}. \tag{3.58}$$

The goal of this section is to find the smallest positively invariant set $\Lambda_\delta(\ell)$ and simultaneously test its trapping property.

Given the present choice of Lyapunov functions $\{h_\delta\}_{\delta\in\mathbb{R}}$, there is actually an equivalent description of the contour sets $\Lambda_\delta(\ell)$. Let us consider the family of sets $\{\Pi_{c_0,c_1}\}_{c_0,c_1\in\mathbb{R}}$ defined by

$$\Pi_{c_0,c_1} := \big\{\mathbf{u}\in\mathbb{R}^2 :\ p_0(u_1) + c_0 \leq u_2 \leq p_1(u_1) + c_1\big\}, \tag{3.59}$$

where $p_i(u_1)$ was defined in (3.16).

Proposition 3.4.1

$$\Lambda_\delta(\ell) = \Pi_{c_0,c_1} \qquad \Longleftrightarrow \qquad c_i = \delta - \tfrac{1}{x_i}\,\ell, \quad i = 0,1 \tag{3.60}$$

$$\Longleftrightarrow \qquad \delta = x_0c_0 - x_1c_1 \quad \textit{and} \quad \ell = x_0x_1(c_0 - c_1). \tag{3.61}$$

This is proven in Appendix-C.12.

In other words, $\{\Lambda_\delta(\ell)\}_{\delta,\ell\in\mathbb{R}}$ and $\{\Pi_{c_0,c_1}\}_{c_0,c_1\in\mathbb{R}}$ are equal families of sets, except that they are described by different parameters. Sets of the form (3.59) were actually considered in [9, 16] as candidates for trapping regions or positively invariant sets. The novelty here is our conceptual sufficient condition to recognize that such a set is positively invariant. According to Propositions 3.2.2, 3.2.7 and 3.4.1, any given set Π_{c_0,c_1} is positively invariant when it includes Υ_δ, where δ is given by (3.61). In fact, we are going to show that this condition is *necessary and sufficient*. This first requires a closer look at the Lyapunov function $h_\delta(\mathbf{u})$.

The expression (3.35) shows that the quadratic part $h^{\mathrm{q}}(\mathbf{u})$ of $h_\delta(\mathbf{u})$ is a one-piece function that does no depend on the variable u_2 and the parameter δ. Meanwhile, the piecewise part $a\big(Dh_\delta(\mathbf{u})\big)$ of $h_\delta(\mathbf{u})$ is continuous and affine in each piece. Based on (3.35) and $\Upsilon_\delta = \Gamma_{Dh_\delta}$, we obtain the following property.

Proposition 3.4.2 *For any given $\delta, \ell \in \mathbb{R}$,*

$$\Lambda_\delta(\ell) \textit{ is positively invariant} \text{ if and only if } \textit{it includes } \Upsilon_\delta.$$

This is proven in Appendix-C.13.

This is based on the fact that $h_\delta(\mathbf{u})$ has a unique global minimum and on the special properties of $\Upsilon_\delta = \Gamma_{Dh_\delta}$ obtained in Section 3.3. With (3.22) and the new notation (3.58), we conclude that ℓ_δ is the smallest value of ℓ such that $\Lambda_\delta(\ell)$ is positively invariant. The next question is whether $\Lambda_\delta(\ell)$ is also a trapping set for every $\ell \geq \ell_h$.

Proposition 3.4.3 *Assume $s > \frac{3}{2}$. For any given bounded interval $D \subset \mathbb{R}$ and $\lambda > 0$,*

$$\exists \varepsilon > 0, \quad \forall \delta \in D, \qquad \Lambda_\delta(\ell_\delta + \lambda) \supset \Upsilon_{h_{\delta,\varepsilon}}(\varepsilon).$$

This is proven in Appendix-C.14.

This proposition implies that for any given $\delta \in \mathbb{R}$, $\Lambda_\delta(\ell)$ is *automatically* a trapping set when $\ell > \ell_h$. Indeed, by taking $\lambda = \ell - \ell_h$, there exists $\varepsilon > 0$ such that $\Lambda_\delta(\ell)$ includes $\Upsilon_{h_{\delta,\varepsilon}}(\varepsilon)$. Then, the trapping property of $\Lambda_\delta(\ell)$ results from Proposition 3.2.5. This does not prove that $\Lambda_\delta := \Lambda_\delta(\ell_h)$ itself is a trapping set. We will say that Λ_δ is only an *asymptotic* trapping set.

The next question is whether there exists a "smallest" set Λ_δ among all $\delta \in \mathbb{R}$. When writing $\Lambda_\delta(\ell) = \Pi_{c_0,c_1}$, (3.58) tells us that ℓ is directly proportional to $c_1 - c_0$, which basically gives the size in u_2 of Π_{c_0,c_1} (up to a constant) as can be seen in (3.59). Meanwhile δ only acts as a global shift of the set in u_2. In the next section, we will show that ℓ_δ has as a function of δ a minimum $\ell^* = \ell_{\delta^*}$. Then, we can already claim that $\Lambda_{\delta^*} = \Lambda_{\delta^*}(\ell^*)$ is the smallest in size among the sets Λ_δ. The following proposition actually states a stronger result.

Proposition 3.4.4 *For* all $\delta \in \mathbb{R}$, $\Lambda_{\delta^*} \subset \Lambda_\delta$.

This is proven in Appendix-C.15.

Therefore, $\Lambda_{\delta^*} = \Lambda_{\delta^*}(\ell^*)$ is the smallest of the asymptotic trapping sets Λ_δ in the strong sense of inclusion.

3.4.2 Analytical derivation of δ^* and ℓ^*

To find δ^* and $\ell^* := \ell_{\delta^*}$, we need to derive explicitly ℓ_δ. From (3.22) and (3.28), we have $\ell_\delta = \sup_{\mathbf{u}\in\Gamma_{Dh_\delta}} h_\delta(\mathbf{u})$. Given the set description (3.48) of Γ_{Dh_δ}, let us define

$$\ell_\delta^i := \sup_{\mathbf{u}\in\triangle\mathbf{a}^i\mathbf{b}^i\mathbf{c}^i} h_\delta(\mathbf{u}).$$

From (3.57) with $\varepsilon = 0$, we have

$$\ell_\delta = \begin{cases} \ell_\delta^0 & , \quad \delta \le \delta^1 \\ \max\left(\ell_\delta^0, \ell_\delta^1\right) & , \quad \delta^1 \le \delta \le \delta^0 \\ \ell_\delta^1 & , \quad \delta \ge \delta^0 \end{cases} \tag{3.62}$$

where $$\delta^i := \delta_0^i = \tfrac{1}{2}(s-\tfrac{1}{2})x_i - \tfrac{1}{4}x. \tag{3.63}$$

Proposition 3.4.5

$$\begin{array}{l} \forall \delta \le \delta^0,\ \ell_\delta^0 = \max\left(h_\delta^1(\mathbf{a}^0), h_\delta^0(\mathbf{b}^0)\right) \\ \forall \delta \ge \delta^1,\ \ell_\delta^1 = \max\left(h_\delta^0(\mathbf{a}^1), h_\delta^1(\mathbf{b}^1)\right) \end{array} . \tag{3.64}$$

This is proven in Appendix-C.16.

Using the explicit expressions of $h_\delta^i(\mathbf{u})$, $\mathbf{a}^i$ and $\mathbf{b}^i$ from (3.31), and (3.55) with $\varepsilon = 0$, we find

$$\begin{array}{r} h_\delta^{\bar{i}}(\mathbf{a}^i) = x_{\bar{i}}\,\delta + \tfrac{1}{2}\left[(\tfrac{1}{4}-x^2)s+x^2\right] \\ h_\delta^i(\mathbf{b}^i) = \frac{(\delta-\mu^i)^2}{2(s-\frac{3}{2})^2} + \tfrac{1}{8}(\tfrac{1}{4}-x^2) \end{array} , \tag{3.65}$$

where

$$\mu^i := \delta^i + (s-\frac{3}{2})\frac{x_i}{2} \tag{3.66}$$

and $\bar{i} := 1-i$. We derive δ^* and ℓ^* in Appendix-C.17 from (3.62) and the subsequent equations. They yield different analytical expressions in three regions of the pair (x, s). The results are given in Table 3.1, where

$$s_1(x) := \tfrac{5}{2} - \tfrac{2x}{1+2x} \qquad \text{and} \qquad s_2(x) := \tfrac{5}{2} + \tfrac{3x}{1-2x}. \tag{3.67}$$

Table 3.1 is derived in Appendix-C.17.

Table 3.1:

range in s	δ^*	ℓ^*
$\frac{3}{2} < s \leq s_1(x)$	$(s-\frac{5}{4})x$	$\frac{1}{8}\left[\left(\frac{s-1}{s-\frac{3}{2}}\right)^2 + (\frac{1}{4}-x^2)\right]$
$s_1(x) \leq s \leq s_2(x)$	$\frac{1}{2}(s-1)x - \frac{1}{4}(s-\frac{5}{2})$	$\frac{1}{16}(5+6x)$
$s_2(x) \leq s$	0	$\frac{1}{2}\left[(\frac{1}{4}-x^2)s + x^2\right]$

3.4.3 Smallest trapping set of $\mathcal{M}_x$

Let us not forget that the original $\Sigma\Delta$ mapping is $\mathcal{M}_x$ as in (2.13). However, from (2.22), the mapping M full describes its dynamics. So S is positively invariant by $\mathcal{M}_x$ if and only if $S - c_x\mathbf{j}$ is positively invariant by M. The same thing can be said for trapping sets. From (3.59), one can easily see that

$$\Pi_{c_0,c_1} - c_x\mathbf{j} = \Pi_{c_0-c_x,c_1-c_x}.$$

Within the family of sets Π_{c_0,c_1}, we conclude that the smallest asymptotic trapping set of $\mathcal{M}_x$ is $\Pi_{c_0^*,c_1^*}$ such that $\Pi_{c_0^*-c_x,c_1^*-c_x} = \Lambda_{\delta^*}(\ell^*)$. Using (3.60) and (2.9), we have

$$c_i^* = \delta^* - \tfrac{x}{a_2} - \tfrac{1}{x_i}\,\ell^*. \tag{3.68}$$

Concerning the bound on u_1, we obtain particularly simple results by considering a trapping set that is even smaller than $\Pi_{c_0^*,c_1^*}$ and is derived from the following proposition.

Proposition 3.4.6 *If S is a trapping set of* M*, then the following subset $S' := \left\{\mathbf{u} \in S : \gamma_1 + x_1 \leq u_1 \leq \gamma_0 + x_0\right\}$ of S is also a trapping set of* M *with $\gamma_0 := \sup_{\mathbf{u}\in S\cap\Omega_0} u_1$ and $\gamma_1 := \inf_{\mathbf{u}\in S\cap\Omega_1} u_1$.*

This is proven in Appendix-C.18.

This in fact directly gives the bound $\max|u_1| = \max_{i=0,1}|\gamma_i + x_i|$. This also applies to $\mathcal{M}_x$ according to (2.22). After deriving γ_i with $S = \Pi_{c_0^*, c_1^*}$, we find

$$\max|u_1| = \begin{cases} \frac{s-1}{2s-3}, & \frac{3}{2} < s \leq s_1(x) \\ \frac{1}{4}(3+2x), & s_1(x) \leq s \end{cases}$$

where $s_1(x)$ is given in (3.67). Note that $\max|u_1|$ no longer depends on s when $s \geq s_1(x)$. The comparison of this result and prior work is shown in [29].

Chapter 4

Tile attractor preliminary

From Chapter 3, we have derived global trapping sets for various configurations. But our final goal is to prove that the attractor inside these trapping sets is a single tile.

4.1 The tiling Theorem

Firstly, we formally define the concepts of *attractor* and *tile*.

Definition 4.1.1 *The* a*ttractor of a mapping* M *in a positively invariant set* S *is the set*

$$A := \bigcap_{n\in\mathbb{N}} \mathrm{M}^n(S).$$

Definition 4.1.2 *A* t*ile in* $\mathbb{R}^n$ *is a set* Γ *such that*

$$(i)\ \bigcup_{\mathbf{k}\in\mathbb{Z}^n} \big(\Gamma + \mathbf{k}\big) = \mathbb{R}^n,$$
$$(ii)\ \Gamma \cap \big(\Gamma + \mathbf{k}\big) = \emptyset, \mathbf{k} \in \mathbb{Z}^n \setminus \mathbf{0}.$$

To prove that the attractor of M is a tile, we will use an advanced theorem from [28]. We rewrite the necessary part of the theorem as the following proposition.

Proposition 4.1.3 *The attractor of* M *in a positively invariant set is the union of a finite number of disjointed tiles up to a measure zero set.*

The measure is the *Lebesgue measure*, which is denoted by the function $m(\cdot)$. Actually, the theorem in [28] is more general than Proposition 4.1.3 because it proves the statement on a family of mappings which transform one tile to another. So Proposition 4.1.3 is the consequence of the theorem and the following proposition.

Proposition 4.1.4 *If the set* Γ *is a tile, so is* $\mathcal{M}_x(\Gamma)$.

This has been proven in [28] too.

When we use Proposition 4.1.3, we will basically ignore the measure zero set.

4.2 Area argument for tile attractor

From Proposition 4.1.3, the measure of an attractor must be an integer number. Since we have derived analytical trapping sets, we will intuitively try to prove its attractor is a tile by evaluating the area of the smallest trapping set available. If a trapping set has an area less than 2, the attractor is automatically a single tile.

We start with the smallest trapping sets Λ_{δ^*} given in Chapter 3. The sets are simply bounded by two parabolas. An example is shown in Figure 3.1(b'). We derive the area of the set Λ_ℓ as a function of x, s, and ℓ

$$m(\Lambda_\ell) = \frac{\sqrt{4(5-4s)x^2+4s-1}\,(4x^2-48\ell+s\,(2-8x^2)+1)}{96x^2-24}.$$

With the explicit parameters in Table 3.1, we find that

$$m(\Lambda_{\delta^*}) = \begin{cases} m_1, & \frac{3}{2} < s < s_1 \\ m_x, & s_1 \le s < s_2 \\ m_2, & s \ge s_2 \end{cases}, \tag{4.1}$$

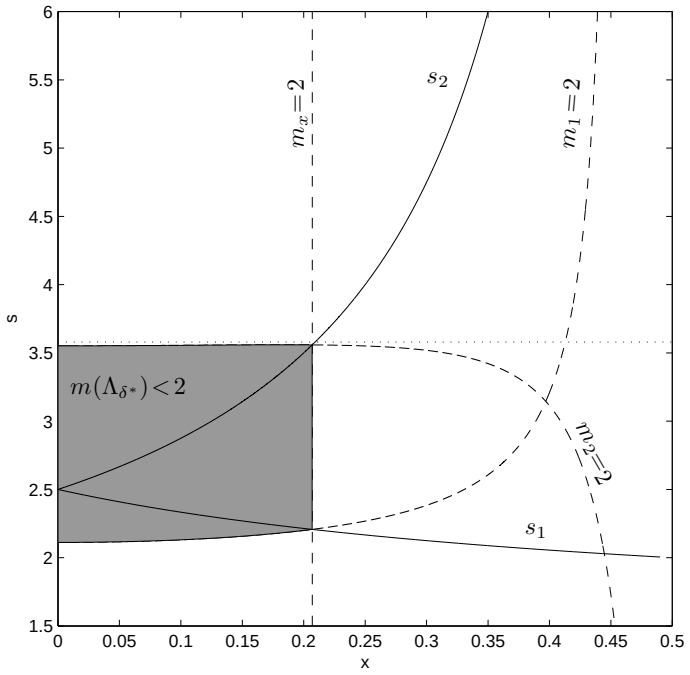


Figure 4.1: With the configuration inside the dark gray region, the area of trapping set Λ_{δ^*} is less than 2. Since (4.1) gives a piecewise definition, there are three dash lines representing the function value of 2 for each piece.

where

$$
\begin{aligned}
m_1 &= \frac{8(s-1)^3}{3(2s-3)^3\,(1-4x^2)},\\
m_x &= \frac{(2x+3)^3}{24-96x^2},\\
m_2 &= \frac{(4(5-4s)x^2+4s-1)^{3/2}}{24(1-4x^2)}.
\end{aligned}
$$

We show in Figure 4.1 the region of the parameters (x, s) leading to a trapping set of area less than 2. Unfortunately, the region is limited. When taking large x, the area is always larger than 2 even there is freedom to choose any s. On the other hand, a set containing no more than one tile may have an area more than 2. So we are going to find the exact necessary condition on a set which can contain two disjointed tiles.

4.3 Tile attractor condition

The goal of this section is to find general conditions of a set that can not contain more than one tile. We will find those configurations such that the trapping set Λ_{δ^*} satisfies these constrains. Then the corresponding attractor must be a tile.

We first introduce special set theoretic tools to recognize the presence of tiles. For any $S \subset \mathbb{R}^2$, we define

$$\mathcal{T}(S) := \overline{\bigcup_{\mathbf{k}\in\mathbb{Z}^2\setminus\{\mathbf{0}\}} (S+\mathbf{k})}. \tag{4.2}$$

Proposition 4.3.1 *Let S be a subset of $\mathbb{R}^2$.*

(i) S is a tile if and only if $\mathcal{T}(S) = S$.

(ii) S contains a tile if and only if $\mathcal{T}(S) \subset S$.

(iii) S contains two disjointed tiles if and only if $\mathcal{T}(S) = \emptyset$.

This is proven in Appendix-D.1 based on the following properties

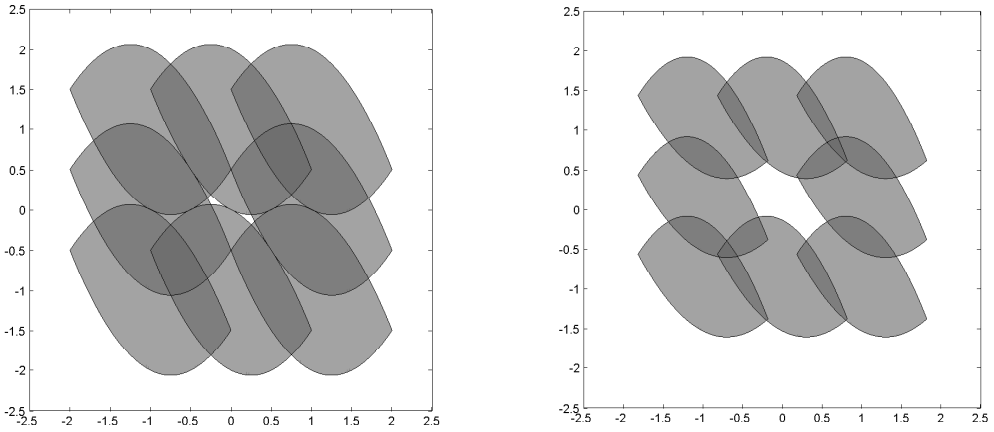
$$\mathcal{T}(A\cup B) = \mathcal{T}(A)\cap\mathcal{T}(B) \qquad \text{and} \qquad A\subset B \;\Rightarrow\; \mathcal{T}(B)\subset\mathcal{T}(A). \tag{4.3}$$

The above property (iii) has the following consequence.

Proposition 4.3.2 *For any set $S \subset \mathbb{R}^2$ such that $m(\mathcal{T}(S)) > 0$, S* cannot *contain two disjointed tiles up to a measure zero set.*

This is proven in Appendix-D.2.

Using Proposition 4.3.2, one can prove that the attractor of M is a tile by simply evaluating $m(\mathcal{T}(S))$. In Figure 4.2, we show examples that the set $\mathcal{T}(S)$ has no measure zero.

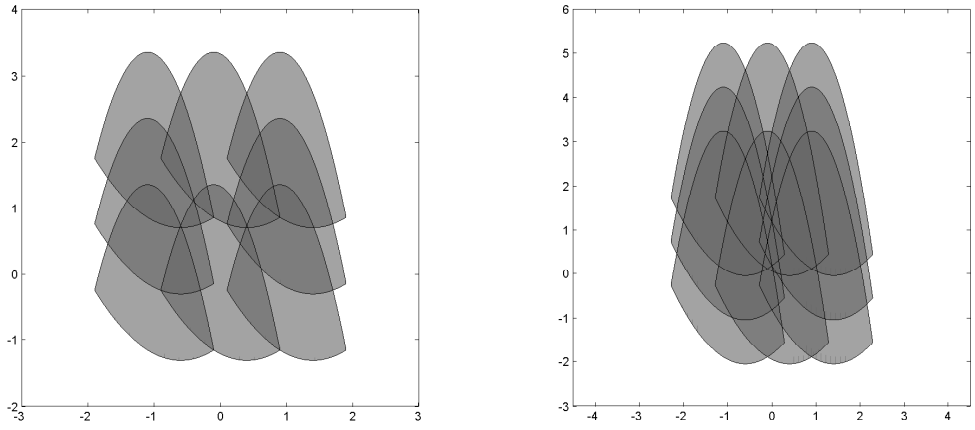


(a) (b)

Figure 4.2: Examples of $\mathcal{T}(\Lambda_{\delta^*})$ apparently have nonzero measure under configurations: (a) $x = 0$ and $s = 2$; (b) $x = 0.1$ and $s = 3$. The set $\mathcal{T}(\Lambda_{\delta^*})$ is left blank in the middle.

The value of $m(\mathcal{T}(\Lambda_{\delta^*}))$ can be algebraically derived. However, it is not worth to do the derivation as we numerically observe that $\mathcal{T}(\Lambda_{\delta^*}) = \emptyset$ when x is larger than certain value even for any s. Some examples are shown in 4.3. We find the configuration region such that $m(\mathcal{T}(\Lambda_{\delta^*})) > 0$ by experiments. The results is shown in Figure 4.4.

It appears that the derived smallest trapping set Λ_{δ^*} by using Lyapunov function is still too big to identify whether it contains one tile or more. We need to find smaller trapping sets. In the next chapter, we create new tools to do this by studying the fundamental dynamics.



(a) (b)

Figure 4.3: Examples of empty $\mathcal{T}(\Lambda_{\delta^*})$: (a) $x = 0.3$ and $s = 3$; (b) $x = 0.3$ and $s = 10$. No blank region is left in the middle.

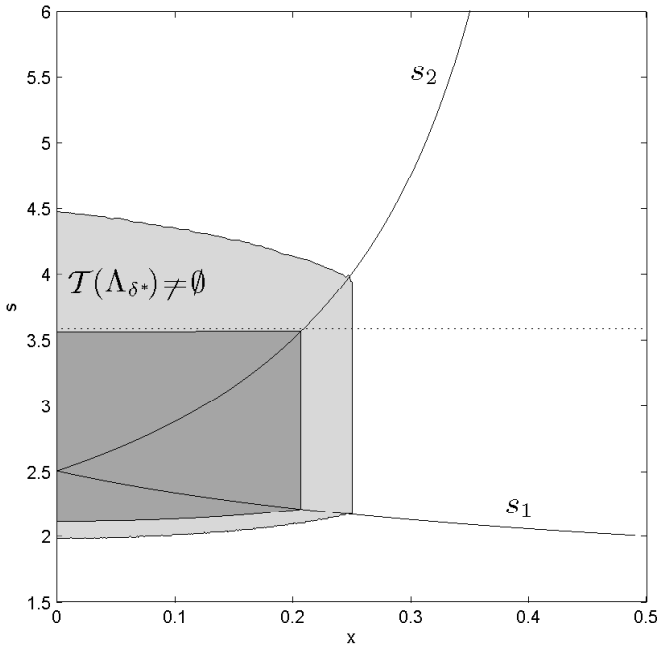


Figure 4.4: We numerically examine the measure of $\mathcal{T}(\Lambda_{\delta^*})$ and show those configurations which give nonzero measure with light gray. According to Proposition 4.3.2, Under these configurations, the attractor is a tile. Compared to the region(dark gray) from the area argument, the tile attractor is proven in more configurations with this method.

Chapter 5

Fundamental theorem of dynamics

In this chapter, we study the details of the dynamical behavior inside a positively invariant set. The goal is to establish a technique which enables us to obtain smaller trapping sets inside. In stead of study the dynamics of single point, we proceed with a set and study the process that it is transformed into the attractor.

Firstly, we use (2.12) define notation $\mathbf{u}[n]$ recursively with $\mathbf{u}[0] \in \mathbb{R}^2$ for a given input $x[n]$. Explicitly,

$$\mathbf{u}[n+1] := \mathcal{M}_{x[n]}\mathbf{u}[n]. \tag{5.1}$$

With this, define mappings

$$\tilde{\mathcal{M}}_n\mathbf{u} := \mathbf{u}[n], \text{ with } \mathbf{u}[0] := \mathbf{u}, \tag{5.2}$$

$$\tilde{\mathcal{M}}_{-n}(\mathbf{u}) := \{\mathbf{v} \in \mathbb{R}^2 \times \mathbb{R}^2 : \tilde{\mathcal{M}}_n\mathbf{v} = \mathbf{u}\}. \tag{5.3}$$

Note that $\tilde{\mathcal{M}}_{-n}(\mathbf{u})$ is usually a set including more than one point.

The mapping $\tilde{\mathcal{M}}_n$ can operate on a set as well as on a single point. For a given set S_0, $\tilde{\mathcal{M}}_n$ generates the following sequence of sets

$$S_n := \tilde{\mathcal{M}}_n(S_0).$$

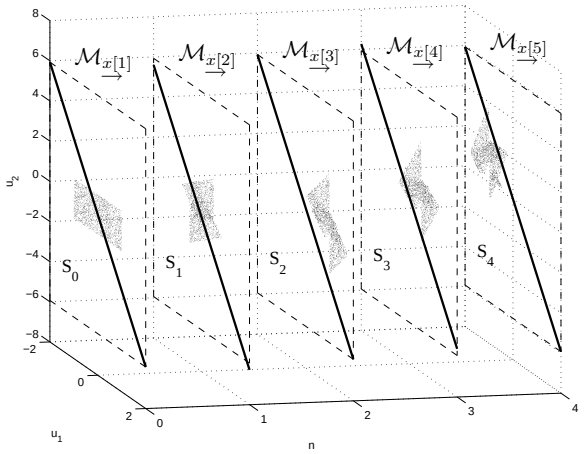


Figure 5.1: A sequence of sets is generated as $\tilde{\mathcal{M}}_n(S_0)$, with the set S_0 defined as $[-1, 1] \times [-1, 1]$.

An example is shown in Figure 5.1. Following the order of n, the sets S_n are presented as s sequence of slices in the 3D space. Note that the mapping $\mathcal{M}_{x[n]}$ changes according to $x[n]$.

5.1 Dynamics of pairs of points

There are two approaches to study the attraction of a mapping. One is to take any point outside and study its dynamics. The other is to take the whole positively invariant set and study how the set is transformed into the attractor. We now consider the second method.

From (2.12), one easily see that the mapping $\mathcal{M}_x$ is a piecewise function and each of two pieces is an area-preserve mapping. Then the area reduction of a set after applying the mapping $\mathcal{M}_x$ is purely because of the overlap between two parts of the set mapped by different pieces. So what sets will have overlap after mapping is of our interest. However, instead of an arbitrary set, we study a set containing only two points. The reason is that the piecewise mapping $\mathcal{M}_x$ has only two pieces.

An overlapped point has exact two distinct previous images which are mapped by different M_i.

The following results in this section is general. They apply to the mapping (2.13) where $\{\Omega_0, \Omega_1\}$ is any arbitrary partition of $\mathbb{R}^2$. To clearly associate with the mapping $\mathcal{M}_x$, we denote the partition as $\{\Omega_0^x, \Omega_1^x\}$. In (2.13), e.g., $\Omega_i^x = \Omega_i + c_x\,\mathbf{j}$.

In the space $\mathbb{R}^2 \times \mathbb{R}^2$, for any $x \in \mathbb{R}$, define:

$$\begin{aligned} \Pi_{-1}^x &:= \Omega_1^x \times \Omega_0^x, \\ \Pi_0^x &:= (\Omega_0^x \times \Omega_0^x) \cup (\Omega_1^x \times \Omega_1^x), \\ \Pi_1^x &:= \Omega_0^x \times \Omega_1^x. \end{aligned} \tag{5.4}$$

Since $\{\Omega_0^x, \Omega_1^x\}$ is a partition of $\mathbb{R}^2$, $\{\Pi_{-1}^x, \Pi_0^x, \Pi_1^x\}$ is a partition of $\mathbb{R}^2 \times \mathbb{R}^2$.

For any $\boldsymbol{w} = (\mathbf{u}, \mathbf{v}) \in \mathbb{R}^2 \times \mathbb{R}^2$, define:

$$\mathcal{M}_x\boldsymbol{w} := \big(\mathcal{M}_x\mathbf{u}, \mathcal{M}_x\mathbf{v}\big), \tag{5.5}$$

$$\mathbf{p}_{\mathrm{u}}(\boldsymbol{w}) := \mathbf{u} \quad \text{and} \quad \mathbf{p}_{\mathrm{v}}(\boldsymbol{w}) := \mathbf{v}, \tag{5.6}$$

$$\mathbf{d}(\boldsymbol{w}) := \mathbf{v} - \mathbf{u}, \tag{5.7}$$

$$d_1(\boldsymbol{w}) := v_1 - u_1 \quad \text{and} \quad d_2(\boldsymbol{w}) := v_2 - u_2. \tag{5.8}$$

Note that $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$.

With the above new definition , we have

$$\forall k \in \{-1, 0, 1\}, \quad \forall \boldsymbol{w} \in \Pi_k^x, \qquad \mathbf{d}(\mathcal{M}_x\boldsymbol{w}) = \mathrm{L}\,\mathbf{d}(\boldsymbol{w}) - k\,\mathbf{i}. \tag{5.9}$$

Similarly to the mapping of a single point, for a given $\boldsymbol{w}[0] \in \mathbb{R}^2 \times \mathbb{R}^2$, recursively define the sequence of pairs:

$$\boldsymbol{w}[n+1] := \mathcal{M}_{x[n]}\boldsymbol{w}[n]. \tag{5.10}$$

Following (5.10), define

$$\tilde{\mathcal{M}}_n \boldsymbol{w} := \boldsymbol{w}[n], \text{ with } \boldsymbol{w}[0] := \boldsymbol{w}, \tag{5.11}$$

$$\tilde{\mathcal{M}}_{-n}(\boldsymbol{w}) := \{\boldsymbol{w}' \in \mathbb{R}^2 \times \mathbb{R}^2 : \tilde{\mathcal{M}}_n \boldsymbol{w}' = \boldsymbol{w}\}. \tag{5.12}$$

Proposition 5.1.1

(i) If $d_1(\boldsymbol{w}[0]) \in \mathbb{Z}$*, then for all* $n \in \mathbb{N}$*,* $d_1(\boldsymbol{w}[n]) \in \mathbb{Z}$*.*

(ii) If $\mathbf{d}(\boldsymbol{w}[0]) \in \mathbb{Z}^2$*, then for all* $n \in \mathbb{N}$*,*$\mathbf{d}(\boldsymbol{w}[n]) \in \mathbb{Z}^2$*.*

(iii) If $\mathbf{d}(\boldsymbol{w}[0]) \notin \mathbb{Z}^2$*, then for all* $n \in \mathbb{N}$*,*$\mathbf{d}(\boldsymbol{w}[n]) \notin \mathbb{Z}^2$*.*

This is proven in Appendix-E.1.

The property (iii) in this proposition tells us that if any pair of points initially not having integer vector difference, they will never have integer vector difference with iteration. Hence they never be mapped into one point. So when we are looking for the reason of attraction, it might not be very useful to study the dynamics of such pair of points. On the other hand, the dynamics of such pair of points may tell us other information, such as the structure of an attractor. According to Proposition 4.1.4, if a set contains a tile, all its forwarding images will.

5.2 Fundamental dynamics of $\mathcal{M}_x$

Because of Proposition 5.1.1(iii), we now focus on the dynamics of those pairs of points with integer vector difference. Consider a set $S \subset \mathbb{R}^2$, define

$$\mathcal{W}_S := \left\{\boldsymbol{w} \in S^2 : \quad \mathbf{d}(\boldsymbol{w}) \in \mathbb{Z}^2\right\}.$$

We partition $\mathcal{W}_S$ into the following two sets,

$$\mathcal{X}_S := \big\{\boldsymbol{w} \in \mathcal{W}_S : \quad \exists n_0 \in \mathbb{N}, \forall n \geq n_0, \tilde{\mathcal{M}}_n \boldsymbol{w} \in \Pi_0^{x[n]}\big\} \qquad \text{and} \qquad \mathcal{Y}_S := \mathcal{W}_S \setminus \mathcal{X}_S. \tag{5.13}$$

Qualitatively, $\mathcal{X}_S$ is the set of pairs $(\mathbf{u}, \mathbf{v}) \in \mathcal{W}_S$ such that $\tilde{\mathcal{M}}_n \mathbf{u}$ and $\tilde{\mathcal{M}}_n \mathbf{v}$ are simultaneously in $\Omega_0^{x[n]}$ or in $\Omega_1^{x[n]}$ at every instant n larger than some $n_0 \geq 0$.

5.2.1 Dynamics of pairs in $\mathcal{X}_S$

The set $\mathcal{X}_S$ happens to include the following important subset

$$\mathcal{X}_S^0 := \big\{\boldsymbol{w} \in \mathcal{X}_S : \quad \exists n \in \mathbb{N},\ \mathbf{d}(\tilde{\mathcal{M}}_n \boldsymbol{w}) = \mathbf{0}\big\}. \tag{5.14}$$

Indeed, if $\boldsymbol{w} = (\mathbf{u}, \mathbf{v})$ and $\mathbf{d}(\tilde{\mathcal{M}}_{n_0} \boldsymbol{w}) = \mathbf{0}$, then $\tilde{\mathcal{M}}_{n_0} \mathbf{u} = \tilde{\mathcal{M}}_{n_0} \mathbf{v}$. This implies that $\tilde{\mathcal{M}}_{n_0} \mathbf{u} = \tilde{\mathcal{M}}_{n_0} \mathbf{v}$ for all $n \geq n_0$, and obviously $(\tilde{\mathcal{M}}_n \mathbf{u}, \tilde{\mathcal{M}}_n \mathbf{v}) \in \Pi_0$. Under certain weak conditions, we are going to show that $\mathcal{X}_S^0$ actually occupies most the "area" of $\mathcal{X}_S$. We formalize this by first defining the complementary set

$$\mathcal{X}_S^{\emptyset} := \mathcal{X}_S \setminus \mathcal{X}_S^0. \tag{5.15}$$

We then reduce it to a subset of $\mathbb{R}^2$ by performing its projection $\mathbf{p}_{\mathrm{u}}(\mathcal{X}_S^{\emptyset})$ onto the first vector component of the pairs. We finally measure its area, which we denote by $m\big(\mathbf{p}_{\mathrm{u}}(\mathcal{X}_S^{\emptyset})\big)$.

From (5.14) and (5.15),

$$\mathcal{X}_S^{\emptyset} = \big\{\boldsymbol{w} \in \mathcal{X}_S : \quad \forall n \in \mathbb{N},\ \mathbf{d}(\tilde{\mathcal{M}}_n \boldsymbol{w}) \neq \mathbf{0}\big\} \tag{5.16}$$

We take a pair $\boldsymbol{w} \in \mathcal{X}_S^{\emptyset}$. According to (5.16), $\boldsymbol{w} \in \mathcal{X}_S$. Then there exists $n_0 \in \mathbb{N}$ such that $\tilde{\mathcal{M}}_n \boldsymbol{w} \in \Pi_0^{x[n]}$ for all $n \geq n_0$. Now consider a subset of $\mathcal{X}_S^{\emptyset}$ such that $n_0 = 0$.

Define:

$$\Phi_S := \left\{ \boldsymbol{w} \in \mathcal{X}_S^{\emptyset} : \quad \tilde{\mathcal{M}}_n \boldsymbol{w} \in \Pi_0^{x[n]} \text{ for all } n \in \mathbb{N} \right\}. \tag{5.17}$$

This implies that

$$\mathcal{X}_S^{\emptyset} = \mathcal{W}_S \cap \bigcup_{n\in\mathbb{N}} \tilde{\mathcal{M}}_{-n}(\Phi_S). \tag{5.18}$$

Proposition 5.2.1 *For a given pair $\boldsymbol{w} \in \mathbb{R}^2 \times \mathbb{R}^2$, assume that for all $n \in \mathbb{N}$, $\tilde{\mathcal{M}}_n \boldsymbol{w} \in \Pi_0^{x[n]}$ and $\mathbf{p}_{\mathrm{u}}(\tilde{\mathcal{M}}_n \boldsymbol{w}) - \mathbf{p}_{\mathrm{v}}(\tilde{\mathcal{M}}_n \boldsymbol{w}) \in B$, where B is a bounded set. Then by necessity $d_1(\boldsymbol{w}) = 0$ and $\mathbf{d}(\tilde{\mathcal{M}}_n \boldsymbol{w}) = \mathbf{d}(\boldsymbol{w})$ for all $n \in \mathbb{N}$.*

This is proven in Appendix-E.2.

Proposition 5.2.2 *Assume that for all $n \in \mathbb{N}$, the partition boundary of $\{\Omega_0^{x[n]}, \Omega_1^{x[n]}\}$ in mapping $\mathcal{M}_x$ is a continuous function. For a measurable set S, define $S_n := \tilde{\mathcal{M}}_n(S)$. Assume that there exists a bounded set B such that $\forall \mathbf{u}, \mathbf{v} \in S_n$, $\mathbf{u} - \mathbf{v} \in B$, then*

$$m(\mathbf{p}_{\mathrm{u}}(\Phi_S)) = 0. \tag{5.19}$$

This is proven in Appendix-E.3.

From (5.18), (5.5) and (5.6),

$$\mathbf{p}_{\mathrm{u}}(\mathcal{X}_S^{\emptyset}) = S \cap \bigcup_{n\in\mathbb{N}} \tilde{\mathcal{M}}_{-n}(\mathbf{p}_{\mathrm{u}}(\Phi_S)). \tag{5.20}$$

When S satisfies conditions of Proposition 5.2.2, $\mathbf{p}_{\mathrm{u}}(\Phi_S)$ is a measure zero set. Since the right hand side of (5.20) is the union or intersection of countable many measure zero set, it is measure zero, too.

$$m(\mathbf{p}_{\mathrm{u}}(\mathcal{X}_S^{\emptyset})) = 0. \tag{5.21}$$

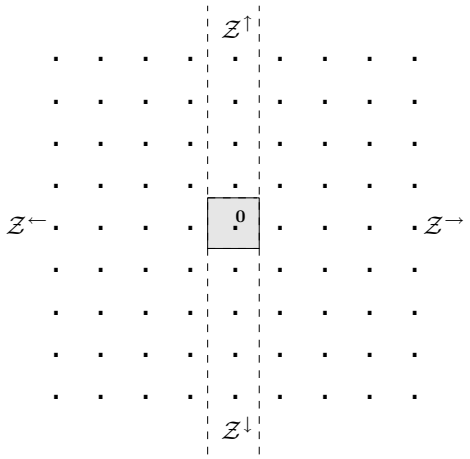


Figure 5.2: Partition of $\mathbb{Z}^2 \setminus \{\mathbf{0}\}$.

5.2.2 Dynamics of pairs in $\mathcal{Y}_S$

Since $\mathcal{X}_S^{\emptyset} \subset \mathcal{X}_S$, $\mathcal{Y}_S$ is disjointed with $\mathcal{X}_S^{\emptyset}$. As a particular consequence, for all $\boldsymbol{w} \in \mathcal{Y}_S$, $\mathbf{d}(\boldsymbol{w}) \in \mathbb{Z}^2 \setminus \{\mathbf{0}\}$. Let us partition $\mathcal{Z} := \mathbb{Z}^2 \setminus \{\mathbf{0}\}$ into the following four sets

$$\mathcal{Z}^{\leftarrow} := \mathbb{Z}^- \times \mathbb{Z}, \qquad \mathcal{Z}^{\rightarrow} := \mathbb{Z}^+ \times \mathbb{Z}, \qquad \mathcal{Z}^{\downarrow} := \{0\} \times \mathbb{Z}^-, \qquad \mathcal{Z}^{\uparrow} := \{0\} \times \mathbb{Z}^+.$$

We show this partition in Figure 5.2. Next, we are going to discover the dynamics of $\mathrm{M}^n\boldsymbol{w}$ over $\mathcal{Y}_S$.

Firstly, we define

$$\mathcal{Y}_S^{-} := \mathcal{Y}_S \cap \mathcal{Z}^{\leftarrow}, \qquad \mathcal{Y}_S^{+} := \mathcal{Y}_S \cap \mathcal{Z}^{\rightarrow}, \qquad \mathcal{Y}_S^{\downarrow} := \mathcal{Y}_S \cap \mathcal{Z}^{\downarrow}, \qquad \mathcal{Y}_S^{\uparrow} := \mathcal{Y}_S \cap \mathcal{Z}^{\uparrow}. \tag{5.22}$$

Theorem 5.2.3 *Under the same conditions as in Proposition 5.2.2,*

(i) For all $\boldsymbol{w} \in \mathcal{Y}_S^{-} \cup \mathcal{Y}_S^{\uparrow}$, $\exists n_0 \in \mathbb{N}$, $\forall n = 0, \ldots, n_0 - 1$, $\tilde{\mathcal{M}}_n \boldsymbol{w} \in \mathcal{Y}_S^{-} \cup \mathcal{Y}_S^{\uparrow}$ *and* $\tilde{\mathcal{M}}_{n_0} \boldsymbol{w} \in \mathcal{Y}_S^{\downarrow}$,

(ii) For all $\boldsymbol{w} \in \mathcal{Y}_S^{+} \cup \mathcal{Y}_S^{\downarrow}$, $\exists n_0 \in \mathbb{N}$, $\forall n = 0, \ldots, n_0 - 1$, $\tilde{\mathcal{M}}_n \boldsymbol{w} \in \mathcal{Y}_S^{+} \cup \mathcal{Y}_S^{\downarrow}$ *and* $\tilde{\mathcal{M}}_{n_0} \boldsymbol{w} \in \mathcal{Y}_S^{\uparrow}$.

This is proven in Appendix-E.4.

Graphically, this theorem implies that the integer valued vector $\mathbf{d}(\tilde{\mathcal{M}}_n \boldsymbol{w})$ has a sort of anticlockwise and perpetual motion around the zero vector. The following proposition tells more details.

Proposition 5.2.4 *Under the same conditions as in Theorem 5.2.3,*

(i) For all $\boldsymbol{w} \in \mathcal{Y}_S$, $\exists n \in \mathbb{N}$, $\mathbf{d}(\tilde{\mathcal{M}}_n \boldsymbol{w}) \in \{-1\} \times -\mathbb{N}$,

(ii) For all $\boldsymbol{w} \in \mathcal{Y}_S$, $\exists n \in \mathbb{N}$, $\mathbf{d}(\tilde{\mathcal{M}}_n \boldsymbol{w}) \in \{1\} \times \mathbb{N}$.

This is proven in Appendix-E.5.

5.3 DC inputs case

With a constant input x, the mapping $\tilde{\mathcal{M}}_n$ in Chapter 5 is simplified as M^n, where M is defined in (2.21). The sets $\mathcal{X}_S$ and $\mathcal{X}_S^0$ are simplified too.

$$\mathcal{X}_S = \big\{ \boldsymbol{w} \in \mathcal{W}_S : \exists n_0 \in \mathbb{N}, \forall n \geq n_0, \mathrm{M}^n \boldsymbol{w} \in \Pi_0 \big\},$$

$$\mathcal{X}_S^0 = \big\{ \boldsymbol{w} \in \mathcal{X}_S : \exists n_0 \in \mathbb{N}, \forall n \geq n_0, \mathbf{d}(\mathrm{M}^n \boldsymbol{w}) = \mathbf{0} \big\}.$$

Set $\mathcal{Y}_S = \mathcal{W}_S \setminus \mathcal{X}_S$ and $\mathcal{X}_S^{\emptyset} = \mathcal{X}_S \setminus \mathcal{X}_S^0$ are defined accordingly.

Take a positively invariant set S, assume that it is bounded and measurable, for any $n \in \mathbb{N}$, $S_n := \mathrm{M}^n(S) \subset S$, which is bounded. The partition boundary of $\{\Omega_0, \Omega_1\}$ is clearly a continuous function. So all conditions in Proposition 5.2.3, 5.2.4 and (5.21) are met. Therefore, simplified propositions for DC input case will follow from them, respectively. Next, we propose a fundamental theorem based on these fundamental dynamics propositions.

Theorem 5.3.1 *Consider a mapping* M *as in* (2.21) *with a continuous function to separate the partition* $\{\Omega_0, \Omega_1\}$*. Assume that there are two bounded and measurable positively invariant sets* S *and* S'*, and* S' *contains at least one tile. Also assume that* S' *is in the form of* $S' = \{(u_1, u_2) \in S : u_2 \geq f(u_1)\}$ *or* $S' = \{(u_1, u_2) \in S : u_2 < f(u_1)\}$*, where* f *is a real function. Then for all* $\mathbf{u} \in S \setminus \mathbf{p}_{\mathrm{u}}(\mathcal{X}_S^{\emptyset})$*, there exists* $n \in \mathbb{N}$ *such that* $\mathrm{M}^n\mathbf{u} \in S'$*. And set* $\mathbf{p}_{\mathrm{u}}(\mathcal{X}_S^{\emptyset})$ *is measure zero.*

This is proven in Appendix-E.6.

Theorem 5.3.1 tells us that for any positively invariant set, once we can find a function to partition it into two subsets and one of them is positively invariant, then the positively invariant subset is a trapping set. This is the new tool that enables us to obtain smaller trapping sets.

This page is left blank on purpose !

Chapter 6

Global tile attractor in DC inputs case

A new tool, Theorem 5.3.1 from the last chapter, enables us to find smaller trapping set from a bigger one. As a stepping stone, we first derive a global trapping set by using the knowledge of Chapter 3. Then we use the new tool to obtain smaller trapping set. Our final goal will be achieved when new trapping set can not contain more than one tile.

6.1 A global trapping set

We define

$$T := \{\mathbf{u} : \quad t(\mathbf{u}) = 0\}, \tag{6.1}$$

where t is defined in (2.8) and

$$T^i := M_i(T), \qquad \Omega_1^i := \mathrm{M}_i(\Omega_1) \tag{6.2}$$

We consider a special point $\mathbf{p}$ defined by

$$\mathbf{p} := T^1 \cap \mathrm{M}_0^{-1}(T^1). \tag{6.3}$$

For any parameters x and s, we derive a global trapping set as like in Chapter 3. We use a Lyapunov function h defined in (3.20), where $h^i := p_i$, where p_i is defined in (3.16). We then have the following proposition.

Proposition 6.1.1

$$h(\mathbf{p}) > \ell_h \tag{6.4}$$

under condition

Condition 6.1.2

$$s > \frac{5}{2}.$$

This is proven in Appendix-F.1.

It follows from Proposition 3.2.3 and 6.1.1 that the set

$$R_0 := \Lambda_h(h(\mathbf{p}))$$

is a global trapping set. An example is shown in Figure 6.1(a).

6.2 Smaller global trapping set

We use Theorem 5.3.1 to derive trapping sets in R_0. We want to make the trapping sets as small as impossible. Let us first introduce the notation

$$G_\ell := \big\{\mathbf{u} :\ h_1(\mathbf{u}) < \ell\big\}. \tag{6.5}$$

The set $G = G_\ell$ satisfies the properties

$$\forall d \geq 0, \qquad G - d\mathbf{j} \ \subset\ G \qquad \text{and} \qquad \overline{G} + d\mathbf{j} \ \subset\ \overline{G}, \tag{6.6}$$

$$\mathrm{M}_1(G) = G \qquad \text{and} \qquad \mathrm{M}_1(\overline{G}) = \overline{G}, \tag{6.7}$$

where $\bar{G} := \mathbb{R}^2 \setminus G$.

Proposition 6.2.1 *The set*

$$R_1 := R_0 \cap \big(\Omega_1^1 \cap \overline{G_e} \cap \mathrm{M}_0(\overline{G_e})\big) \tag{6.8}$$

is positively invariant with

$$e := \min_{\mathbf{v}\in \mathrm{M}_0(\Omega_1^1)} h_1(\mathbf{v}). \tag{6.9}$$

This is proven in Appendix-F.2.

Let us define $\mathbf{d}$ to be the point of $\partial G_e \cap T$ of minimal abscissa u_1 and the point

$$\mathbf{c} := \mathrm{M}_0\mathbf{d}. \tag{6.10}$$

Proposition 6.2.2 *The set*

$$R_2 := R_1 \cap G_{e'} \tag{6.11}$$

is positively invariant with

$$e' := \max\big(h_1(\mathbf{c}), h_1(\mathbf{b}_0)\big). \tag{6.12}$$

This is proven in Appendix-F.3.

Proposition 6.2.3 *For almost all* $\mathbf{u} \in \mathbb{R}^2$*, there exists* n *such that* $\mathrm{M}^n\mathbf{u} \in R_2$.

Proof: According to Proposition 3.2.3, R_0 traps any point within finite iterations. Due to (6.6), we have $F+\mathbf{j} \subset F$ and $G_{e'} - \mathbf{j} \subset G_{e'}$ with $F := \Omega_1^1 \cap \overline{G_e} \cap \mathrm{M}_0(\overline{G_e})$. Since both R_0 and R_1 are positively invariant, Theorem 5.3.1 applies.

■

6.3 Tile attractor

We recall from Proposition 4.3.1 that $m(\mathcal{T}(R_2) > 0$ is a sufficient condition to prevent R_2 to contain two disjointed tiles. This measure is unfortunately not easy to derive directly.

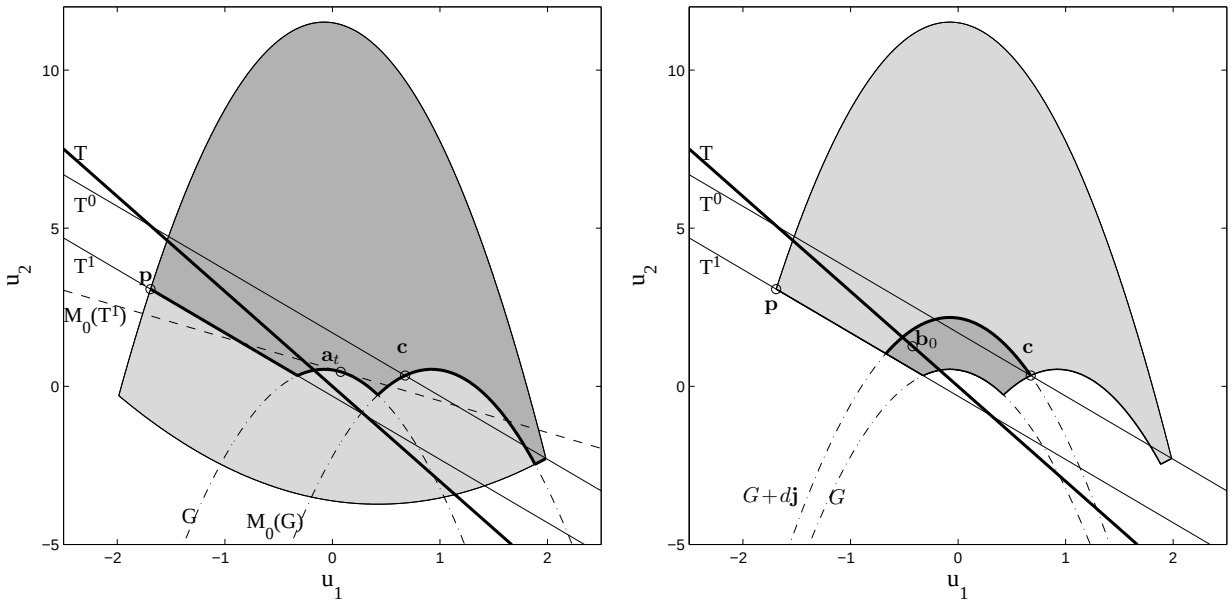


Figure 6.1: Global trapping sets: (a) R_0; (b) R_1.

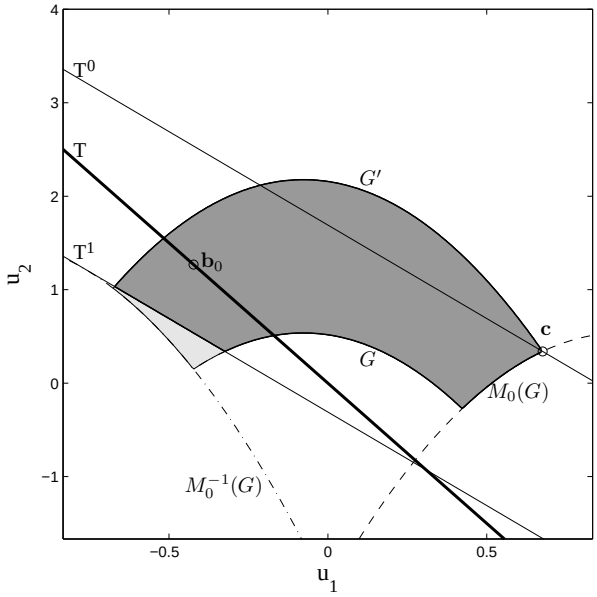


Figure 6.2: Global trapping set R_2 (dark gray area) and superset $\mathcal{Q}_d(G_e)$ (whole gray area).

We proceed by finding a superset R of R_2 such that $m(\mathcal{T}(R))$ can be derived. Having $m(\mathcal{T}(R)) > 0$ will be a sufficient condition for $m(\mathcal{T}(R_2)) > 0$, since $\mathcal{T}(R) \subset \mathcal{T}(R_2)$. For any $S \subset \mathbb{R}^2$ and $(a, b) \in \mathbb{R}^2$, let us introduce the notation

$$S_{a,b} := S + (a, b).$$

The following proposition provides a particularly convenient superset of R_2.

Proposition 6.3.1

$$R_2 \subset \mathcal{Q}_d(G_e) \tag{6.13}$$

where for any $S \subset \mathbb{R}^2$,

$$\mathcal{Q}_d(S) := \overline{S}_{-1,1} \cap \overline{S} \cap \overline{S}_{1,0} \cap S_{0,d} \tag{6.14}$$

and

$$d := -\frac{1}{x_1}(e' - e). \tag{6.15}$$

Proof: From (6.8) and (6.11), $R_2 \subset R$ with

$$R := \left(\Omega_1^1 \cap \overline{G_e} \cap \mathrm{M}_0(\overline{G_e})\right) \cap G_{e'}.$$

By definition of e in (6.9), $h_1(\mathbf{v}) \geq e$ for all $\mathbf{v} \in \mathrm{M}_0(\Omega_1^1)$. So $M_0(\Omega_1^1) \subset \overline{G_e}$. Meanwhile, $G_{e'} = G_e - \frac{1}{x_1}(e' - e)\mathbf{j}$ due to (6.5) and (3.13). So

$$R \ \subset \ \mathrm{M}_0^{-1}(\overline{G_e}) \cap \overline{G_e} \cap \mathrm{M}_0(\overline{G_e}) \cap \left(G_e + d\mathbf{j}\right).$$

with d defined in (6.15). Next, using (2.16) and (6.7), we get $\mathrm{M}_0(G_e) = G_e + \mathbf{i}$. This in turn implies that $\mathrm{M}_0^{-1}(G) = G - \mathrm{L}^{-1}\mathbf{i} = G - \mathbf{i} + \mathbf{j}$. ■

The set $\mathcal{Q}_d(G_e)$ is shown in Figure 6.2.

Proposition 6.3.2 *For any $d \in [1,2)$ and any set of the type $G = G_\ell$,*

$$\mathcal{T}\big(\mathcal{Q}_d(G)\big) \supset \mathcal{Q}_{2-d}(G) + (d-1)\mathbf{j}. \tag{6.16}$$

This is proven[1] in Appendix-F.4.

From (6.13), one can always write $R_2 \subset \mathcal{Q}_d(G_e)$ with $d = \max(\frac{e'-e}{-x_1}, 1)$. Assume that $\frac{e'-e}{-x_1} < 2$. Then $d \in [1,2)$ and Proposition 6.3.2 implies that $\mathcal{T}(R_2) \supset \mathcal{Q}_{2-d}(G_e) + (d-1)\mathbf{j}$. It is easy to see from (6.14) that $m\big(\mathcal{Q}_c(G)\big) > 0$ if and only if $c > 0$. Since $2 - d > 0$, then $m(\mathcal{T}(R_2)) > 0$ and Proposition 4.3.2 implies that R_2 cannot contain two tiles up to a set of measure 0. This finally proves that the attractor is a single tile since Proposition 4.1.3. The following proposition shows the actual range of configurations we have proved to guarantee tile attractor. As shown in Figure 6.4 it covers most configurations we are interested in.

Proposition 6.3.3 *Under Conditions 2.4.4 and 6.1.2, the values e and e' of* (6.9) *and* (6.12) *satisfy $\frac{e'-e}{-x_1} < 2$ if and only if $s_1(x) < s(x) < s_2(x)$, where*

$$s_1(x) := \frac{2 - 3\,x - \sqrt{1-2\,x}}{\frac{1}{2} - x} \qquad \textit{and} \qquad s_2(x) := \frac{1 - x + \sqrt{1-2\,x}}{\frac{1}{2} - x}. \tag{6.17}$$

This is proven in AppendixF.5

[1]The two sets of (6.16) can actually be proven to be equal (see Appendix F.6). This implies as a particular consequence that $\mathcal{T}(\mathcal{Q}_1(G)) = \mathcal{Q}_1(G)$, which in turn implies the interesting result that $\mathcal{Q}_1(G)$ is an exact tile.

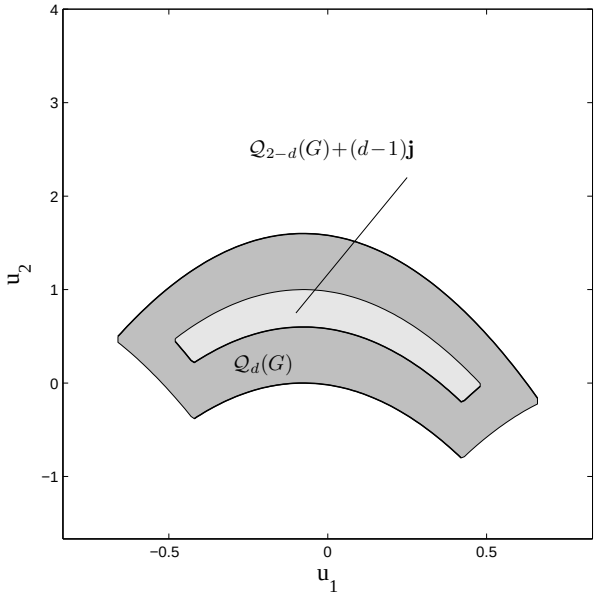


Figure 6.3: Set $\mathcal{Q}_d(G)$ (which includes both dark gray and light gray area) and its subset $\mathcal{Q}_{2-d}(G) + (d-1)\mathbf{j}$ (light gray area only).

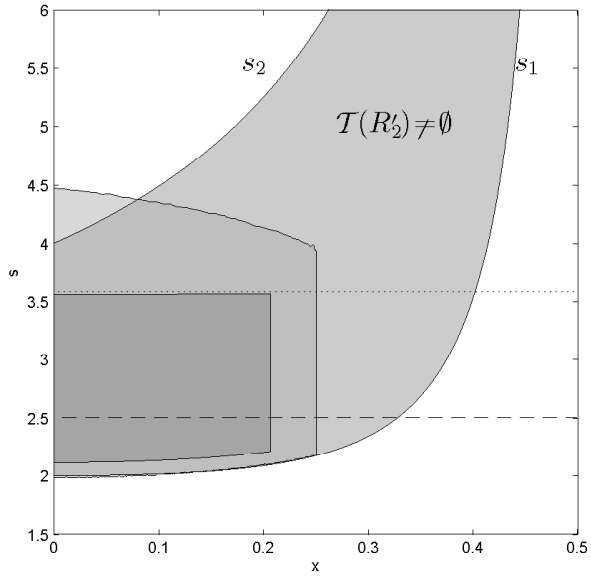


Figure 6.4: Range of configurations such that the attractor of the mapping defined by a gray marked configuration is a single tile.

This page is left blank on purpose !

Chapter 7

AC inputs case

In this chapter, we study the attraction of second order $\Sigma\Delta$ modulation with time varying inputs. Since it is proven that global stability does not hold for arbitrary time varying singles [15] [25], we only focus on the case where the input $x[n]$ is a finite sum of sinusoids. We will prove global stability of such inputs and the existence of *tile attractor*.

7.1 Dynamics equations of AC inputs

Assume the input $x[n]$ is of the form

$$x[n] = x + \tilde{x}[n], \tag{7.1}$$

where

$$\tilde{x}[n] = \sum_{k=1}^{K} a_k \cos(\omega_k n + \theta_k). \tag{7.2}$$

We extend the method introduced in [9] for this type of inputs. The method is based on the fact that the recurrence equation

$$\mathbf{y}[n+1] = \mathrm{L}\mathbf{y}[n] + \tilde{x}[n]\,\mathbf{i} \tag{7.3}$$

has a bounded solution in $\mathbf{y}[n]$. Indeed, it is shown in Appendix-G.1 that

$$\mathbf{y}[n] := (y_1[n], y_2[n])^\top \tag{7.4}$$

with

$$y_1[n] = \sum_{k=1}^{K} \frac{-a_k}{2\sin(\omega_k/2)} \cos(\omega_k n + \theta_k - \tfrac{\omega_k+\pi}{2})$$
$$y_2[n] = \sum_{k=1}^{K} \frac{a_k}{4\sin^2(\omega_k/2)} \cos(\omega_k n + \theta_k - \omega_k - \pi)$$

satisfies (7.3). Then, by defining the new state variable

$$\mathbf{v}[n] := \mathbf{u}[n] - \mathbf{y}[n], \tag{7.5}$$

and by taking the difference between (2.5) and (7.3), one obtains the recurrence equation

$$\mathbf{v}[n+1] := \mathrm{L}\mathbf{v}[n] + (x - q[n])\,\mathbf{i}. \tag{7.6}$$

This looks like a constant-input state equation. This is however not the case because $q[n]$ follows the rule of (2.7) instead of (2.20). We also have

$$\begin{aligned} t(\mathbf{u}[n]) \geq c_{x[n]} \quad &\Leftrightarrow \quad t(\mathbf{v}[n] + \mathbf{y}[n]) \geq c_{x[n]} \\ &\Leftrightarrow \quad t(\mathbf{v}[n]) \geq c_{x[n]} - y_2[n] - s\,y_1[n]. \end{aligned}$$

This leads to

$$q[n] = \begin{cases} -\frac{1}{2}, & t(\mathbf{v}[n]) < d[n] \\ \frac{1}{2}, & t(\mathbf{v}[n]) \geq d[n] \end{cases} \tag{7.7}$$

with

$$d[n] := c_{x[n]} - y_2[n] - s\,y_1[n]. \tag{7.8}$$

This is equivalent to

$$q[n] = \begin{cases} -\frac{1}{2}, & t(\mathbf{v}[n] - d[n]\,\mathbf{j}) < 0 \\ \frac{1}{2}, & t(\mathbf{v}[n] - d[n]\,\mathbf{j}) \geq 0 \end{cases} \tag{7.9}$$

From (7.6), (2.6), and (2.15), the new state variable is also equal to

$$\begin{aligned}\mathbf{v}[n+1] &= \mathrm{L}\mathbf{v}[n] + (x - q[n])\,\mathbf{i} - d[n]\,\mathbf{j} + d[n]\,\mathbf{j} \\ &= \mathrm{L}(\mathbf{v}[n] - d[n]\,\mathbf{j}) + (x - q[n])\,\mathbf{i} + d[n]\,\mathbf{j}. \end{aligned} \tag{7.10}$$

We compare the right hand side of (7.10) and (7.9) to (2.21), then obtain

$$\mathbf{v}[n+1] = \mathrm{M}\big(\mathbf{v}[n] - d[n]\,\mathbf{j}\big) + d[n]\,\mathbf{j}. \tag{7.11}$$

Let us change variable again,

$$\boldsymbol{w}[n] := \mathbf{v}[n] - d[n].$$

Then (7.11) follows from (7.8), (2.9), and (7.3) that

$$\boldsymbol{w}[n+1] = \mathrm{M}\boldsymbol{w}[n] + (d[n] - d[n+1])\,\mathbf{j}, \tag{7.12}$$

We then define

$$\mathbb{M}_n\,\mathbf{u} := \mathbb{T}_n\,\mathrm{M}\,\mathbf{u} \tag{7.13}$$

where $\mathbb{T}\mathbf{u} := \mathbf{u} + (d[n] - d[n+1])\,\mathbf{j}$. The equation (7.12) is rewritten as

$$\boldsymbol{w}[n+1] = \mathbb{M}_n\boldsymbol{w}[n]. \tag{7.14}$$

As shown in (7.13), the dynamics of AC inputs is very similar to that of DC inputs. For any $n \in \mathbb{N}$, the mapping $\mathbb{M}_n$ is decomposed as a nonlinear mapping M followed by a linear translation $\mathbb{T}_n$. Moreover, M is defined by the DC component x. And $\mathbb{T}_n$ is defined by the AC components $\tilde{x}[n]$. An intuition is to prove the stability of (7.12) by finding a sequence of sets $S_{n+1} := \mathbb{T}_n(S_n)$, for $n \in \mathbb{N}$ such that every set S_n is positively invariant by M. So the analysis of the stability of DC inputs case is actually crucial.

7.2 Global stability

We first introduce some notation for AC input case. We define a "positively invariant sequence", any set sequence Γ_n such that

$$\mathbb{M}_n(\Gamma_n) \subset \Gamma_{n+1}, \ \textit{for all } n \in \mathbb{N}. \tag{7.15}$$

Note that the "positively invariance" here is in the sense of sequence because a sequence of mappings is involved. Next, we define an "invariant sequence", any sequence Γ_n such that for all $n \in \mathbb{N}$,

(i) $\Gamma_{n+1} = \mathbb{M}_n(\Gamma_n)$,

(ii) $m(\Gamma_n) = m(\Gamma_{n+1})$,

where $m(.)$ denote the function of measure. In an invariant sequence, the set at each instance may change because the mapping changes. But the condition (ii) requires that all sets are "irreducible" in the sense of measure. For a given sequence S_n that satisfies the above (i), any sequence Γ_n such that for all $n \in \mathbb{N}$,

(i) Γ_n is positively invariant,

(ii) $\forall \mathbf{u} \in S_0, \exists n \in \mathbb{N}$, $\mathbb{M}_n \mathbf{u} \in \Gamma_n$,

Γ_n is called a "trapping sequence" of S_n. Finally, an "attractor sequence" is defined as a trapping and invariant sequence.

It is the time to analyze the stability of the sequence of mappings $\mathbb{M}_n$. By comparing (7.13) with (7.15), we find that a sequence S_n such that for all $n \in \mathbb{N}$,

[i] $\mathrm{M}(S_n) \subset S_n$,

[ii] $\mathbb{T}(S_n) = S_{n+1}$,

is a positively invariant sequence since

$$\mathbb{M}(S_n) = \mathbb{T}(\mathrm{M}(S_n)) \subset \mathbb{T}(S_n) = S_{n+1}.$$

This requires however our Lyapunov-function based framework, without which such a sequence is difficult to find. We have the following proposition.

Proposition 7.2.1

$$S_n := \Lambda_{-d[n]}(\ell), \ \textit{where } \ell \geq \tilde{\ell} \tag{7.16}$$

is a positively invariant sequence.

Proof: Under our framework, especially, Proposition 3.4.3, we easily build a family of positively invariant sets

$$\left\{\Lambda_{h_{d[n]}}(\ell)\right\}_{n\in\mathbb{N},\ell\geq\ell_{d[n]}}$$

based on a family of Lyapunov functions $h_{d[n]}$. Since $d[n]$ is bounded, the following value exists and unique:

$$\tilde{\ell} := \sup_{n\in\mathbb{N}} \ell_{d[n]}. \tag{7.17}$$

Therefore, for all $\ell \geq \tilde{\ell}$, the set $\Lambda_{h_{d[n]}}(\ell)$ is positively invariant by M. We then choose from the family a sequence of sets as in (7.16) which automatically satisfies above [i]. From (3.58), (3.2), and (3.32), we obtain that

$$\begin{aligned}
\Lambda_\delta(\ell) &= \{\mathbf{u} \in \mathbb{R}^2 : \quad h_\delta(\mathbf{u}) \leq \ell\} \\
&= \{\mathbf{u} \in \mathbb{R}^2 : \quad h_0(\mathbf{u} - \delta\,\mathbf{j}) \leq \ell\} \\
&= \{\mathbf{u} \in \mathbb{R}^2 : \quad h_0(\mathbf{u}\,\mathbf{j}) \leq \ell\} + \delta\,\mathbf{j} \\
&= S + \delta\,\mathbf{j},
\end{aligned}$$

where $S := \Lambda_0(\ell)$. By applying this to (7.16), we have

$$S_n = S - d[n]\,\mathbf{j}.$$

The sequence S_n clearly satisfies above [ii]. So the proof is complete.

■

The next question is whether S_n is a trapping sequence or not? We have the following proposition to show it is when $\ell > \tilde{\ell}$.

Proposition 7.2.2 *Assume that $\ell > \tilde{\ell}$, S_n defined as in Proposition 7.2.1 is a global trapping sequence.*

Proof: According to Proposition 7.2.1, S_n is a positively invariant sequence. So we only need to prove that for any sequence $\mathbf{u}[n+1] := \mathbb{M}_n\mathbf{u}[n]$, with $\mathbf{u}[0] \in \mathbb{R}^2$, there exists $n \in \mathbb{N}$ such that $\mathbf{u}_n \in S_n$.

We first according to (3.58) and (3.2), rewrite the set S_n

$$S_n = \big\{\mathbf{u} : \quad h_{d[n]}(\mathbf{u}) \le \ell\big\}. \tag{7.18}$$

We then define a Lyapunov sequence for the sequence $\mathbf{u}[n]$ as

$$h[n] := h_{d[n]}(\mathbf{u}[n]). \tag{7.19}$$

The following result is obtained from (7.19) and (3.32).

$$\begin{aligned} h[n+1] &= h_{d[n+1]}(\mathbb{M}_n\mathbf{u}[n]) \\ &= h_{d[n+1]}(\mathrm{M}\mathbf{u}[n] + d[n]\,\mathbf{j} - d[n+1]\,\mathbf{j}) \\ &= h_0(\mathrm{M}\mathbf{u}[n] + d[n]\,\mathbf{j}) \\ &= h_{d[n]}(\mathrm{M}\mathbf{u}[n]). \end{aligned}$$

It follows that

$$\Delta[n] := h[n+1] - h[n] = \Delta_{h_{d[n]}}(\mathbf{u}[n]).$$

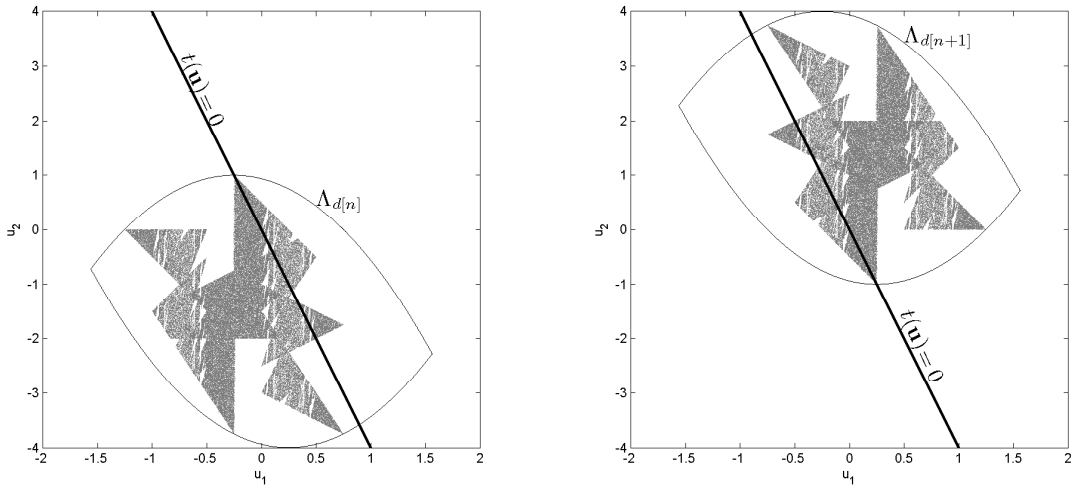
When $\mathbf{u}[n]$ is out of the trapping set S_n, $\Delta[n] \leq -\varepsilon$ according to Proposition 3.4.3, where ε is a constant which determined by ℓ and h. Therefore, $h[n] \leq h[0] - n\varepsilon$, which guarantees the existence of n such that $h[n] \leq \ell$. It finally proves the proposition with (7.19) and (7.18).

■

Note that the existence of trapping sets only depends on the DC component x and the mapping parameter s. The AC components have control on the size and position of the sets. In other words, the system can be stable even if the input is a high amplitude sinusoid signal.

7.3 Application

From last section, the size of the positively invariant sets mainly depends on the interval D, which bounds the amplitude of $d[n]$. According to (7.8) and (7.4), the amplitude of $d[n]$ is usually large when $\tilde{x}$ contains low frequency components. So the sets should give more accurate estimation of state variable bounds on bandpass $\Sigma\Delta$ modulation than lowpass $\Sigma\Delta$ modulation. We show an example in Figure 7.1 to illustrate how tight the trapping sets we derived vs. the invariant sets generated by experiments.



(a) (b)

Figure 7.1: Consider the time varying input $x[n] = cos(n\pi)$, which is a 1 and -1 alternative sequence. Since global stability has been proven. Any point must be mapped into a trapping sequence, e.g. S_n. In the space after changing variables, the trapping sequence are shown as well as the experimental invariant sequence. Both have a period of 2. (a) at time n; (b) at time $n + 1$.

7.4 Tile attractor

We have derived some trapping sets for the attractors in time varying inputs case. But many sets in the trapping sequence are large enough to contain more than one tile. According to some experiments, the attractor sometimes looks occupying a big area although it might not actually fill it fully. The example shown in Figure 7.1 is such an example. It requires more advanced technique to generate really tight positively invariant sets. Note that the DC-input based Lyapunov functions is sufficiently proven the stability but may not bound the attractor effectively. One may seek for Lyapunov functions that genuinely deal with the time-varying dynamics.

On the other hand, when a trapping set has measure less than 2, according to Proposition 4.1.3, we can conclude that the attractor sequence is consisted of a sequence of single tiles. This is indeed guaranteed when the input signal has small enough

amplitude. It concludes the existence of tile attractor in AC inputs case.

This page is left blank on purpose !

Chapter 8

Discussion and future research

In this thesis, we establish a framework based on Lyapunov functions to study the stability of $\Sigma\Delta$ modulation. Philosophically speaking, adopting a Lyapunov function is choosing some relation of order of the state space, and use it as a reference to measure the evolution of the mapping M. This order is artificial, but it permits the replacement of difficult recursive and discrete reasonings by pure function analysis. By choosing Lyapunov functions that are quadratic in u_1 and piecewise linear (affine) in u_2, we have actually reproduced the type of algebra that appeared in the previous research on the stability of $\Sigma\Delta$ modulation. But, we have either re-established previously derived results with a more efficient and tractable language, or, improved and generalized them within the same complexity of functions. With the same methodology, the next step will be to allow more freedom in the design of the Lyapunov functions, to obtain trapping sets that are closer to the actual attractor of the transformation M. In particular, one can include into $h(\mathbf{u})$ more knowledge about the dynamics close to the attractor. A goal will not just be to obtain better bounds, but also to understand more theoretically the phenomenon of attraction. Another direction of research is to keep using simple analytical Lyapunov functions for the mere prediction of stability in more difficult systems, such as higher order systems.

Another contribution in this thesis is that we show a fundamental dynamical theorem in $\Sigma\Delta$ modulation by studying the the dynamics within a positively invariant set. This theorem enables us to construct trapping sets within a given positively invariant set. With this technique we finally prove that the global attractor is a single tile. This method may be generalized to higher order systems too. On the other hand, we may still concentrate on second order $\Sigma\Delta$ modulation but try to explicitly derive the attractor with the fundamental dynamical knowledge.

In this thesis, we prove that the attractor of second order $\Sigma\Delta$ modulation is a tile for a major portion of the configuration space in DC inputs case. A proof that can cover all configurations is still unavailable.

Appendix

A Proofs for propositions of Chapter 1

A.1 Proof of Proposition 1.3.1

Since $|x| > \frac{1}{2}$, both x_0 and x_1 have same sign as x. Then with (1.7), $\mathrm{M}^n(u)$ is a strictly monotonic sequence and at each step it changes at least $\min(|x_1|, |x_0|) > 0$ in absolute value. So it is impossible to be stable. In other words, the stable region is an empty set.

A.2 Proof of Proposition 1.3.2

First of all, $a < -x < b$ is implied by $x \in (-\frac{1}{2}, \frac{1}{2})$ and $a \leq -\frac{1}{2} < \frac{1}{2} \leq b$. Then from (1.7),

$$\begin{aligned} \mathrm{M}([a,b)) &= \mathrm{M}_0([a,b) \cap \Omega_0) \cup \mathrm{M}_1([a,b) \cap \Omega_1) \\ &= ([a,-x) + x_0) \cup ([-x,b)) + x_1) \\ &= [a + x_0, \tfrac{1}{2}) \cup [-\tfrac{1}{2}, b + x_1). \end{aligned} \tag{1}$$

Use $a \leq -\frac{1}{2} < \frac{1}{2} \leq b$ again, $a + \frac{1}{2} \leq b - \frac{1}{2}$. It follows that $a + x_0 \leq b + x_1$ by adding x to both side. Therefore, the right hand side of (1) is actually one interval as shown in (1.9).

A.3 Proof of Proposition 1.3.3

Firstly, we assume $a \leq -\frac{1}{2} < \frac{1}{2} \leq b$, according to Proposition 1.3.2,

$$\mathrm{M}([a,b)) = [\min(a + x_0, -\tfrac{1}{2}), \max(b + x_1, \tfrac{1}{2})).$$

Condition $x \in (-\frac{1}{2}, \frac{1}{2})$ implies that $x_1 < 0 < x_0$. It follows from $a \leq -\frac{1}{2} < \frac{1}{2} \leq b$ that $\begin{cases} a < a + x_0 \\ a \leq -\frac{1}{2} \end{cases}$ and $\begin{cases} b > b + x_1 \\ b \geq \frac{1}{2} \end{cases}$. Therefore, $\mathrm{M}([a,b)) \subset [a,b)$.

On the other hand, let $\mathrm{M}([a,b)) \subset [a,b)$. But we assume $a > -\frac{1}{2}$. Then nonempty set $[-\frac{1}{2}, a)$ and $[a,b)$ are disjointed. Therefore, $[a,b)$ can not have any intersection with $\mathrm{M}^{-1}([-\frac{1}{2}, a))$, which is the previous image of $[-\frac{1}{2}, a)$. Since

$$\mathrm{M}^{-1}([-\tfrac{1}{2}, a)) = (\mathrm{M}_1^{-1}([-\tfrac{1}{2}, a)) \cap \Omega_1) \cup (\mathrm{M}_0^{-1}([-\tfrac{1}{2}, a)) \cap \Omega_0)$$

and

$$\mathrm{M}_1^{-1}([-\tfrac{1}{2}, a)) \cap \Omega_1 = [-\tfrac{1}{2} - x_1, a - x_1) \cap [-x, +\infty) = [-x, a - x_1),$$

which leads to

$$[a,b) \cap [-x, a - x_1) = \emptyset.$$

This implies that $b \leq -x$ or $a \geq a - x_1$. However, we know from $x_1 < 0$ that $a < a - x_1$, which forces $b \leq -x$. Therefore, $[a,b) \subset \Omega_0$. This follows that $\mathrm{M}([a,b)) = \mathrm{M}_0([a,b)) = [a,b) + x_0$, which is impossible to be a subset of $[a,b)$ since $x_0 \neq 0$. To avoid this contradiction $a \leq -\frac{1}{2}$ must be true. Similarly, we can prove that $b \geq \frac{1}{2}$.

A.4 Proof of Proposition 1.3.4

Define

$$[a_n, b_n) := \mathrm{M}^n\left([a,b)\right). \tag{2}$$

Recursively using Proposition 1.3.3, we have that

$$\begin{cases} a_n \leq -\frac{1}{2} \\ b_n \geq \frac{1}{2} \end{cases}.$$

According to Proposition 1.3.2:

$$a_n = \min(a_{n-1} + x_0, -\tfrac{1}{2}) = \min(a_0 + nx_0, -\tfrac{1}{2}),$$

$$b_n = \max(b_{n-1} + x_1, \tfrac{1}{2}) = \max(b_0 + nx_1, \tfrac{1}{2}).$$

It is clear that the iteration will end up with $[a_n, b_n) = [-\frac{1}{2}, \frac{1}{2})$ within finite steps on n.

B Proofs for propositions of Chapter 2

B.1 Proof of Proposition 2.3.1

From (2.14),

$$\mathrm{M}_i(\mathbf{u} + \mathbf{v}) = \mathrm{M}_i\mathbf{u} + \mathbf{L}\mathbf{v}.$$

This becomes (2.18) when letting $\mathbf{v} := d\mathbf{j}$ and the fact that $\mathbf{L}\mathbf{j} = \mathbf{j}$. Since M_i is invertible as shown in (2.17), equation (2.19) is implied by (2.18), (2.17) and the fact $\mathbf{L}^{-1}\mathbf{j} = \mathbf{j}$.

B.2 Proof of Proposition 2.4.3

When $x > \frac{1}{2}$, according to (2.14), both x_0 and x_1 are positive and $x_0 > x_1$. For any give point $\mathbf{u}$, by applying (2.21), we have that $(\mathrm{M}^n\mathbf{u})_1 \geq u_1 + n\,x_1$, which is apparently unbounded when n tends to infinity. A similar result is obtained when $x < -\frac{1}{2}$. So the statable region is an empty set when $|x| > \frac{1}{2}$.

When $x = \frac{1}{2}$, according to (2.14), $x_0 = 1$ and $x_1 = 0$. For any given $\mathbf{u}$, by applying (2.21), we have that $(\mathrm{M}^n\mathbf{u})_1 = u_1 + k$, where k is a number to count how many times that $\mathrm{M}^{n'}\mathbf{u}$ belongs to Ω_0, where $n' \leq n$. Assume that $\mathrm{M}^n\mathbf{u}$ is bounded when n tends to infinity, k must be finite. So there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $(\mathrm{M}^n\mathbf{u})_1 = u_1 + k_m$ is constant and $\mathrm{M}^n\mathbf{u} \in \Omega_1$, where k_m is the maximum value of k. Then we apply (2.21) again and obtain that for all $n \in \mathbb{N}$, $(\mathrm{M}^{n_0+n}\mathbf{u})_2 = (\mathrm{M}^{n_0}\mathbf{u})_2 + n(u_1 + k_m)$. When $\mathrm{M}^n\mathbf{u}$ is bounded for all n, $u_1 + k_m$ must be 0. Since $(\mathrm{M}^{n+1}\mathbf{u} - \mathrm{M}^n\mathbf{u})_1 \in \{0, 1\}$,

we conclude that $u_1 \in \mathbb{Z}$, which force $\mathbf{u}$ can only taken from a measure zero set. A similar conclusion is obtained when $x = -\frac{1}{2}$. The proposition is finally proven.

C Proofs for propositions of Chapter 3

C.1 Proof of Proposition 3.2.2

Assume that $\mathrm{M}(\Upsilon_h) \subset \Lambda_h(\ell)$. For any $\mathbf{u} \in \Upsilon_h$, $\mathrm{M}\mathbf{u} \in \Lambda_h(\ell)$. Then $h(\mathbf{u}) < h(\mathrm{M}\mathbf{u}) \leq \ell$. This proves (i). Consider now any $\mathbf{u} \in \Lambda_h(\ell)$. If $\mathbf{u} \in \Upsilon_h$, we already know that $\mathrm{M}\mathbf{u} \in \Lambda_h(\ell)$. If $\mathbf{u} \notin \Upsilon_h$, then $h(\mathrm{M}\mathbf{u}) \leq h(\mathbf{u}) \leq \ell$, which implies that $\mathrm{M}\mathbf{u} \in \Lambda_h(\ell)$. So $\mathrm{M}(\Lambda_h(\ell)) \subset \Lambda_h(\ell)$.

C.2 Proof of Proposition 3.2.4

Let us call h_{min} a lower bound of $h(\mathbf{u})$. For any $\mathbf{u} \notin \Upsilon_h(\varepsilon)$, $\Delta h(\mathbf{u}) \leq -\varepsilon$. Now, for any $k \geq 0$, $h(\mathrm{M}^k\mathbf{u}) = h(\mathbf{u}) + \sum_{j=0}^{k-1} \Delta h(\mathrm{M}^j\mathbf{u})$. Consider $\mathbf{u}$ such that $\mathrm{M}^k\mathbf{u} \notin \Upsilon_h(\varepsilon)$ for all $k \geq 0$. Then $h_{min} \leq h(\mathrm{M}^k\mathbf{u}) \leq h(\mathbf{u}) - k\varepsilon$ for every $k \geq 0$, which is impossible.

C.3 Proof of Proposition 3.2.6

From (2.14), $h^i(\mathrm{M}_i\mathbf{u}) = \alpha_i\big((u_2+u_1)-g_i(u_1+x_i)\big)$. Then, $\Delta_i h^i(\mathbf{u}) = h^i(\mathrm{M}_i\mathbf{u})-h^i(\mathbf{u}) = \alpha_i\big(u_1 - g_i(u_1 + x_i) + g_i(u_1)\big)$. So $\Delta_i h^i(\mathbf{u}) = -\varepsilon$ if and only if $g_i(u_1)$ is a solution to the equation

$$g_i(u_1 + x_i) = g_i(u_1) + u_1 + \tfrac{\varepsilon}{\alpha_i}. \tag{3}$$

Define $\beta_i(u_1) := g_i(u_1) - f_i(u_1 + \frac{\varepsilon}{\alpha_i})$. Since $f_i(u_1)$ satisfies (3.14), $f_i(u_1 + x_i + \frac{\varepsilon}{\alpha_i}) - f_i(u_1 + \frac{\varepsilon}{\alpha_i}) = u_1 + \frac{\varepsilon}{\alpha_i}$. Then $\beta_i(u_1 + x_i) - \beta_i(u_1) = g_i(u_1 + x_i) - g_i(u_1) - (u_1 + \frac{\varepsilon}{\alpha_i})$. So $g_i(u_1)$ satisfies (3) if and only if $\beta_i(u_1 + x_i) = \beta_i(u_1)$.

C.4 Proof of Proposition 3.2.7

Assume that $\mathrm{M}(\Upsilon_h) \subset \Lambda_h(\ell)$. Then, for any $\mathbf{u} \in \Upsilon_h$, $h(\mathbf{u}) < h(\mathrm{M}\mathbf{u}) \leq \ell$. So $\Upsilon_h \subset \Lambda_h(\ell)$. Conversely, assume that $\Upsilon_h \subset \Lambda_h(\ell)$. Consider $\mathbf{u}_0 \in \Upsilon_h$. Since $t(\mathbf{u})$ is affine, there exists $\mathbf{v}_0$ such that $t(\mathbf{u}_\alpha) = t(\mathbf{u}_0)$ for all $\alpha \in \mathbb{R}$, with $\mathbf{u}_\alpha := \mathbf{u}_0 + \alpha\mathbf{v}_0$. Let i be the index of $\{0,1\}$ such that $\mathbf{u} \in \Omega_i$. Then $\mathbf{u}_\alpha \in \Omega_i$ for all $\alpha \in \mathbb{R}$. This implies in particular that $\Delta h(\mathbf{u}_\alpha) = \Delta_i h(\mathbf{u}_\alpha)$ for all $\alpha \in \mathbb{R}$. Since M_i is continuous and $h(\mathbf{u})$ is convex and thus continuous, $\Delta h(\mathbf{u}_\alpha)$ is a continuous function of α. So there exists a largest interval $I = (\alpha_0, \alpha_1)$ that contains 0 and such that $\mathbf{u}_\alpha \in \Upsilon_h$ for all $\alpha \in I$. By necessity, $\mathbf{u}_{\alpha_j} \in \overline{\Upsilon_h} \in \Lambda_\ell$, which implies that $h(\mathrm{M}_i\mathbf{u}_{\alpha_j}) = h(\mathrm{M}\mathbf{u}_{\alpha_j}) \leq \ell$, for both $j = 0, 1$. By taking $\theta := \frac{\alpha_1}{\alpha_1 - \alpha_0} \in [0,1]$, we have $\mathbf{u}_0 = \theta\mathbf{u}_{\alpha_0} + (1-\theta)\mathbf{u}_{\alpha_1}$. By linearity of M_i and convexity of h, $h(\mathrm{M}\mathbf{u}) = h(\mathrm{M}_i\mathbf{u}) = h\big(\theta\mathrm{M}_i\mathbf{u}_{\alpha_0} + (1-\theta)\mathrm{M}_i\mathbf{u}_{\alpha_1}\big) \leq \theta h(\mathrm{M}_i\mathbf{u}_{\alpha_0}) + (1-\theta)h(\mathrm{M}_i\mathbf{u}_{\alpha_1}) = \theta h(\mathbf{u}_{\alpha_0}) + (1-\theta)h(\mathbf{u}_{\alpha_1}) \leq \ell$. We have thus proved that $\mathrm{M}(\Upsilon_h) \subset \Lambda_h(\ell)$. The equality (3.22) is a trivial consequence.

C.5 Proof of Proposition 3.2.8

Using (3.19), (3.20) and (3.7),

$$
\begin{aligned}
&\Delta_0 h(\mathbf{u}) - \Delta_0 h^0(\mathbf{u}) \\
&= \big(h(\mathrm{M}_0\mathbf{u}) - h(\mathbf{u})\big) - \big(h^0(\mathrm{M}_0\mathbf{u}) - h^0(\mathbf{u})\big) = \big(h(\mathrm{M}_0\mathbf{u}) - h^0(\mathrm{M}_0\mathbf{u})\big) - \big(h(\mathbf{u}) - h^0(\mathbf{u})\big) \\
&= \big(\max(h^0(\mathrm{M}_0\mathbf{u}), h^1(\mathrm{M}_0\mathbf{u})) - h^0(\mathrm{M}_0\mathbf{u})\big) - \big(\max(h^0(\mathbf{u}), h^1(\mathbf{u})) - h^0(\mathbf{u})\big) \\
&= \max\big(0, Dh(\mathrm{M}_0\mathbf{u})\big) - \max\big(0, Dh(\mathbf{u})\big).
\end{aligned}
$$

With the inequality $0 \leq \max\big(0, Dh(\mathbf{u})\big)$, we then have

$$\begin{aligned}
&\Delta_0 h(\mathbf{u}) > \Delta_0 h^0(\mathbf{u}) \\
&\Leftrightarrow \ \max\big(0, Dh(\mathrm{M}_0\mathbf{u})\big) > \max\big(0, Dh(\mathbf{u})\big) \ \Leftrightarrow \ Dh(\mathrm{M}_0\mathbf{u}) > \max\big(0, Dh(\mathbf{u})\big) \\
&\Leftrightarrow \ \mathrm{M}_0 Dh(\mathbf{u}) = Dh(\mathrm{M}_0\mathbf{u}) > 0 \text{ and } \Delta_0 Dh(\mathbf{u}) = Dh(\mathrm{M}_0\mathbf{u}) - Dh(\mathbf{u}) > 0.
\end{aligned}$$

This proves (3.24). The proof of (3.25) is similar.

C.6 Proof of Proposition 3.2.9

For any $\mathbf{u} \in \Omega_0 = S_t^-$, $\Delta h(\mathbf{u}) = \Delta_0 h(\mathbf{u})$ from (3.23) and $\Delta_0 h^0(\mathbf{u}) = -\varepsilon$ by assumption. So using (3.24) and the notation of (3.26) and (3.27), we have

$$\forall \mathbf{u} \in \Omega_0, \qquad \Delta h(\mathbf{u}) > -\varepsilon \quad \Leftrightarrow \quad \Delta_0 h(\mathbf{u}) > \Delta_0 h^0(\mathbf{u}) \quad \Leftrightarrow \quad \mathbf{u} \in S^+_{\mathrm{M}_0 Dh} \cap S^+_{\Delta_0 Dh}.$$

Similarly, when $\mathbf{u} \in \Omega_1 = \overline{S_t^+}$, $\Delta h(\mathbf{u}) > -\varepsilon$ if and only if $\mathbf{u} \in S^-_{\mathrm{M}_1 Dh} \cap S^-_{\Delta_1 Dh}$. This proves (3.28).

C.7 Proof of Proposition 3.3.1

Consider $\mathbf{u} \in \mathbb{R}^2$. From (3.33), $Dh_{\delta,\varepsilon}(\mathbf{u} - d\mathbf{j}) = Dh_{\delta,\varepsilon}(\mathbf{u}) - d$ for any $d \in \mathbb{R}$. Let us choose $d = Dh_{\delta,\varepsilon}(\mathbf{u})$. Then $Dh_{\delta,\varepsilon}(\mathbf{u} - d\mathbf{j}) = 0$ and $h^0_{\delta,\varepsilon}(\mathbf{u} - d\mathbf{j}) = h^1_{\delta,\varepsilon}(\mathbf{u} - d\mathbf{j})$. Let i be the index of $\{0, 1\}$ such that $\mathbf{u} \in H_i$. Then, by using (3.31), we have

$$h_{\delta,\varepsilon}(\mathbf{u}) - h_{\delta,\varepsilon}(\mathbf{u} - d\mathbf{j}) = h^i_{\delta,\varepsilon}(\mathbf{u}) - h^i_{\delta,\varepsilon}(\mathbf{u} - d\mathbf{j}) = -x_i\, d = -x_i\, Dh_{\delta,\varepsilon}(\mathbf{u}).$$

In fact, $i = 1$ if and only if $Dh_{\delta,\varepsilon}(\mathbf{u}) \geq 0$. Since $x_1 < 0 < x_0$, it is then easy to see that $-x_i Dh_{\delta,\varepsilon}(\mathbf{u}) = a\big(Dh_{\delta,\varepsilon}(\mathbf{u})\big)$. Next, for any $\mathbf{v}$ such that $Dh_{\delta,\varepsilon}(\mathbf{v}) = 0$, it is easy to derive that $h_{\delta,\varepsilon}(\mathbf{v}) = h^{\mathrm{q}}_{\varepsilon}(\mathbf{v})$. Then, $h_{\delta,\varepsilon}(\mathbf{u} - d\mathbf{j}) = h^{\mathrm{q}}_{\varepsilon}(\mathbf{u} - d\mathbf{j}) = h^{\mathrm{q}}_{\varepsilon}(\mathbf{u})$ since $h^{\mathrm{q}}_{\varepsilon}(\mathbf{u})$ does not depend on u_2.

C.8 Proof of Proposition 3.3.2 and 3.3.3

These propositions are based on elementary properties of linear algebra. For the reader's convenience, we first review the needed properties.

Proposition H.1 *Let* $\mathbf{v}_1$, $\mathbf{v}_2$ *and* $\mathbf{v}_3$ *be three vectors of* $\mathbb{R}^2$, *and define* $d_{ij} := \det(\mathbf{v}_i, \mathbf{v}_j)$. *Then, the vector* $\mathbf{v}_0 := d_{23}\,\mathbf{v}_1 + d_{31}\,\mathbf{v}_2 + d_{12}\,\mathbf{v}_3$ *is zero.*

Proof: If, $d_{12} = d_{23} = d_{31} = 0$, clearly $\mathbf{v}_0 = \mathbf{0}$. If not, we can assume without loss of generality that $d_{12} \neq 0$. Then, the equation $\alpha\mathbf{v}_1 + \beta\mathbf{v}_2 = \mathbf{v}_3$ yields the solutions $\alpha = \frac{d_{32}}{d_{12}}$ and $\beta = \frac{d_{13}}{d_{12}}$. This implies that $\mathbf{v}_0 = \mathbf{0}$. ■

Next consider three affine functions

$$f_i(\mathbf{u}) = \langle \mathbf{u}, \mathbf{v}_i \rangle + c_i \tag{4}$$

where $\langle \mathbf{u}, \mathbf{v} \rangle := u_1 v_1 + u_2 v_2$ and $c_i \in \mathbb{R}$ for $i = 1, 2, 3$. As a trivial consequence of Proposition H.1,

$$\forall \mathbf{u} \in \mathbb{R}^2, \qquad d_{23}\, f_1(\mathbf{u}) + d_{31}\, f_2(\mathbf{u}) + d_{12}\, f_3(\mathbf{u}) = c_0 \tag{5}$$

where $c_0 := d_{23}\, c_1 + d_{31}\, c_2 + d_{12}\, c_3$. Define the set

$$\triangle := S^-_{f_1} \cap S^-_{f_2} \cap S^-_{f_3}. \tag{6}$$

We assume from now on that d_{12}, d_{23} and d_{31} are nonzero.

Proposition H.2 *The set* $\triangle$ *is bounded* if and only if d_{12}, d_{23} *and* d_{31} *have the same sign.*

Proof: From (5), we have

$$f_3(\mathbf{u}) = \tfrac{c_0}{d_{12}} - \tfrac{d_{23}}{d_{12}}\, f_1(\mathbf{u}) - \tfrac{d_{31}}{d_{12}}\, f_2(\mathbf{u}). \tag{7}$$

Assume that d_{12}, d_{23} and d_{31} have the same sign. We have in particular, $\frac{d_{23}}{d_{12}} > 0$ and $\frac{d_{31}}{d_{12}} > 0$. For any $\mathbf{u} \in \triangle$, (7) implies that $\frac{c_0}{d_{12}} < f_3(\mathbf{u}) < 0$. It can be shown similarly that $f_1(\mathbf{u})$ and $f_2(\mathbf{u})$ must be both lower bounded and upper bounded. Since for example, $\mathbf{v}_1$ and $\mathbf{v}_2$ are linearly independent, this is sufficient to bound any vector $\mathbf{u}$ of $\triangle$. Assume now that d_{12} has an opposite sign to both d_{23} and d_{31}, implying that $-\frac{d_{23}}{d_{12}} > 0$ and $-\frac{d_{31}}{d_{12}} > 0$. Consider the set $S := \big\{\mathbf{u} \in \mathbb{R}^2 : f_1(\mathbf{u}) < \min(0, \frac{c_0}{d_{23}})$ and $f_2(\mathbf{u}) < 0\big\}$. Because $(\mathbf{v}_1, \mathbf{v}_2)$ are linearly independent, S is easily seen to be unbounded. One also easily finds from (7) that any $\mathbf{u} \in S$ satisfies $f_3(\mathbf{u}) < 0$. So $S \subset \triangle$. This makes $\triangle$ unbounded. By similar proof, the same result is obtained in the two other cases where d_{12}, d_{23} and d_{31} are not of the same sign. ■

For any permutation (i, j, k) of $(1, 2, 3)$, call $\mathbf{p}_i$ the unique point of the intersection $O_{f_j} \cap O_{f_k}$. In other words, $\mathbf{p}_i$ satisfies

$$f_j(\mathbf{p}_i) = f_k(\mathbf{p}_i) = 0. \tag{8}$$

Proposition H.3 *Assume that d_{12}, d_{23} and d_{31} have the same sign. Then, the following propositions are equivalent:*

(i) $\triangle \neq \emptyset$,

(ii) $\exists i \in \{1, 2, 3\}$, $f_i(\mathbf{p}_i) < 0$,

(iii) $\forall i \in \{1, 2, 3\}$, $f_i(\mathbf{p}_i) < 0$,

(iv) $\overline{\triangle} = \triangle \mathbf{p}_1\mathbf{p}_2\mathbf{p}_3$.

Proof: Using (5) and (8), we have

$$\forall \mathbf{u} \in \mathbb{R}^2, \qquad d_{23}\, f_1(\mathbf{u}) + d_{31}\, f_2(\mathbf{u}) + d_{12}\, f_3(\mathbf{u}) = d_{23}\, f_1(\mathbf{p}_1) = d_{31}\, f_2(\mathbf{p}_2) = d_{12}\, f_3(\mathbf{p}_3) = c_0. \tag{9}$$

(i) $\Rightarrow$ (ii): Assuming (i) implies that there exists $\mathbf{u}$ such that $f_i(\mathbf{u}) < 0$ for all $i = 1, 2, 3$. Since d_{12}, d_{23} and d_{31} are nonzero with the same sign, it is clear from (9) that $f_1(\mathbf{p}_1) < 0$, which implies (ii).

(ii) $\Rightarrow$ (iii): This implication is again clear from (9) since d_{12}, d_{23} and d_{31} are nonzero with the same sign.

(iii) $\Rightarrow$ (iv): Assume (iii). For any permutation (i, j, k) of $(1, 2, 3)$, $\mathbf{p}_i$ does not belong to the straight line O_{f_i} since $f_i(\mathbf{p}_i) \neq 0$, while $\mathbf{p}_j, \mathbf{p}_k \in O_{f_i}$ due to (8). This implies that $\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p}_3$ are not aligned. Then, for any $\mathbf{u} \in \mathbb{R}^2$, there exists a unique triplet $(\theta_1, \theta_2, \theta_3)$ such that $\mathbf{u} = \sum_{i=1}^3 \theta_i \mathbf{p}_i$ with $\sum_{i=1}^3 \theta_i = 1$. Consider any $j \in \{1, 2, 3\}$. Since $f_j(\mathbf{u})$ is affine, $f_j(\mathbf{u}) = \sum_{i=1}^3 \theta_i\, f_j(\mathbf{p}_i)$. Due to (8), then $f_j(\mathbf{u}) = \theta_j\, f_j(\mathbf{p}_j)$. Since by assumption $f_j(\mathbf{p}_j) < 0$, $f_j(\mathbf{u}) < 0$ if and only if $\theta_j > 0$. This proves that $\triangle$ is equal to the interior of $\triangle \mathbf{p}_1\mathbf{p}_2\mathbf{p}_3$. This implies (iv).

(iv) $\Rightarrow$ (i): If $\triangle = \emptyset$, then $\overline{\triangle} = \emptyset$, which is impossible since $\triangle \mathbf{p}_1\mathbf{p}_2\mathbf{p}_3$ is never empty.

■

Let us now prove Proposition 3.3.2 and 3.3.3. The set Γ_f^0 of (3.30) is equal to the set $\triangle$ of (6) when taking

$$f_1(\mathbf{u}) := -\Delta_0 f(\mathbf{u}), \qquad f_2(\mathbf{u}) := -\mathrm{M}_0 f(\mathbf{u}) \qquad \text{and} \qquad f_3(\mathbf{u}) := t(\mathbf{u}). \tag{10}$$

By looking at at (3.39), (3.40) and (2.8), these function are of the form (4) with

$$\begin{array}{lll} \mathbf{v}_1 := -(1, 0), & \mathbf{v}_2 := -(a+1, 1), & \mathbf{v}_3 := (s, 1), \\ c_1 := -a x_0, & c_2 := -(b + a x_0) & c_3 := 0. \end{array}$$

With this, $d_{12} = 1$, $d_{23} = s - (a+1)$ and $d_{31} = 1$. Assuming that $s \neq a+1$, Proposition H.2 then implies that Γ_f^0 is bounded if and only if $s > a+1$. Next, the points $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$ coincide with $\mathbf{a}^0, \mathbf{b}^0, \mathbf{c}^0$, respectively. Then, under the condition $s > a+1$, Proposition H.3 is applicable and implies (3.42). The set Γ_f^1 of (3.30) is treated in a similar manner, with just a minor difference. By taking the functions $f_i(\mathbf{u})$ of (4) with

$$\begin{array}{lll} \mathbf{v}_1 := (1,0), & \mathbf{v}_2 := (a+1,1), & \mathbf{v}_3 := -(s,1), \\ c_1 := ax_1, & c_2 := b + ax_1 & c_3 := 0, \end{array}$$

we obtain with (6) a set $\triangle$ that is the interior of Γ_f^1 due to the closure performed on S_t^+ in (3.30). This however does not prevent the same results. Assuming that $s \neq a+1$, Γ_f^1 is bounded if and only if $s > a+1$ according to Proposition H.2. Then, assuming $s > a+1$, Proposition H.3 implies (3.43).

While the proof of Proposition 3.3.3 has been completed, we have only proved Proposition 3.3.2 assuming that $s \neq a+1$. In the special case $s = a+1$, one can prove that Γ_f^i is bounded if and only if it is empty, for each $i = 0, 1$ (for the sake of briefness, we will not show the details here). Meanwhile, one can also show that Γ_f^0 and Γ_f^1 cannot be simultaneously empty due to the assmption $a > 0$. This concludes the proof of Proposition 3.3.2.

C.9 Proof of Proposition 3.3.4

Suppose that Γ_f^0 or Γ_f^1 is empty. Then, Γ_f is either empty or reduced to one triangular set. In either case, it is obviously connected. Assume now that both Γ_f^0 and Γ_f^1 are non-empty. Define the point $\mathbf{g} := O_f \cap O_t$ (see Figure 3.2(b)). We have $\mathbf{a}^0, \mathbf{b}^0, \mathbf{g} \in O_t$ with $f(\mathbf{b}^0) < 0$, $f(\mathbf{g}) = 0$ and $f(\mathbf{a}^0) > 0$. Since $f(\mathbf{u})$ is affine, then $\mathbf{g}$ must belong to

the open segment $(\mathbf{b}^0, \mathbf{a}^0)$, i.e., there exists $\theta \in (0,1)$ such that $\mathbf{g} = (1-\theta)\mathbf{b}^0 + \theta\mathbf{a}^0$. Since $\mathrm{M}_0 f(\mathbf{u})$ and $\Delta_0 f(\mathbf{u})$ are also affine functions, we obtain with (3.46), $\mathrm{M}_0 f(\mathbf{u}_\theta) = (1-\theta)\mathrm{M}_0 f(\mathbf{b}^0) + \theta \mathrm{M}_0 f(\mathbf{a}^0) = \theta f(\mathbf{a}^0) > 0$ and $\Delta_0 f(\mathbf{u}_\theta) = (1-\theta)\Delta_0 f(\mathbf{b}^0) + \theta \Delta_0 f(\mathbf{a}^0) = -(1-\theta) f(\mathbf{b}^0) > 0$. This implies that $\mathbf{g} \in S^+_{\mathrm{M}_0 f} \cap S^+_{\Delta_0 f} \cap O_t \subset \overline{\Gamma^0_f}$. Similarly, one proves that $\mathbf{g} \in S^-_{\mathrm{M}_1 f} \cap S^-_{\Delta_1 f} \cap O_t \subset \Gamma^1_f$. Then $\mathbf{g} \in \overline{\Gamma^0_f} \cap \Gamma^1_f$. This is sufficient to make $\Gamma_f = \Gamma^0_f \cup \Gamma^1_f$ connected.

C.10 Proof of Proposition 3.3.5

The relations (3.45) can be rewritten in the form $\mathrm{M}_i f(\mathbf{u}) - \mathrm{M}_{\bar{i}} f(\mathbf{u}) = (-1)^i a = \Delta_i f(\mathbf{u}) - \Delta_{\bar{i}} f(\mathbf{u})$ for both $i = 0, 1$. Consider the mid-point $\mathbf{c} := \frac{1}{2}(\mathbf{c}^0 + \mathbf{c}^1)$. Since $\mathrm{M}_i f(\mathbf{u})$ is an affine function and $\mathrm{M}_i f(\mathbf{c}^i) = 0$ for both $i = 0, 1$, then $\mathrm{M}_i f(\mathbf{c}) = \frac{1}{2}(\mathrm{M}_i f(\mathbf{c}^0) + \mathrm{M}_i f(\mathbf{c}^1)) = \frac{1}{2}\mathrm{M}_i f(\mathbf{c}^{\bar{i}}) = (-1)^i \frac{a}{2}$. Similarly, $\Delta_i f(\mathbf{c}) = (-1)^i \frac{a}{2}$. Because $a > 0$, then $\mathbf{c} \in S^+_{\mathrm{M}_0 f} \cap S^+_{\Delta_0 f} \cap S^-_{\mathrm{M}_1 f} \cap S^-_{\Delta_1 f}$. If $t(\mathbf{c}) < 0$, $\mathbf{c} \in \Gamma^0_f$, otherwise $c \in \Gamma^1_f$. This proves that Γ_f is non-empty.

Assume that $\Gamma^0_f = \emptyset$. Either because the definitions of (3.41) or the equivalences of (3.42), $\mathrm{M}_0 f(\mathbf{u}) \leq 0$, $\Delta_0 f(\mathbf{u}) \leq 0$ and $t(\mathbf{u}) \geq 0$ for all $\mathbf{u} \in \{\mathbf{a}^0, \mathbf{b}^0, \mathbf{c}^0\}$. Since $\mathrm{M}_1 f(\mathbf{u}) < \mathrm{M}_0 f(\mathbf{u})$ and $\Delta_1 f(\mathbf{u}) < \Delta_0 f(\mathbf{u})$ from (3.45) and the assumption $a > 0$, then we also have $\mathrm{M}_1 f(\mathbf{u}) \leq 0$ and $\Delta_1 f(\mathbf{u}) \leq 0$ for all $\mathbf{u} \in \{\mathbf{a}^0, \mathbf{b}^0, \mathbf{c}^0\}$. This implies that $\{\mathbf{a}^0, \mathbf{b}^0, \mathbf{c}^0\} \subset \overline{\Gamma^1_f} = \triangle \mathbf{a}^1 \mathbf{b}^1 \mathbf{c}^1$ since $\Gamma^1_f \neq \emptyset$. This proves that $\triangle \mathbf{a}^i \mathbf{b}^i \mathbf{c}^i \subset \triangle \mathbf{a}^{\bar{i}} \mathbf{b}^{\bar{i}} \mathbf{c}^{\bar{i}}$ in the case $i = 0$. The case $i = 1$ is proven in a similar manner.

C.11 Proof of Proposition 3.3.6

According to (3.55), $\mathbf{a}^i$ depends on ε, while $\mathbf{b}^i$ and $\mathbf{c}^i$ depend on both δ and ε. To emphasize this dependence, let us write $\mathbf{a}^i_\varepsilon$, $\mathbf{b}^i_{\delta,\varepsilon}$ and $\mathbf{c}^i_{\delta,\varepsilon}$. The continuous dependence

of these points with ε in the neighborhood of 0 is clear since $d_\varepsilon > 0$ for all $\varepsilon \in [0, \varepsilon_s)$. Moreover, the differences $\mathbf{a}^i_\varepsilon - \mathbf{a}^i_0$ and $\mathbf{c}^i_{\delta,\varepsilon} - \mathbf{c}^i_{\delta,0} = \mathbf{a}^i_\varepsilon - \mathbf{a}^i_0 - (\delta^i_\varepsilon - \delta^i_0)\mathbf{j}$ do not depend on δ. Meanwhile, $\mathbf{b}^i_{\delta,\varepsilon} - \mathbf{b}^i_{\delta,0} = \mathbf{a}^i_\varepsilon - \mathbf{a}^i_0 - c_{\delta,\varepsilon}\,(1,-s)$, where $c_{\delta,\varepsilon} := \frac{1}{d_\varepsilon}\,(\delta^i_\varepsilon - \delta) - \frac{1}{d_0}\,(\delta^i_0 - \delta)$. One easily bounds $c_{\delta,\varepsilon}$ as $|c_{\delta,\varepsilon}| \leq \frac{1}{|d_\varepsilon|}\,|\delta^i_\varepsilon - \delta^i_0| + \left|\frac{1}{d_\varepsilon} - \frac{1}{d_0}\right|\,|\delta^i_0 - \delta|$. The right hand side goes to 0 uniformly with $\delta \in D$ when ε goes to 0.

C.12 Proof of Proposition 3.4.1

A point $\mathbf{u}$ belongs to Λ_δ if and only if $h^i_\delta(\mathbf{u}) \leq \ell$ for both $i \in \{0,1\}$. From (3.31), $h^i_\delta(\mathbf{u}) = x_i\big(p_i(u_1) - u_2 + \delta\big)$. Since $x_1 < 0 < x_0$, $\mathbf{u} \in \Lambda_\delta$ if and only if $p_0(u_1) + \delta - \frac{1}{x_0}\,\ell \leq u_2 \leq p_1(u_1) + \delta - \frac{1}{x_1}\,\ell$. The equivalence between (3.60) and (3.61) is obtained with (2.14).

C.13 Proof of Proposition 3.4.2

Suppose that $\Lambda_\delta(\ell) \supset \Upsilon_\delta$. From Proposition 3.2.7 we also have $\Lambda_\delta(\ell) \supset \mathrm{M}(\Upsilon_\delta)$. So $\Lambda_\delta(\ell)$ is positively invariant according to Proposition 3.2.2.

Conversely, assume that $\Lambda_\delta(\ell)$ is positively invariant. For all $\mathbf{u}$ such that $h_\delta(\mathbf{u}) = \ell$, we must have $\Delta h_\delta(u) \leq 0$. Since $\Upsilon_\delta = \Gamma_{Dh_\delta}$, it is connected according to Proposition 3.3.4. So we have either $\Upsilon_\delta \cap \Lambda_\delta(\ell) = \emptyset$ or $\Upsilon_\delta \subset \Lambda_\delta(\ell)$. Note from (3.35) that $h_\delta(\mathbf{u})$ has a unique global minimum $\mathbf{u}_\mathrm{m}$, which is the $\mathbf{u}$ point of O_{Dh} such that $u_1 = 0$. One one hand $\mathbf{u}_\mathrm{m} \in \Lambda_\delta(\ell)$. On the other hand, $\mathbf{u}_\mathrm{m} \in \Upsilon_\delta$ because the only way to obtain $\Delta h_\delta(\mathbf{u}_\mathrm{m}) \leq 0$ would be to have $\mathrm{M}\mathbf{u}_\mathrm{m} = \mathbf{u}_\mathrm{m}$, which is not possible. So $\Upsilon_\delta \subset \Lambda_\delta(\ell)$.

C.14 Proof of Proposition 3.4.3

In this proof, we will use the notation $\sup f(S) := \sup_{\mathbf{u}\in S} f(\mathbf{u})$ for any real function $f(\mathbf{u})$ and any set $S \subset \mathbb{R}^2$. Then (3.22) is rewritten as $\ell_\delta = \sup h_\delta(\Upsilon_{h_\delta})$. For any $\varepsilon \geq 0$, let us define $\ell_{\delta,\varepsilon} := \sup h_\delta(\Upsilon_{h_{\delta,\varepsilon}}(\varepsilon))$. Let us prove that $\ell_{\delta,\varepsilon}$ uniformly converges to $\ell_{\delta,0}$ when ε goes to 0 and $\delta \in D$.

From (3.28) and the new notation of (3.58), $\Upsilon_{h_{\delta,\varepsilon}}(\varepsilon) = \Gamma_{Dh_{\delta,\varepsilon}}$ and $\Upsilon_{h_\delta} = \Upsilon_{h_{\delta,0}}(0) = \Gamma_{Dh_{\delta,0}}$. With (3.32), we then also have $\ell_\delta = \sup h_0(\Gamma_{Dh_{\delta,0}} - \delta\mathbf{j})$ and $\ell_{\delta,\varepsilon} = \sup h_0(\Gamma_{Dh_{\delta,\varepsilon}} - \delta\mathbf{j})$. By Proposition 3.3.6, $\Gamma_{Dh_{\delta,\varepsilon}}$ is the union of two triangles whose vertices uniformly converge to those of $\Gamma_{Dh_{\delta,0}}$ when ε goes to 0 and $\delta \in D$. The same thing can be said of their shifted versions by $-\delta\mathbf{j}$. Now, because $h_0(\mathbf{u})$ is convex, the supremum of $h_0(\Gamma_{Dh_{\delta,\varepsilon}} - \delta\mathbf{j})$ is achieved in the set of its vertices. With $\delta \in D$ and ε limited to a neighborhood of 0 (for example $[0, \frac{\varepsilon_s}{2}]$ where ε_s is defined in Proposition 3.3.6), the set $\Gamma_{Dh_{\delta,\varepsilon}} - \delta\mathbf{j}$ can be easily shown to be enclosed in a bounded region $B \subset \mathbb{R}^2$. Because of the convexity of $h_0(\mathbf{u})$, there exists $\alpha \geq 0$ such that $|h_0(\mathbf{v}) - h_0(\mathbf{u})| \leq \alpha\|\mathbf{v} - \mathbf{u}\|$ for all $\mathbf{u}, \mathbf{v} \in B$ and some norm $\|\cdot\|$ of $\mathbb{R}^2$. This is sufficient to prove that $\ell_{\delta,\varepsilon}$ uniformly converges to ℓ_δ when ε goes to 0 and $\delta \in D$.

Consequently, for any $\lambda > 0$, there exists $\varepsilon > 0$ such that for any $\delta \in D$, $\ell_{\delta,\varepsilon} \leq \ell_\delta + \lambda$. For this $\varepsilon > 0$, we have $\Upsilon_{h_{\delta,\varepsilon}}(\varepsilon) \subset \Lambda_\delta(\ell_\delta + \lambda)$ for all $\delta \in D$.

C.15 Proof of Proposition 3.4.4

We first need to establish the following result.

Proposition O.4 *For any two pairs (δ, ℓ) and (δ', ℓ') of $\mathbb{R}^2$, there exists $(\delta'', \ell'') \in \mathbb{R}^2$ such that $\Lambda_\delta(\ell) \cap \Lambda_{\delta'}(\ell') = \Lambda_{\delta''}(\ell'')$. If moreover $\Lambda_\delta(\ell) \not\subseteq \Lambda_{\delta'}(\ell')$, then $\ell'' < \ell$.*

Proof: According to (3.60), $\Lambda_\delta(\ell) = \Pi_{c_0,c_1}$ and $\Lambda_{\delta'}(\ell') = \Pi_{c'_0,c'_1}$, where $c_i = \delta - \frac{1}{x_i}\,\ell$ and $c'_i = \delta' - \frac{1}{x_i}\,\ell'$. Then $\Lambda_\delta(\ell) \cap \Lambda_{\delta'}(\ell') = \Pi_{c''_0,c''_1}$ where $c''_0 = \max(c_0, c'_0)$ and $c''_1 = \min(c_1, c'_1)$. According to (3.61), $\Pi_{c''_0,c''_1} = \Lambda_{\delta''}(\ell'')$ where $\delta'' = x_0 c''_0 - x_1 c''_1$ and $\ell'' = -x_0 x_1 (c''_1 - c''_0)$. From (3.61), we have $c_1 - c_0 = -\frac{\ell}{x_0 x_1}$. Then,

$-\frac{\ell''}{x_0 x_1} = c''_1 - c''_0 = \min(c_1, c'_1) - \max(c_0, c'_0) \leq \min(c'_1 - c_0, c_1 - c'_0) = \min(c'_1 - c_1, c_0 - c'_0) - \frac{\ell}{x_0 x_1}$. Assume that $\Lambda_\delta(\ell) \not\subseteq \Lambda_{\delta'}(\ell')$. Then $\Pi_{c_0,c_1} \not\subseteq \Pi_{c'_0,c'_1}$, which implies that $c'_1 < c_1$ or $c'_0 > c_0$, and hence $\min(c'_1 - c_1, c_0 - c'_0) < 0$. Since $-x_0 x_1 > 0$, we obtain $\ell'' < \ell$. ■

We now prove Proposition 3.4.4. Consider a set $\Lambda_\delta(\ell)$ that is positively invariant. Assume that $\Lambda_{\delta^*}(\ell^*) \not\subset \Lambda_\delta(\ell)$. Then, according to Proposition O.4, there exists $(\delta'', \ell'') \in \mathbb{R}^2$ such that $\Lambda_{\delta''}(\ell'') = \Lambda_{\delta^*}(\ell^*) \cap \Lambda_\delta(\ell)$ with $\ell'' < \ell^*$. Since $\Upsilon_{\delta^*} \subset \Lambda_{\delta^*}(\ell^*)$ and $\Upsilon_\delta \subset \Lambda_\delta(\ell)$, we know from Proposition 3.3.5 that $\Lambda_{\delta''}(\ell'')$ contains $\mathbf{a}^0$ and $\mathbf{a}^1$. So $\Lambda_{\delta''}(\ell'')$ is then positively invariant as a non-empty intersection between two positively invariant sets. So $\ell'' \geq \ell_{\delta''}$ according to Proposition 3.2.7. but $\ell_{\delta''} \geq \ell_{\delta^*} = \ell^*$, which is contradictory. So $\Lambda_{\delta^*} = \Lambda_{\delta^*}(\ell^*)$ must be included in $\Lambda_\delta(\ell)$.

C.16 Proof of Proposition 3.64

Because $h_\delta(\mathbf{u})$ is a convex function, then $\ell^i_\delta = \max\big(h_\delta(\mathbf{a}^i), h_\delta(\mathbf{b}^i), h_\delta(\mathbf{c}^i)\big)$. According to (3.41), $\mathbf{a}^i$ and $\mathbf{c}^i$ belong to $R_{\Delta_i Dh}$ and thus have the same u_1 value due to (3.40). Since $h^{\mathrm{q}}_0(\mathbf{u})$ given in (3.37) depends only on u_1, then $h^{\mathrm{q}}_0(\mathbf{a}^i) = h^{\mathrm{q}}_0(\mathbf{c}^i)$. Next, from (3.44), $Dh_\delta(\mathbf{c}^i) = \mathrm{M}_i Dh_\delta(\mathbf{c}^i) - \Delta_i Dh_\delta(\mathbf{c}^i) = 0$ by definition of $\mathbf{c}^i$ in (3.41). Then, using (3.35) and (3.37), we have $h_\delta(\mathbf{a}^i) - h_\delta(\mathbf{c}^i) = v\big(Dh_\delta(\mathbf{a}^i)\big) - v\big(Dh_\delta(\mathbf{c}^i)\big) = v\big(Dh_\delta(\mathbf{a}^i)\big) \geq 0$. This reduces ℓ^i_δ to $\max\big(h_\delta(\mathbf{a}^i), h_\delta(\mathbf{b}^i)\big)$. Consider the case $i = 0$ and $\delta < \delta^0$. From (3.56) and (3.47), we have $\Gamma^0_{Dh_\delta} \neq \emptyset$ and $Dh_\delta(\mathbf{b}^0) < 0 < Dh_\delta(\mathbf{a}^0)$. This implies from

Table 1:

(i,j)	$\delta^* = \delta_{ij}$	$\ell^* = m_{ij}$	range in s
(b,b)	$(s-\frac{5}{4})x$	$\frac{1}{8}\big[\big(\frac{s-1}{s-\frac{3}{2}}\big)^2 + (\frac{1}{4}-x^2)\big]$	$\frac{3}{2} < s \le s_1(x)$
(b,a)	$\frac{1}{2}(s-1)x - \frac{1}{4}(s-\frac{5}{2})$	$\frac{1}{16}(5+6x)$	$s_1(x) \le s \le s_2(x)$
(a,a)	0	$\frac{1}{2}\big[(\frac{1}{4}-x^2)s + x^2\big]$	$s_2(x) \le s$
(a,b)	$\frac{1}{2}(s-1)x + \frac{1}{4}(s-\frac{5}{2})$	$\frac{1}{16}(5-6x)$	$\emptyset$

(3.20) and (3.19) that $h_\delta(\mathbf{a}^0) = h^1_\delta(\mathbf{a}^0)$ and $h_\delta(\mathbf{b}^0) = h^0_\delta(\mathbf{b}^0)$. This is also true at $\delta = \delta_0$ by continuity. The case $i = 1$ is proven in a similar manner.

C.17 Derivation of Table 3.1

We in this section derive a more complete Table 1 which includes Table 3.1. The meaning of (i,j), δ_{ij}, m_{ij} and of the last line of Table 1 is explained later.

The value $h^{\bar{i}}_\delta(\mathbf{a}^i)$ of (3.65) is a linear function of δ that is strictly decreasing with $i = 0$ and strictly increasing with $i = 1$ due to the inequalities $x_1 < 0 < x_0$. Meanwhile, $h^i_\delta(\mathbf{b}^i)$ from (3.65) is quadratic of positive curvature and minimum located at $\delta = \mu^i$. Since $x_1 < 0 < x_0$ and $s > \frac{3}{2}$, (3.66) implies that $\mu^1 < \delta^1$ and $\delta^0 < \mu^0$. Therefore, $h^0_\delta(\mathbf{b}^0)$ is strictly decreasing on $(-\infty, \delta^0]$ and $h^1_\delta(\mathbf{b}^1)$ is strictly decreasing on $[\delta^1, \infty)$. We conclude from (3.64) that ℓ^0_δ is strictly decreasing on $(-\infty, \delta^0]$ and ℓ^0_δ is strictly increasing on $[\delta^1, \infty)$. As a consequence, the minimum of ℓ_δ must be achieved at some $\delta^* \in [\delta^1, \delta^0]$. The following proposition gives special conditions under which δ^* is easy to find.

Proposition Q.5 *On an interval $[\alpha, \beta]$, let $\{\grave{g}_i(\delta)\}_{i\in I}$ be a finite family of strictly decreasing functions and $\{\acute{g}_j(\delta)\}_{j\in J}$ be a finite family of strictly increasing functions. For each $(i,j) \in I \times J$, we assume moreover that the equation $\grave{g}_i(\delta) = \acute{g}_j(\delta)$ has a*

solution $\delta_{ij} \in [\alpha, \beta]$ *and we define* $m_{ij} := \grave{g}_i(\delta_{ij}) = \acute{g}_j(\delta_{ij})$. *Then the function*

$$g(\delta) := \max\Big(\max_{i\in I} \grave{g}_i(\delta), \max_{j\in J} \acute{g}_j(\delta)\Big) \tag{11}$$

is minimized in $[\alpha, \beta]$ *at* $\delta = \delta_{i^*j^*}$ *of minimal value* $m_{i^*j^*}$, *where* (i^*, j^*) *is such that* $m_{i^*j^*} = \max_{(i,j)\in I\times J} m_{ij}$.

Proof: For any $\delta \in [\alpha, \delta_{i^*j^*}]$, $g(\delta) \geq \grave{g}_{i^*}(\delta) > \grave{g}_{i^*}(\delta_{i^*j^*}) = m_{i^*j^*}$ and for any $\delta \in [\delta_{i^*j^*}, \beta]$, $g(\delta) \geq \acute{g}_{j^*}(\delta) > \acute{g}_{j^*}(\delta_{i^*j^*}) = m_{i^*j^*}$. Then, $g(\delta) \geq m_{i^*j^*}$ for all $\delta \in [\alpha, \beta]$. Next, for any $i \in I$ and $j \in J$,

$$\begin{aligned}
m_{ij^*} \leq m_{i^*j^*} &\Rightarrow \acute{g}_{j^*}(\delta_{ij^*}) \leq \acute{g}_{j^*}(\delta_{i^*j^*}) \Rightarrow \delta_{ij^*} \leq \delta_{i^*j^*} \\
&\Rightarrow \grave{g}_i(\delta_{i^*j^*}) \leq \grave{g}_i(\delta_{ij^*}) = \acute{g}_{j^*}(\delta_{ij^*}) \leq \acute{g}_{j^*}(\delta_{i^*j^*}) = m_{i^*j^*}, \\
m_{i^*j} \leq m_{i^*j^*} &\Rightarrow \grave{g}_{i^*}(\delta_{i^*j}) \leq \grave{g}_j(\delta_{i^*j^*}) \Rightarrow \delta_{i^*j} \geq \delta_{i^*j^*} \\
&\Rightarrow \acute{g}_j(\delta_{i^*j^*}) \leq \acute{g}_j(\delta_{i^*j}) = \grave{g}_{i^*}(\delta_{i^*j}) \leq \grave{g}_{i^*}(\delta_{i^*j^*}) = m_{i^*j^*}.
\end{aligned}$$

Then, $g(\delta_{i^*j^*}) \leq m_{i^*j^*}$. But $m_{i^*j^*} = g_{i^*}(\delta_{i^*j^*}) \leq g(\delta_{i^*j^*})$. So $g(\delta_{i^*j^*}) = m_{i^*j^*}$. ■

The function ℓ_δ of (3.62) on (δ^1, δ^0) is of the form (11) if we take $(\alpha, \beta) = (\delta^1, \delta^0)$, $I = J := \{a, b\}$ (where a and b are abstract indices) and

$$\grave{g}_a(\delta) := h^1_\delta(\mathbf{a}^0), \quad \grave{g}_b(\delta) := h^0_\delta(\mathbf{b}^0),$$

$$\acute{g}_a(\delta) := h^0_\delta(\mathbf{a}^1), \quad \acute{g}_b(\delta) := h^1_\delta(\mathbf{b}^1).$$

We already know that $\grave{g}_i(\delta)$ is strictly decreasing and $\acute{g}_i(\delta)$ is strictly increasing for any $i, j \in I$. We solve the equation $\grave{g}_i(\delta) = \acute{g}_j(\delta)$ for each $(i, j) \in I^2$ and give in Table 1 the analytical expression of its solution δ_{ij} together with the value $m_{ij} := \grave{g}_i(\delta_{ij}) = \acute{g}_j(\delta_{ij})$. It can be checked that $\delta_{ij} \in [\delta^1, \delta^0]$ for all $i, j \in I$, given the constraint

$(x, s) \in [0, \frac{1}{2}) \times (\frac{3}{2}, \infty)$. So Proposition Q.5 is applicable. Then, for any given (x, s), the next step is to find the pair (i, j) of I^2 that is equal to $(\overset{*}{i}, \overset{*}{j})$. As $x \geq 0$, we always have $m_{ba} \geq m_{ab}$. So $(a, b) \neq (\overset{*}{i}, \overset{*}{j})$ in any case. Next, m_{bb} is easily shown to be strictly decreasing with s when $s > \frac{3}{2}$, while m_{aa} is obviously strictly increasing with s since $|x| < \frac{1}{2}$. For any given $x \in [0, \frac{1}{2})$, when solving the equations $m_{bb} = m_{ba}$ and $m_{aa} = m_{ba}$ in s with $s > \frac{3}{2}$, we find respectively $s = s_1(x)$ and $s = s_2(x)$ where $s_1(x)$ and $s_2(x)$ are given in (3.67). Clearly $s_1(x) \leq s_2(x)$. We conclude from these derivations the region of (x, s) where $(i, j) = (\overset{*}{i}, \overset{*}{j})$ for each (i, j) of I^2 and the result is reported in the last column of Table 1.

C.18 Proof of Proposition 3.4.6

First, we prove that S' is positively invariant. Take any $\mathbf{u} \in S' \cap \Omega_0$. Then $(\mathrm{M}\mathbf{u})_1 = (\mathrm{M}_0\mathbf{u})_1 = u_1 + x_0$ and $\gamma_1 + x_1 \leq u_1 \leq \gamma_0$. This follows from $x_0 > 0$ that $\gamma_1 + x_1 \leq (\mathrm{M}\mathbf{u})_1 \leq \gamma_0 + x_0$. Because S is trapping set, $\mathrm{M}\mathbf{u} \in S$. Therefore, $\mathrm{M}\mathbf{u} \in S'$, too. Similarly, we can prove the case of $\mathbf{u} \in S' \cap \Omega_1$.

We prove the follow inequality before proving points in $S \setminus S'$ will be trapped into S':

$$\gamma_0 - \gamma_1 \geq x_1. \tag{12}$$

Take a point $\mathbf{u} \in \{\mathbf{u} \in \Omega_1 \cap S : \quad \gamma_1 \leq u_1 < \gamma_1 - x_1\}$. Upon the existence of γ_1, $\mathbf{u}$ exists. Then $(\mathrm{M}\mathbf{u})_1 = (\mathrm{M}_1\mathbf{u})_1 = u_1 + x_1$, which means $\gamma_1 + x_1 \leq (\mathrm{M}\mathbf{u})_1 < \gamma_1$. This follows from the fact $\mathrm{M}\mathbf{u} \in S$ that $\mathrm{M}\mathbf{u} \in \Omega_0$. Therefore, $\gamma_1 + x_1 \leq (\mathrm{M}\mathbf{u})_1 \leq \gamma_0$, which leads to (12).

Let us define sets

$$S_1 := \{\mathbf{u} \in S : \quad u_1 > \gamma_0 + x_0\},$$

$$S_0 := \{\mathbf{u} \in S : \quad u_1 < \gamma_1 + x_1\}.$$

It is clear that $S_i \in \Omega_i \cap S$, for $i = 0, 1$. And $\{S', S_0, S_1\}$ forms a partition of S. Following (12), we have

$$\gamma_0 + x_0 - (\gamma_1 + x_1) = \gamma_0 - \gamma_1 + 1 \geq x_1 + 1 = x_0 \geq \max(|x_0|, |x_1|).$$

Therefore, $\mathrm{M}(S_1) \subset S_1 \cup S'$ and $\mathrm{M}(S_0) \subset S_0 \cup S'$.

If we assume that for all $n \in \mathbb{N}$, $\mathbf{u} \in S_1$. But $(\mathrm{M}^n\mathbf{u})_1 = u_1 + nx_1$ can not be lower bounded when n tends to infinity as keeping inside S_1. So $\mathbf{u}$ must be trapped into S'. Similarly, $\mathbf{u} \in S_0$ will be trapped into S', too.

D Proofs for propositions of Chapter 4

D.1 Proof of Proposition 4.3.1

The forward implications:

(i) If S is a tile, then $\bigcup_{\mathbf{k}\in\mathbb{Z}^2}(S+\mathbf{k}) = \mathbb{R}^2$ while S and $\bigcup_{\mathbf{k}\in\mathbb{Z}^2\setminus\{\mathbf{0}\}}(S+\mathbf{k})$ are disjointed. This implies that $\mathcal{T}(S) = S$.

(ii) If S contains a tile Γ, then $\mathcal{T}(S) \subset \mathcal{T}(\Gamma) = \Gamma \subset S$.

(iii) If S contains two disjointed tiles Γ_1 and Γ_2, then $\mathcal{T}(S) \subset \mathcal{T}(\Gamma_1 \cup \Gamma_2) = \mathcal{T}(\Gamma_1) \cap \mathcal{T}(\Gamma_2) = \Gamma_1 \cap \Gamma_2 = \emptyset$.

The backward implications:

(i) Assume that $\mathcal{T}(S) = S$. Then $\bigcup_{\mathbf{k}\in\mathbb{Z}^2}(S+\mathbf{k}) = \mathbb{R}^2$ while S and $\bigcup_{\mathbf{k}\in\mathbb{Z}^2\setminus\{\mathbf{0}\}}(S+\mathbf{k})$ are disjointed. Since S and $S+\mathbf{k}$ are disjointed for any $\mathbf{k} \in \mathbb{Z}^2 \setminus \{\mathbf{0}\}$, then $S+\mathbf{k}$ and

$S + \mathbf{k}'$ are disjointed for any distinct $\mathbf{k}, \mathbf{k}' \in \mathbb{Z}^2$. Thus, S is a tile.

(ii) Assume that $\mathcal{T}(s) \subset S$. This implies that $\bigcup_{\mathbf{k}\in\mathbb{Z}^2}(S + \mathbf{k}) = \mathbb{R}^2$. For every $\mathbf{u} \in [0, 1)^2$, one can choose one vector $\mathbf{k_u} \in \mathbb{Z}^2$ such that $\mathbf{u} \in S + \mathbf{k_u}$. The set $\Gamma := \left\{\mathbf{u} - \mathbf{k_u} : \mathbf{u} \in [0, 1)^2\right\}$ is obviously a subset of S. It is easy to show that Γ is also a tile.

(iii) Assume that $\mathcal{T}(s) = \emptyset$. This implies that $\bigcup_{\mathbf{k}\in\mathbb{Z}^2\setminus\{\mathbf{0}\}}(S + \mathbf{k}) = \mathbb{R}^2$. Consider a given $\mathbf{u} \in [0, 1)^2$. One can choose one vector $\mathbf{k_u} \in \mathbb{Z}^2 \setminus \{\mathbf{0}\}$ such that $\mathbf{u} \in S + \mathbf{k_u}$. Now, since $\mathbb{R}^2 + \mathbf{k_u} = \mathbb{R}^2$, we also have $\bigcup_{\mathbf{k}\in\mathbb{Z}^2\setminus\{\mathbf{k_u}\}}(S + \mathbf{k}) = \mathbb{R}^2$. We can choose $\mathbf{k}'_\mathbf{u} \in \mathbb{Z}^2 \setminus \{\mathbf{k_u}\}$ such that $\mathbf{u} \in S + \mathbf{k}'_\mathbf{u}$. One can easily show that the two sets $\Gamma := \left\{\mathbf{u} - \mathbf{k_u} : \mathbf{u} \in [0, 1)^2\right\}$ and $\Gamma' := \left\{\mathbf{u} - \mathbf{k}'_\mathbf{u} : \mathbf{u} \in [0, 1)^2\right\}$ are two disjointed tiles that are subsets of S.

D.2 Proof of Proposition 4.3.2

Consider a set S that contains two disjointed tiles up to a set of measure 0. This amounts to saying that with a zero-measure set S_0, which could be empty, $S \cup S_0$ contains two disjointed tiles. According to Proposition 4.3.1, $\mathcal{T}(S \cup S_0) = \emptyset$. From (4.3), we have $\mathcal{T}(S) \cap \mathcal{T}(S_0) = \emptyset$, which implies that $\mathcal{T}(S) \subset \overline{\mathcal{T}(S_0)} = \bigcup_{\mathbf{k}\in\mathbb{Z}^2\setminus\{\mathbf{0}\}}(S_0 + \mathbf{k})$. Since $m(S_0) = 0$, then $m(\mathcal{T}(S)) = 0$.

E Proofs for propositions of Chapter 5

E.1 Proof of Proposition 5.1.1

Recursively using (5.9), for all $n \in \mathbb{N}$,

$$d_1(\boldsymbol{w}[n]) = d_1(\boldsymbol{w}[0]) + \sum_{k=0}^{n-1} \mathbf{i}_k, \tag{13}$$

$$d_2(\boldsymbol{w}[n]) = d_2(\boldsymbol{w}[0]) + n\, d_1(\boldsymbol{w}[0]), \tag{14}$$

where

$$\forall k, \mathbf{i}_k \in \{-1, 0, 1\}.$$

Therefore, property (i) and (ii) are clearly true.

For (iii), if $\mathbf{d}(\boldsymbol{w}[0]) \notin \mathbb{Z}^2$, then $d_1(\boldsymbol{w}[0]) \notin \mathbb{Z}$ or $d_1(\boldsymbol{w}[0]) \in \mathbb{Z}$ but $d_2(\boldsymbol{w}[0]) \notin \mathbb{Z}$. In the first case, from 13, $d_1(\boldsymbol{w}[n]) \notin \mathbb{Z}$ for all $n \in \mathbb{N}$. In the second case, $n\, d_1(\boldsymbol{w}[0]) \in \mathbb{Z}$, then from 14, $d_2(\boldsymbol{w}[n]) \notin \mathbb{Z}$ for all $n \in \mathbb{N}$. Hence, (iii) is proven.

E.2 Proof of Proposition 5.2.1

By applying (5.9), for any $x \in \mathbb{R}$, $\mathbf{d}(\mathcal{M}_x \boldsymbol{w}) = \big(d_1(\boldsymbol{w}),\ d_1(\boldsymbol{w}){+}d_2(\boldsymbol{w})\big)$ since $k = 0$. We then obtain that $\mathbf{d}(\tilde{\mathcal{M}}_n \boldsymbol{w}) = \big(d_1(\boldsymbol{w}),\ d_2(\boldsymbol{w}){+}n\, d_1(\boldsymbol{w})\big)$ by repeating same argument. And $\tilde{\mathcal{M}}_n \boldsymbol{w}$ must remain in $S_n \times S_n$ for all $n \in \mathbb{N}$. Since $(S_n)_2$ is bounded, $n\, d_1(\boldsymbol{w})$ remains bounded of all n, which forces $d_1(\boldsymbol{w})$ to be 0. This finally implies that $\mathbf{d}(\tilde{\mathcal{M}}_n \boldsymbol{w}) = \mathbf{d}(\boldsymbol{w})$.

E.3 Proof of Proposition 5.2.2

We first introduce two propositions.

Proposition C.1 *For a given set S, assume that there exists a bounded set B such that for all $n \in \mathbb{N}$, for all $\mathbf{u}, \mathbf{v} \in S_n, \mathbf{u} - \mathbf{v} \in B$, then*

$$\mathbf{p}_{\mathrm{u}}(\Phi_S) = \bigcup_{k\in\mathbb{Z}\setminus\{0\}} \Phi_S^k. \tag{15}$$

where

$$\Phi_S^k := \big\{\mathbf{u} \in S : \mathbf{u} + k\mathbf{j} \in S \text{ and } \tilde{\mathcal{M}}_n(\mathbf{u}, \mathbf{u} + k\mathbf{j}) \in \Pi_0^{x[n]} \text{ for all } n \in \mathbb{N}\big\}. \tag{16}$$

Proof: According to (5.17) and Lemmas 5.2.1, for any $\boldsymbol{w} \in \Phi_S$, $d_1(\boldsymbol{w}) = 0$. Therefore, Φ_S can be rewritten as

$$\Phi_S = \big\{\boldsymbol{w} \in \mathcal{W}_S : \quad \mathbf{d}(\boldsymbol{w}) \in \{0\} \times \mathbb{Z} \setminus \{0\} \quad \text{and} \quad \tilde{\mathcal{M}}_n \boldsymbol{w} \in \Pi_0^{x[0]} \text{ for all } n \in \mathbb{N}\big\}.$$

So Φ_S yields the following direct characterization,

$$\mathbf{p}_{\mathrm{u}}(\Phi_S) = \big\{\mathbf{u} \in S : \exists k \in \mathbb{Z}\backslash\{0\}, \quad \mathbf{u}+k\mathbf{j} \in S \text{ and } \tilde{\mathcal{M}}_n(\mathbf{u}, \mathbf{u}+k\mathbf{j}) \in \Pi_0^{x[n]} \text{ for all } n \in \mathbb{N}\big\}.$$

This clearly implies (15).

■

Proposition C.2 *For any given $k \in \mathbb{Z}$, Φ_S^k satisfies the following properties:*

(i) $\tilde{\mathcal{M}}_n(\Phi_S^k) \subset \big(\Omega_0^{x[n]} \cap (\Omega_0^{x[n]} - k\mathbf{j})\big) \cup \big(\Omega_1^{x[n]} \cap (\Omega_1^{x[n]} - k\mathbf{j})\big)$ for all $n \in \mathbb{N}$,

(ii) if for all $n \in \mathbb{N}$, $\Omega_0^{x[n]}$, $\Omega_1^{x[n]}$ and S are measurable, Φ_S^k is measurable.

Proof:

(i) When $\mathbf{u} \in \Phi_S^k$, we from (16) have $\tilde{\mathcal{M}}_n(\mathbf{u}, \mathbf{u}+k\mathbf{j}) \in \Pi_0^{x[n]} = (\Omega_0^{x[n]})^2 \cup (\Omega_1^{x[n]})^2$. For any given $i \in \{0, 1\}$, $\tilde{\mathcal{M}}_n(\mathbf{u}, \mathbf{u}+k\mathbf{j}) \in (\Omega_i^{x[n]})^2$ if and only if $\mathbf{u} \in \Omega_i^{x[n]} \cap (\Omega_i^{x[n]} - k\mathbf{j})$. This completes the proof of (i).

(ii) Take any $\mathbf{u} \in \Phi_S^k$, then for all $n \in \mathbb{N}$, for all $i \in \{0,1\}$ such that

$$\begin{aligned} & \tilde{\mathcal{M}}_n(\mathbf{u}, \mathbf{u}+k\mathbf{j}) \in (\Omega_i^{x[n]})^2 \\ \Leftrightarrow \quad & \mathbf{u} \in \tilde{\mathcal{M}}_{-n}(\Omega_i^{x[n]}) \text{ and } \mathbf{u}+k\mathbf{j} \in \tilde{\mathcal{M}}_{-n}(\Omega_i^{x[n]}) \\ \Leftrightarrow \quad & \mathbf{u} \in \tilde{\mathcal{M}}_{-n}(\Omega_i^{x[n]}) \cap \big(\tilde{\mathcal{M}}_{-n}(\Omega_i^{x[n]}) - k\mathbf{j}\big). \end{aligned}$$

Since $\Pi_0^{x[n]} = (\Omega_0^{x[n]})^2 \cup (\Omega_1^{x[n]})^2$, we conclude from (16) that

$$\Phi_S^k = S \cap (S - k\mathbf{j}) \cap \bigcap_{n\in\mathbb{N}} \bigcup_{i=\{0,1\}} \big(\tilde{\mathcal{M}}_{-n}(\Omega_i^{x[n]}) \cap (\tilde{\mathcal{M}}_{-n}(\Omega_i^{x[n]}) - k\mathbf{j})\big). \qquad (17)$$

For any measurable set $B \subset \mathbb{R}^2$, $\mathcal{M}_x^{-1}(B) = \big(\mathrm{M}_0^{-1}(B) \cap \Omega_0\big) \cup \big(\mathrm{M}_1^{-1}(B) \cap \Omega_1\big)$, which is measurable. By induction $\tilde{\mathcal{M}}_{-n}(B)$ is then also measurable for all $n \in \mathbb{N}$. The expression (17) then shows that Φ_S^k is a countable union and intersection of measurable sets. So Φ_S^k is measurable.

■

According to Proposition C.1 and set theory, we only need to prove that for all k, $m(\Phi_S^k) = 0$. Consider any $n \in \mathbb{N}$, from property (i) of Proposition C.2, for any integer $k > 0$, $\tilde{\mathcal{M}}_n(\Phi_S^k) \subset (\Omega_0^{x[n]} - k\mathbf{j}) \cup \Omega_1^{x[n]}$. The two sets $(\Omega_0^{x[n]} - k\mathbf{j}) = \big\{(u_1, u_2) : u_2 < t_{x[n]}(u_1) - k\big\}$ and $\Omega_1^{x[n]} = \big\{(u_1, u_2) : u_2 \geq t_{x[n]}(u_1)\big\}$ are disconnected because $t_{x[n]}(u)$ is a continuous function. Therefore, any connected subset S_c of $\tilde{\mathcal{M}}_n(\Phi_S^k)$ is by necessity included either in $(\Omega_0^{x[n]} - k\mathbf{j}) \subset \Omega_0^{x[n]}$ or in $\Omega_1^{x[n]}$. As a consequent, there exists $i \in \{0,1\}$ such that $S_c \in \Omega_i^{x[n]}$. One easily finds the same result as if $k < 0$.

Consider a given nonzero integer k. Since $t_{x[n]}(u)$ is continuous for all n, $\Omega_0^{x[n]}$ and $\Omega_1^{x[n]}$ are measurable. From property (ii) of Proposition C.2 and the fact S is measurable, Φ_S^k is measurable. In other words, $m(\Phi_S^k) \geq 0$. Assume that $m(\Phi_S^k) > 0$. We must

be able to find one nonempty open ball $R \subset \Phi_S^k$. Since R is a connected set, there exists $i_0 \in \{0,1\}$ such that $R \in \Omega_{i_0}^{x[0]}$. Suppose that $\tilde{\mathcal{M}}_n(R)$ has been proven to be a connected nonempty open set at some $n \geq 0$. Since $\tilde{\mathcal{M}}_n(R) \subset \tilde{\mathcal{M}}_n(\Phi_S^k)$, there exists $i_n \in \{0,1\}$ such that $\tilde{\mathcal{M}}_n(R) \subset \Omega_{i_n}^{x[n]}$. Then, $\tilde{\mathcal{M}}_{n+1}(R)$ is the transformation of $\tilde{\mathcal{M}}_n(R)$ by M_{i_n}, which is an affine mapping. So $\tilde{\mathcal{M}}_{n+1}(R)$ is a connected nonempty open set.

In this induction, we have proved that $\tilde{\mathcal{M}}_n(R)$ is either a subset of $\Omega_0^{x[n]}$ or a subset of $\Omega_1^{x[n]}$ for each $n \in \mathbb{N}$. Then, for any given $\boldsymbol{w} \in R^2$, $\boldsymbol{w} \in R^2 \subset (\Omega_0^{x[n]})^2 \cup (\Omega_1^{x[n]})^2 = \Pi_0^{x[n]}$ for all $n \in \mathbb{N}$. Since $\tilde{\mathcal{M}}_n(R) \subset S_n$ is bounded, $\boldsymbol{w}$ satisfies the conditions of Proposition 5.2.1, which implies that $d_1(\boldsymbol{w}) = 0$. We may conclude that $(R)_1$ is a constant. But this is impossible since R is a nonempty open set. To avoid this contradiction, for any $k \in \mathbb{Z} \setminus \{0\}$, $m(\Phi_S^k)$ must be 0. This with (15) follows that $\mathbf{p}_{\mathrm{u}}(\Phi_S)$ is of zero measure.

E.4 Proof of Theorem 5.2.3

Consider $\boldsymbol{w} \in \mathcal{Y}_S^{\uparrow}$, $\mathbf{d}(\boldsymbol{w}) = k\mathbf{j}$ where k is some positive integer. Since all threshold are continuous functions, if $\mathbf{p}_{\mathrm{u}}(\boldsymbol{w}) \in \Omega_1^{x[0]}$ then $\mathbf{p}_{\mathrm{v}}(\boldsymbol{w}) = \mathbf{p}_{\mathrm{u}}(\boldsymbol{w}) + k\mathbf{j} \in \Omega_1^{x[0]}$. Therefore, $\boldsymbol{w} \notin \Pi_{-1}^{x[0]}$. So $\boldsymbol{w} \in \Pi_0^{x[0]} \cup \Pi_1^{x[0]}$. Then from (5.9), $\mathbf{d}(\mathcal{M}_x \boldsymbol{w}) = \mathbf{d}(\boldsymbol{w}) - k\mathbf{i}$ where $k \in \{0,1\}$. When $k = 0$, $\mathbf{d}(\mathcal{M}_x \boldsymbol{w}) = \mathbf{d}(\boldsymbol{w}) \in \{0\} \times \mathbb{Z}$ and when $k = 1$, $\mathbf{d}(\mathcal{M}_x \boldsymbol{w}) = \mathbf{d}(\boldsymbol{w}) - \mathbf{i} \in \mathbb{Z}^- \times \mathbb{Z}$. This implies that $\mathcal{M}_x \boldsymbol{w} \in \mathcal{Y}_S^{\uparrow} \cup \mathcal{Y}_S^{-}$.

Next, consider $\boldsymbol{w} \in \mathcal{Y}_S^{\uparrow} \cap \mathcal{Y}_S^{-}$. Assume that $\tilde{\mathcal{M}}_n \boldsymbol{w} \in \mathcal{Y}_S^{\uparrow} \cap \mathcal{Y}_S^{-}$ for all $n \in \mathbb{N}$. Then $d_1(\tilde{\mathcal{M}}_n \boldsymbol{w}) \leq 0$ for all $n \in \mathbb{N}$. According to (5.9), $d_2(\tilde{\mathcal{M}}_{n+1} \boldsymbol{w}) = d_2(\tilde{\mathcal{M}}_n \boldsymbol{w}) + d_1(\tilde{\mathcal{M}}_n \boldsymbol{w})$. The sequence $d_2(\tilde{\mathcal{M}}_n \boldsymbol{w})$ is then non-increasing with n. Now, whenever $\tilde{\mathcal{M}}_n \boldsymbol{w} \in \mathcal{Y}_S^{-}$,

$d_1(\tilde{\mathcal{M}}_n\boldsymbol{w}) \leq -1$. Consider the set $N := \left\{n \in \mathbb{N} : \tilde{\mathcal{M}}_n\boldsymbol{w} \in \mathcal{Y}_S^-\right\}$. If it is infinite, the sequence $d_2(\tilde{\mathcal{M}}_n\boldsymbol{w})$ will by necessity diverge to $-\infty$, which is impossible since $d_2(\tilde{\mathcal{M}}_n\boldsymbol{w}) = (\tilde{\mathcal{M}}_n\mathbf{v})_2 - (\tilde{\mathcal{M}}_n\mathbf{u})_2 > -\xi$. Assume now that N is finite and call n_1 its largest number. Then for all $n > n_1$, $\tilde{\mathcal{M}}_n\boldsymbol{w} \in \mathcal{Y}_S^\uparrow$ which implies that $d_1(\tilde{\mathcal{M}}_n\boldsymbol{w}) = 0$. According to (5.9), because $d_1(\tilde{\mathcal{M}}_n\boldsymbol{w})$ keeps same value, we conclude that $\tilde{\mathcal{M}}_n\boldsymbol{w} \in \Pi_0$ for all $n > n_1$. But this is impossible since $\boldsymbol{w} \notin \mathcal{X}_S$. Therefore, there exists $n \in \mathbb{N}$ such that $\tilde{\mathcal{M}}_n\boldsymbol{w} \notin \mathcal{Y}_S^\uparrow \cap \mathcal{Y}_S^-$. Let us call n_0 the smallest of such integers. By necessity, $\tilde{\mathcal{M}}_{n_0-1}\boldsymbol{w} \in \mathcal{Y}_S^-$ (otherwise $\tilde{\mathcal{M}}_{n_0}\boldsymbol{w}$ would belong to $\mathcal{Y}_S^\uparrow \cup \mathcal{Y}_S^-$, the reason was shown in the previous paragraph). So $d_1(\tilde{\mathcal{M}}_{n_0-1}\boldsymbol{w}) \leq -1$. Then from (5.9), $d_1(\tilde{\mathcal{M}}_{n_0}\boldsymbol{w}) \leq 0$. Now, $d_1(\tilde{\mathcal{M}}_{n_0}\boldsymbol{w})$ must be 0 because $\tilde{\mathcal{M}}_{n_0}\boldsymbol{w} \notin \mathcal{Y}_S^-$. Meanwhile we must have $d_2(\tilde{\mathcal{M}}_{n_0}\boldsymbol{w}) < 0$, because $\tilde{\mathcal{M}}_{n_0}\boldsymbol{w} \notin \mathcal{Y}_S^\uparrow$ and $\mathbf{d}(\boldsymbol{w}) \neq \mathbf{0}$ (since $\boldsymbol{w} \notin \mathcal{X}_0$). So $\tilde{\mathcal{M}}_{n_0}\boldsymbol{w} \in \mathcal{Y}_S^\downarrow$.

This proves (i). The proof of (ii) is similar.

E.5 Proof of Proposition 5.2.4

From Theorem 5.2.3, there exists $n \in \mathbb{N}$ such that

$$\mathbf{d}(\tilde{\mathcal{M}}_n\boldsymbol{w}) \in \mathcal{Y}_S^\uparrow \cup \mathcal{Y}_S^-,$$

$$\mathbf{d}(\tilde{\mathcal{M}}_{n+1}\boldsymbol{w}) \in \mathcal{Y}_S^\downarrow.$$

According to (5.9), $\mathbf{d}(\mathcal{M}_x(\mathcal{Y}_S^\uparrow)) \cap \mathbf{d}(\mathcal{Y}_S^\downarrow) = \emptyset$. So $(\mathbf{d}(\tilde{\mathcal{M}}_n\boldsymbol{w}), \mathbf{d}(\tilde{\mathcal{M}}_{n+1}\boldsymbol{w})) \in \mathcal{Y}_S^- \times \mathcal{Y}_S^\downarrow$. So there exists $u_1 \leq -1, u_2 \in \mathbb{R}$ and $k \leq -1$ such that

$$\mathbf{d}(\tilde{\mathcal{M}}_n\boldsymbol{w}) = (u_1, u_2),$$

$$\mathbf{d}(\tilde{\mathcal{M}}_{n+1}\boldsymbol{w})) = (0, k).$$

Since there exists $k \in \{-1, 0, 1\}, \tilde{\mathcal{M}}_n \boldsymbol{w} \in \Phi_k^{x[n]}$, according to (5.9),

$$u_1 + k = 0,$$

$$u_1 + u_2 = k.$$

Then the only possible case is $k = 1$ and $u_1 = -1$. So $u_2 = k + 1 \leq 0$.

This proves (i). The proof of (ii) is similar.

E.6 Proof of Theorem 5.3.1

We assume $S' = \{(u_1, u_2) \in S : u_2 \geq f(u_1)\}$. Consider $\mathbf{u} \in S \setminus \mathbf{p}_{\mathrm{u}}(\mathcal{X}_S^{\emptyset})$. Since S' contains at least one tile, there exists $\mathbf{v} \in S'$ such that, $\mathbf{v} - \mathbf{u} \in \mathbb{Z}^2$. Define $\boldsymbol{w} := (\mathbf{u}, \mathbf{v})$. Then $\boldsymbol{w} \in \mathcal{W}_S \setminus \mathcal{X}_S^{\emptyset}$. According to definitions from (5.13) to (5.14), $\{\mathcal{X}_S^0, \mathcal{Y}_S\}$ is a partition of $\mathcal{W}_S \setminus \mathcal{X}_S^{\emptyset}$. If $\boldsymbol{w} \in \mathcal{X}_S^0$, there exists n such that $\mathrm{M}^n\mathbf{u} = \mathrm{M}^n\mathbf{v}$, the proposition is clearly true. We assume $\boldsymbol{w} \in \mathcal{Y}_S$. According to Theorem 5.2.3 in the DC inputs case, there exists $n \in \mathbb{N}$ such that $\mathrm{M}^n\boldsymbol{w} \in \mathcal{Y}_S^{\downarrow}$. This implies that $(\mathrm{M}^n\mathbf{u})_2 > (\mathrm{M}^n\mathbf{v})_2 \geq f((\mathrm{M}^n\mathbf{v})_1) = f((\mathrm{M}^n\mathbf{u})_1)$. Since $\mathrm{M}^n\mathbf{u} \in S$, we conclude that $\mathrm{M}^n\mathbf{u} \in \{(u_1, u_2) \in S : u_2 \geq f(u_1)\} = S'$. Next, $m(\mathbf{p}_{\mathrm{u}}(\mathcal{X}_S^{\emptyset})) = 0$ is a consequent of Proposition 5.2.2 when positively invariant set S satisfies all conditions in the proposition.

Similarly, we can prove the theorem in the case $S' = \{(u_1, u_2) \in S : u_2 < f(u_1)\}$. Therefore, the theorem is proven.

F Proofs for propositions of Chapter 6

F.1 Proof of Proposition 6.1.1

We first introduce a proposition to show where $\mathbf{p}$ is located.

Proposition A.1 *The point* $\mathbf{p}$ *defined in* (6.3) *belongs to* H_1 *which is determined by* h*, where* h *is defined in* (3.20)*, and* $h^i := p_i$*, where* p_i *is defined in* (3.16)*.*

Proof: According to (6.3) and (2.14), we have

$$(\mathbf{p})_1 = -(s-1)x_0,$$

$$(\mathbf{p})_2 = -(s-1)(p_1 - x_1).$$

It follows from (3.20), (3.33), (2.14) and writing $s = \frac{5}{2} + \zeta$, we find

$$h^1(\mathbf{p}) - h^0(\mathbf{p}) = \tfrac{1}{4}(4s^2 - 6s + 1)x + \tfrac{1}{4}(2s^2 - 7s + 5) = \tfrac{1}{4}(4\zeta^2 + 14\zeta + 11)x + \tfrac{1}{4}(2\zeta^2 + 3\zeta).$$

Under Conditions 2.4.4 and 6.1.2, $x \geq 0$ and $\zeta > 0$. So $h^1(\mathbf{p}) - h^0(\mathbf{p}) > 0$. Hence $\mathbf{p} \in H_1$. ■

Since $\mathbf{p} \in H_1$, we just need to prove $h^1(\mathbf{p}) > \ell_h$. Since $\delta = 0$, from (3.65), (3.62), and (3.64), We have $\ell_h = \max(h^0(\mathbf{b}^0), h^0(\mathbf{a}^1))$. By using (3.41) and replacing s by $\frac{5}{2} + \zeta$, we find that

$$\begin{array}{rl} h^1(\mathbf{p}) - h^0(\mathbf{b}^0) = f_1(x,\zeta) := & \frac{1}{2(6\zeta+11)^2}\,(a_1\,x^2 + b_1\,x + c_1) \\ h^1(\mathbf{p}) - h^0(\mathbf{a}^1) = f_2(x,\zeta) := & \frac{1}{2}\,(a_2\,x^2 + b_2\,x + c_2) \end{array}$$

where the coefficients a_i, b_i and c_i are functions of ζ as described in the following table.

i	a_i	b_i	c_i
1	$-(16\zeta^4{+}96\zeta^3{+}212\zeta^2{+}208\zeta{+}77)$	$16\zeta^4{+}112\zeta^3{+}240\zeta^2{+}196\zeta{+}50$	$12\zeta^4{+}48\zeta^3{+}65\zeta^2{+}30\zeta$
2	$-(\zeta^2 + 3\zeta + 2)$	$\zeta^2 + 5\zeta + 5$	$\frac{1}{4}(3\zeta^2 + 5\zeta)$

Under Condition 6.1.2, $\zeta > 0$. One can see by direct inspection that $a_i < 0$ and $c_i > 0$ for both $i = 1, 2$. So for any given $\zeta > 0$, $f_i(x,\zeta)$ is a quadratic functions of x of negative curvature. Given that $x \in [0, \frac{1}{2})$ from Condition 2.4.4, $f_i(x,\zeta) \geq \min\big(f_i(0,\zeta), f_i(\frac{1}{2},\zeta)\big)$. We already have $f_i(0,\zeta) > 0$ since $c_i > 0$. Finally, we find that $f_1(\frac{1}{2},\zeta) = \frac{2}{(6\zeta+11)^2}(4\zeta^4{+}20\zeta^3{+}33\zeta^2{+}19\zeta{+}\frac{23}{16}) > 0$ and $f_2(\frac{1}{2},\zeta) = \frac{1}{2}(\zeta^2{+}3\zeta{+}2) > 0$.

F.2 Proof of Proposition 6.2.1

We first prove the following propositions.

Proposition B.2

$$\mathrm{M}_0(R_0 \cap \Omega_1^1) \subset \Omega_1^1, \tag{18}$$

where $\Omega_1^1 := \mathrm{M}_1(\Omega_1)$.

Proof: Consider any $\mathbf{u} \in \Omega_1^1$ such that $u_1 \geq (\mathbf{p})_1$. With (6.3), since T^1 has more positive slope than $\mathrm{M}_0^{-1}(T^1)$, $\mathbf{u} \in \mathrm{M}_0^{-1}(\Omega_1^1)$. Then $\mathrm{M}_0\mathbf{u} \in \Omega_1^1$. To prove (18), then one just needs to show that

$$\inf\big((R_0 \cap \Omega_1^1)_1\big) = (\mathbf{p})_1. \tag{19}$$

Clearly, $R_0 \subset G_{h(\mathbf{p})}$. So $R_0 \cap \Omega_1^1 \subset G_{h(\mathbf{p})} \cap \Omega_1^1$. Since $\mathbf{p} \in H_1$ according to Proposition A.1, $h(\mathbf{p}) = h^1(\mathbf{p})$, so $\mathbf{p} \in \partial G_{h(\mathbf{p})}$. Since $\partial G_{h(\mathbf{p})}$ is a concave parabola and T^1 is a straight line, the u_1 projection of set $G_{h(\mathbf{p})} \cap \Omega_1^1$ is bounded by the projection of their intersections. Note from (6.3) that $\mathbf{p} \in T^1$. So $\partial G_{h(\mathbf{p})} \cap T^1$ is of the form $\{\mathbf{p}, \mathbf{q}\}$, where $\mathbf{q}$ may or may not be distinct from $\mathbf{p}$. A point $\mathbf{u} = (u_1, u_2)$ belongs to $\{\mathbf{p}, \mathbf{q}\}$ if and only if $h^1(\mathbf{u}) = h^1(\mathbf{p})$ and $u_2 = -(s-1)(u_1 - x_1)$. By applying (3.31), we find that u_1 satisfies the second degree equation $\frac{1}{2}u_1^2 + x_1(s - \frac{3}{2})u_1 + c = 0$, where c is some constant. Since $(\mathbf{p})_1$ and $(\mathbf{q})_1$ are the two roots of this equation, $\frac{1}{2}\big((\mathbf{p})_1 + (\mathbf{q})_1\big) = -x_1(s - \frac{3}{2}) > 0$. But from (3.41), $(\mathbf{p})_1 = (1-s)x_0 < 0$. So $\mathbf{p}$ is the left intersection, which implies (19).

■

Proposition B.3 *The set* $R_0 \cap \Omega_1^1$ *is positively invariant by* M.

Proof:

Now we prove Proposition B.3. First of all R_0 is positively invariant: $\mathrm{M}(R_0) \subset R_0$. So all we need to prove is that

$$M(R_0 \cap \Omega_1^1) \subset \Omega_1^1.$$

Next, from the definition of M in (2.21),

$$\mathrm{M}(R_0 \cap \Omega_1^1) = \mathrm{M}_1(R_0 \cap \Omega_1^1 \cap \Omega_1) \cup \mathrm{M}_0(R_0 \cap \Omega_1^1 \cap \Omega_0). \tag{20}$$

Since

$$R_0 \cap \Omega_1^1 \cap \Omega_1 \subset \Omega_1 \Rightarrow \mathrm{M}_1(R_0 \cap \Omega_1^1 \cap \Omega_1) \subset \Omega_1^1,$$

and

$$R_0 \cap \Omega_1^1 \cap \Omega_0 \subset R_0 \cap \Omega_1^1 \Rightarrow \mathrm{M}_0(R_0 \cap \Omega_1^1 \cap \Omega_0) \subset \mathrm{M}(R_0 \cap \Omega_1^1),$$

with Proposition B.2, two subsets in the right hand side of (20) are subset of Ω_1^1. Hence the proposition is proven.

■

We now prove Proposition 6.2.1. For simpler notation, let us write $G_e = G$. Using (6.7) and (2.16), we have

$$\mathrm{M}_1(\overline{G}) = \overline{G}, \quad \mathrm{M}_0(\overline{G}) = \overline{G}+\mathbf{i} \quad \text{and} \quad \mathrm{M}_1\big(\mathrm{M}_0(\overline{G})\big) = \overline{G}+\mathrm{L}\mathbf{i} = \overline{G}+\mathbf{i}+\mathbf{j} \subset \overline{G}+\mathbf{i} = \mathrm{M}_0(\overline{G}).$$

From (6.9), it is also clear that

$$\mathrm{M}_0(\Omega_1^1) \ \subset \ \overline{G}.$$

By applying all of these relations, together with (18), we have

$$\begin{array}{ll} \mathrm{M}_0(R_1) \subset \mathrm{M}_0(R_0 \cap \Omega_1^1) \subset \Omega_1^1, & \mathrm{M}_1(R_1 \cap \Omega_1) \subset \mathrm{M}_1(\Omega_1) = \Omega_1^1, \\ \mathrm{M}_0(R_1) \subset \mathrm{M}_0(\Omega_1^1) = \mathrm{M}_0(\Omega_1^1) \subset \overline{G}, & \mathrm{M}_1(R_1) \subset \mathrm{M}_1(\overline{G}) = \overline{G}, \\ \mathrm{M}_0(R_1) \subset \mathrm{M}_0(\overline{G}), & \mathrm{M}_1(R_1) \subset \mathrm{M}_1\big(\mathrm{M}_0(\overline{G})\big) \subset \mathrm{M}_0(\overline{G}). \end{array}$$

It follows that

$$\mathrm{M}(R_1) = \mathrm{M}_0(R_1 \cap \Omega_0) \cup \mathrm{M}_1(R_1 \cap \Omega_1) \ \subset \ \Omega_1^1 \cup \overline{G} \cup \mathrm{M}_0(\overline{G}).$$

Since R_0 is positively invariant, we have the additional trivial inclusions $\mathrm{M}(R_1) \subset \mathrm{M}(R_0) \subset R_0$. This completes the proof of the proposition.

F.3 Proof of Proposition 6.2.2

We start with the following preliminary property.

Proposition C.4 *For all* $\mathbf{u} \in R_2 \cap \Omega_0$, $(\mathbf{u})_1 < (\mathbf{d})_1$.

Proof: From (6.8) and (6.11),

$$R_2 \cap \Omega_0 \ \subset \ R_1 \cap \Omega_0 \ \subset \ \Omega_1^1 \cap \overline{G} \cap \Omega_0. \tag{21}$$

Using the explicit descriptions of T and G,

$$\forall \mathbf{u} \in \overline{G} \cap \Omega_0, \qquad (\mathbf{u})_1 < (\mathbf{d})_1 \ \text{ or } \ (\mathbf{u})_1 > -2x_1(s - \tfrac{1}{2}) - (\mathbf{d})_1$$

But with the explicit expression of T and T^1,

$$\forall \mathbf{u} \in \Omega_1^1 \cap \Omega_0, \qquad (\mathbf{u})_1 < -(s-1)x_1$$

With Conditions 2.4.4 and 6.1.2,

$$-(s-1)x_1 < -2x_1(s - \tfrac{1}{2}) - (\mathbf{d})_1.$$

So when $\mathbf{u} \in R_2 \cap \Omega_0$, we obtain $(\mathbf{u})_1 < (\mathbf{d})_1$ due to (21). ■

We now prove Proposition 6.2.2. Consider a given $\mathbf{u} \in R_2 \cap \Omega_0$. There are two cases.

Case 1: $(\mathbf{u})_1 < (\mathbf{b}_0)_1$. From (3.31), $h_1(\mathrm{M}_0\mathbf{u}) - h_1(\mathbf{u}) = u_1 + \frac{1}{2}x_0$. Since $(\mathbf{b}_0)_1 = -\frac{1}{2}x_0$, then $h_1(\mathrm{M}_0\mathbf{u}) < h_1(\mathbf{u}) \leq e'$, since $\mathbf{u} \in G_{e'}$.

Case 2: $(\mathbf{u})_1 \geq (\mathbf{b}_0)_1$. By Proposition C.4, there exists $\theta \in [0, 1)$ such that $(\mathbf{u})_1 = (1-\theta)(\mathbf{b}_0)_1 + \theta(\mathbf{d})_1$. Since $\mathbf{b}_0, \mathbf{d} \in T$, the point $\mathbf{u}_T := (1-\theta)\mathbf{b}_0 + \theta\mathbf{d} \in T$. Since $(\mathbf{u}_T)_1 = (\mathbf{u})_1$ and $\mathbf{u} \in \Omega_0$, then $\mathbf{u} = \mathbf{u}_T - d\mathbf{j}$ with $d > 0$. Then $\mathrm{M}_0\mathbf{u} = \mathrm{M}_0\mathbf{u}_T - d\mathbf{j}$ according to (2.18). Now, $\mathrm{M}_0\mathbf{u}_T = (1-\theta)\mathrm{M}_0\mathbf{b}_0 + \theta\mathrm{M}_0\mathbf{d}$. On the one hand $h_1(\mathrm{M}_0\mathbf{d}) = h_1(\mathbf{c}) \leq e'$. On the other hand, since $\mathbf{b}_0 \in L_0$, then $h_1(\mathrm{M}_0\mathbf{b}_0) - h_1(\mathbf{b}_0) = 0$ according to (3.31), which implies that $h_1(\mathrm{M}_0\mathbf{b}_0) = h_1(\mathbf{b}_0) \leq e'$. By convexity of h_1, $h_1(\mathrm{M}_0\mathbf{u}_T) \leq e'$. Finally $h_1(\mathrm{M}_0\mathbf{u}) = h_1(\mathrm{M}_0\mathbf{u}_T) - d < e'$.

We have thus shown that $\mathrm{M}_0(R_2 \cap \Omega_0) \subset G_{e'}$. Meanwhile, $\mathrm{M}_1(R_2 \cap \Omega_1) \subset \mathrm{M}_1(G_{e'}) = G_{e'}$ since $G_{e'}$ is invariant by M_1. So globally $\mathrm{M}(R_2) \subset G_{e'}$. Finally, $\mathrm{M}(R_2) \subset \mathrm{M}(R_1) \subset R_1$ since R_1 is positively invariant. So $\mathrm{M}(R_2) \subset R_1 \cap G_{e'} = R_2$.

F.4 Proof of Proposition 6.3.2

Proposition D.5 *For any bounded set S, let us define $|S|_1 := \sup_{\mathbf{u}\in S} |u_1|$. Then, for any $d > 0$,*

$$|\mathcal{Q}_d(G)|_1 = \tfrac{1}{2}x_0 - dx_1. \tag{22}$$

Proof: Because of the structure of $\mathcal{Q}_d(G)$ given in (6.14), the value $|\mathcal{Q}_d(G)|_1$ is attained by $|u_1|$ where $\mathbf{u}$ is a point of $\partial(G_{-1,1} \cup G \cup G_{0,1}) \cap \partial G_{0,d}$. Since ∂G and $\partial G_{0,d}$ have no intersection, $\mathbf{u}$ must belong to $(\partial G_{-1,1} \cup \partial G_{0,1}) \cap \partial G_{0,d}$. Now, $\mathbf{u} \in \partial(G+\mathbf{v}) \cap \partial(G+\mathbf{v}')$ if and only if both $\mathbf{u}-\mathbf{v}$ and $\mathbf{u}-\mathbf{v}'$ belong to G. This implies that $h_1(\mathbf{u}-\mathbf{v}) = h_1(\mathbf{u}-\mathbf{v}')$. By applying (3.31), we find that $u_1 = v_1 + v_1' + x_1 + 2x_1\frac{v_2'-v_2}{v_1'-v_1}$.

With $\mathbf{v}' = (0, d)$ and any $\mathbf{v} \in \big\{(-1,1),(1,0)\big\}$, we find $|u_1| = \frac{1}{2}x_0 - dx_1$, using the relation $x_0 = x_1 + 1$ resulting from (2.14). ■

Proposition D.6 *Let us define the set*

$$K := \big\{(0,0),(-1,0),(0,-1),(1,-1)\big\}. \tag{23}$$

For any $d \in [0,2]$ *and* $\mathbf{k} \in \mathbb{Z}^2 \setminus K$, $\mathcal{Q}_1(G) \cap \big(\mathcal{Q}_d(G) + \mathbf{k}\big) = \emptyset$.

Proof: Define the set $N := \big\{(-1,1),(0,0),(1,0)\big\}$. Consider any $\mathbf{n} \in N$ and $c \geq 1$. By successively using (6.14), (6.6) and (6.14) again, we have,

$$\mathcal{Q}_d(G) \ \subset\ \overline{G} + \mathbf{n} \ \subset\ (\overline{G} + \mathbf{j}) - c\mathbf{j} + \mathbf{n} \ \subset\ \overline{\mathcal{Q}_1(G)} - c\mathbf{j} + \mathbf{n}. \tag{24}$$

Meanwhile, since $d \leq 2$, we obtain using the same successive properties,

$$\mathcal{Q}_1(G) \ \subset\ \overline{G} + \mathbf{n} \ \subset\ (\overline{G} + d\mathbf{j}) - \mathbf{j} - c\mathbf{j} + \mathbf{n} \ \subset\ \overline{\mathcal{Q}_d(G)} - c\mathbf{j} + (\mathbf{n} - \mathbf{j}). \tag{25}$$

Now (24) can be rewritten as

$$\mathcal{Q}_1(G) \ \subset\ \overline{\mathcal{Q}_d(G)} + c\mathbf{j} - \mathbf{n}. \tag{26}$$

By restricting c to be an integer and by noticing that $(N - \mathbf{j}) \cup (-N) = K$, (25) and (26) can be easily proven to imply

$$\mathcal{Q}_1(G) \ \subset\ \overline{\mathcal{Q}_d(G)} + \mathbf{k}, \tag{27}$$

for all $\mathbf{k} \in \big(\{-,1,0,1\} \times \mathbb{Z}\big) \setminus K$. Meanwhile, we also have $\mathcal{Q}_d(G) \cap \big(\mathcal{Q}_1(G) + \mathbf{k}\big) = \emptyset$ with any vector $\mathbf{k}$ such that $|k_1| > |\mathcal{Q}_d(G)|_1 + |\mathcal{Q}_1(G)|_1$. Using (22),

$$|\mathcal{Q}_d(G)|_1 + |\mathcal{Q}_1(G)|_1 \leq x_0 - (d+1)x_1 = (2-d)x + \tfrac{d}{2} \leq \tfrac{3}{2}$$

due to the inequalities $1 < d \leq 2$ and $x < \frac{1}{2}$. So we also have (26), for all $\mathbf{k} \in \mathbb{Z}^2$ such that $|k_1| \geq 2$. We conclude that (26) is true for all $\mathbf{k} \in \mathbb{Z}^2 \setminus K$. ■

We now prove Proposition 6.3.2. By Proposition D.6,

$$\begin{aligned}\mathcal{T}\big(\mathcal{Q}_d(G)\big) \supset \mathcal{T}\big(\mathcal{Q}_d(G)\big) \cap \mathcal{Q}_1(G) &= \bigcap_{\mathbf{k}\in\mathbb{Z}^2\setminus\{\mathbf{0}\}} \big(\overline{\mathcal{Q}_d(G)} + \mathbf{k}\big) \cap \mathcal{Q}_1(G) \\ &= \bigcap_{\mathbf{k}\in K\setminus\{\mathbf{0}\}} \big(\overline{\mathcal{Q}_d(G)} + \mathbf{k}\big) \cap \mathcal{Q}_1(G) \\ &\supset \bigcap_{\mathbf{k}\in K\setminus\{\mathbf{0}\}} \big(\overline{G}_{0,d} + \mathbf{k}\big) \cap \mathcal{Q}_1(G), \qquad (28)\end{aligned}$$

since $\mathcal{Q}_d(G) \subset G_{0,d}$. Note that $\mathcal{Q}_1(G) = \bigcap_{\mathbf{k}\in K\setminus\{\mathbf{0}\}} \big(\overline{G}_{0,1} + \mathbf{k}\big) \cap G_{0,1}$ and $\overline{G}_{0,d} \subset \overline{G}_{0,1}$. Then

$$\begin{aligned}\bigcap_{\mathbf{k}\in K\setminus\{\mathbf{0}\}} \big(\overline{G}_{0,d} + \mathbf{k}\big) \cap \mathcal{Q}_1(G) &= \bigcap_{\mathbf{k}\in K\setminus\{\mathbf{0}\}} (\overline{G}_{0,d} + \mathbf{k}) \cap G_{0,1} \\ &= \Big(\bigcap_{\mathbf{k}\in K\setminus\{\mathbf{0}\}} (\overline{G}_{0,1} + \mathbf{k}) \cap G_{0,2-d} \Big) + (d-1)\mathbf{j} \\ &= \mathcal{Q}_{2-d}(G) + (d-1)\mathbf{j}. \qquad (29)\end{aligned}$$

F.5 Proof of Proposition 6.3.3

With (6.12) and the inequality $-x_1 > 0$, $\frac{e'-e}{-x_1} < 2$ if and only if

$$h_1(\mathbf{c}) - e + 2x_1 < 0 \qquad \text{and} \qquad h_1(\mathbf{b}_0) - e + 2x_1 < 0. \qquad (30)$$

Let us evaluate e. Since $T^1 \subset \Omega_1^1$, it is clear from (6.9) that $e \le \min_{\mathbf{v}\in \mathrm{M}_0(T^1)} h_1(\mathbf{v})$. Meanwhile, for any $\mathbf{u} \in \Omega_1^1$, there exists $d > 0$ such that $\mathbf{u} - d\mathbf{j} \in T^1$. With (2.18) and (3.31), we have $\min_{\mathbf{v}\in \mathrm{M}_0(T^1)} h_1(\mathbf{v}) \le h_1(\mathrm{M}_0(\mathbf{u} - d\mathbf{j})) = h_1(\mathrm{M}_0\mathbf{u}) + dx_1 \le h_1(\mathrm{M}_0\mathbf{u})$. Since this is true for any $\mathbf{u} \in \Omega_1^1$, this proves that $e = \min_{\mathbf{v}\in \mathrm{M}_0(T^1)} h_1(\mathbf{v})$. By applying (6.1), (3.33) and (3.31), we find for all $\mathbf{u} \in \mathrm{M}_0(T^1)$ that $h_1(\mathbf{u}) = f(u_1)$ where

$$f(u_1) := \tfrac{1}{2}\, u_1^2 + (s - \tfrac{5}{2})x_1\, u_1 + \big((2-s)x_1 + 2(\tfrac{25}{16} - s)x_1^2\big).$$

By trivial derivation of the minimum a second degree polynomial, we find

$$e = \min_{u_1\in\mathbb{R}} f(u_1) = (2-s)x_1 + \tfrac{1}{2}(1-s)s\, x_1^2. \qquad (31)$$

We recall from (6.10) that $\mathbf{c} = \mathrm{M}_0\mathbf{d}$ where $\mathbf{d}$ is the point of $\partial G_e \cap T$ of minimal abscissa. So $\mathbf{d} = (d_1, d_2)$ satisfies $d_2 = -s\,d_1$ with $h_1(\mathbf{d}) = e$. By applying (3.31), we look for the smallest solution to the equation $\frac{1}{2}(d_1 - \frac{1}{2}x_1)^2 + x_1 s d_1 = e$ and find

$$d_1 = -x_1(s - \tfrac{1}{2}) - \sqrt{-2x_1(s-2)}. \tag{32}$$

Then, $h_1(\mathbf{c}) - e = h_1(\mathrm{M}_0\mathbf{d}) - h_1(\mathbf{d}) = d_1 + \frac{1}{2}x_0$ as a result of (3.31) with $i = 0$. With (32) and (2.14),

$$h_1(\mathbf{c}) - e + 2x_1 = -(x_1 s) - \sqrt{4x_1 - 2(x_1 s)} + 3x_1 + \tfrac{1}{2}.$$

Meanwhile, by trivial derivation of the minimum a second degree polynomial, we find

$$e = \min_{u_1 \in \mathbb{R}} f(u_1) = (2-s)x_1 + \tfrac{1}{2}(1-s)s\,x_1^2. \tag{33}$$

Then the expression of $h_1(\mathbf{b}_0)$ from (3.41) and the relation $x = x_1 + \frac{1}{2}$,

$$h_1(\mathbf{b}_0) - e + 2x_1 = \tfrac{1}{2}\big((x_1 s)^2 + (1 - 2x_1)(x_1 s) + (x_1 + \tfrac{1}{2})^2\big).$$

The two second inequalities of (30) are then respectively equivalent to

$$\begin{array}{rcccl} x_1 - \frac{1}{2} - \sqrt{-2x_1} & < & x_1 s & < & x_1 - \frac{1}{2} + \sqrt{-2x_1}, \\ 3x_1 - \frac{1}{2} - \sqrt{-2x_1} & < & x_1 s & < & 3x_1 - \frac{1}{2} + \sqrt{-2x_1}. \end{array}$$

Since $3x_1 < x_1$, this system of inequalities reduces to

$$x_1 - \tfrac{1}{2} - \sqrt{-2x_1} \;<\; x_1 s \;<\; 3x_1 - \tfrac{1}{2} + \sqrt{-2x_1}.$$

After writing $x_1 = x - \frac{1}{2}$, we obtain $s_1(x) < s < s_2(x)$ where $s_1(x)$ and $s_2(x)$ are given in (6.17).

F.6 Proof of Proposition 6.3.2

Proposition F.7 *Define*

$$B := \bigcup_{i\in\mathbb{Z}} (G + \mathbf{k}_i) \qquad \textit{where} \qquad \forall i \in \mathbb{Z}, \quad \mathbf{k}_i := \begin{cases} (i,1), & i \leq -1 \\ (i,0), & i \geq 0 \end{cases} \tag{34}$$

Then, for any $d \in [1,2]$,

$$\overline{B} \cap B_{0,d} \subset \bigcup_{i\in\mathbb{Z}} \big(\mathcal{Q}_d(G) + \mathbf{k}_i\big). \tag{35}$$

Proof: Take any $\mathbf{u} \in \overline{B} \cap B_{0,d}$. Since $\mathbf{u} \in \overline{B}$,

$$\forall j \in \mathbb{Z}, \qquad \mathbf{u} \in \overline{G} + \mathbf{k}_j. \tag{36}$$

Case 1: $\mathbf{u} \in G_{0,d}$. Then, by applying (36) with $j = -1, 0, 1$, one easily finds from (6.14) that $\mathbf{u} \in \mathcal{Q}_d(G) = \mathcal{Q}_d(G) + \mathbf{k}_0$.

Case 2: $\mathbf{u} \in \overline{G}_{0,d} = \overline{G}_{0,d} + \mathbf{k}_0$. Since $\mathbf{u} \in B_{0,d}$, there exists $i \in \mathbb{Z}$ such that $\mathbf{u} \in G_{0,d} + \mathbf{k}_i$. Because the second component of $\mathbf{k}_i$ is upper bounded and the function $g_1(u_1)$ is strictly decreasing when $u_1 > \frac{1}{2}x_1$, there actually exists a smallest integer i such that $\mathbf{u} \in G_{0,d} + \mathbf{k}_i$. On the one hand $i \neq 0$ by assumption, and on the other hand, $\mathbf{u} \in \overline{G}_{0,d} + \mathbf{k}_{i-1}$. Now, $\mathbf{k}_{i-1} = \mathbf{k}_i + (-1, 0)$ from (34) and $\overline{G}_{0,d} \subset \overline{G}_{0,1}$ from (6.6). So $\mathbf{u} \in \overline{G}_{-1,1} + \mathbf{k}_i$. From (36), $\mathbf{u} \in \overline{G} + \mathbf{k}_{i+1}$. When $i \neq -1$, $\mathbf{k}_{i+1} = \mathbf{k}_i + (1, 0)$. So $\mathbf{u} \in \overline{G}_{1,0} + \mathbf{k}_i$. When $i = -1$, this last relation is also true since $\mathbf{u} \in \overline{G}_{0,d} \subset \overline{G}_{0,1} = \overline{G}_{1,0} + \mathbf{k}_{-1}$ since $\mathbf{k}_{-1} = (-1, 1)$. Finally, (36) implies $\mathbf{u} \in \overline{G} + \mathbf{k}_i$. We have thus verified all the conditions required by (6.14) to have $\mathbf{u} \in \mathcal{Q}_d(G) + \mathbf{k}_i$.

■

Given (28) and (29), we only need to prove that

$$\mathcal{T}\big(\mathcal{Q}_d(G)\big) \;\subset \bigcap_{\mathbf{k}\in K\setminus\{\mathbf{0}\}} \big(\overline{G}_{0,d}+\mathbf{k}\big)\cap \mathcal{Q}_1(G). \tag{37}$$

For any $\mathbf{u}\in\mathbb{R}^2$, there exists a smallest integer n such that $u\in B_{0,n+1}$, implying that $u\in\overline{B}_{0,n}$. Then $\mathbb{R}^2$. By applying (35) at $d=1$, we find that

$$\begin{aligned}\mathbb{R}^2 \subset \bigcup_{n\in\mathbb{Z}} (\overline{B_{0,n}}\cap B_{0,n+1}) &\subset \bigcup_{i\in\mathbb{Z},n\in\mathbb{Z}} \big(\mathcal{Q}_1(G)+\mathbf{k}_i+n\mathbf{j}\big)\\ &= \bigcup_{\mathbf{k}\in\mathbb{Z}^2}\big(\mathcal{Q}_1(G)+\mathbf{k}\big) = \mathcal{Q}_1(G)\cup\overline{\mathcal{T}(\mathcal{Q}_1(G))}.\end{aligned}$$

Since $\mathcal{Q}_1(G)\subset\mathcal{Q}_d(G)$, then $\mathcal{T}(\mathcal{Q}_d(G))\subset\mathcal{T}(\mathcal{Q}_1(G))\subset\mathcal{Q}_1(G)$.

Next, one can easily show that

$$\begin{aligned}\bigcup_{\mathbf{k}\in K\setminus\{\mathbf{0}\}}(G_{0,d}+\mathbf{k}) \;\subset\; B_{0,d-1} \;&=\bigcup_{n\in\mathbb{N}^*}(\overline{B}_{0,d-1}\cap B_{0,d})-n\mathbf{j}\\ &\subset \bigcup_{n\in\mathbb{N}^*}(\overline{B}\cap B_{0,d})-n\mathbf{j} \;\subset \bigcup_{i\in\mathbb{Z},n\in\mathbb{N}^*}\big(\mathcal{Q}_d(G)+\mathbf{k}_i-n\mathbf{j}\big),\end{aligned}$$

where the last inclusion was obtained from (35). Note that for all $i\in\mathbb{Z}$ and $n\in\mathbb{N}^*$, $\mathbf{k}_i-n\mathbf{j}\neq\mathbf{0}$. This implies that $\bigcup_{\mathbf{k}\in K\setminus\{\mathbf{0}\}}(G_{0,d}+\mathbf{k}) \;\subset\; \overline{\mathcal{T}(\mathcal{Q}_d(G))}$. This proves (37).

G Proofs for propositions of Chapter 7

G.1 Bounded solution of (7.3)

By writing $\mathbf{y}[n]=(y_1[n],y_2[n])^\top$, it is easy to see that (7.3) is equivalent to

$$\begin{aligned}y_1[n+1]-y_1[n]&=\tilde{x}[n]\\ y_2[n+1]-y_2[n]&=y_1[n].\end{aligned} \tag{38}$$

Assume that $\tilde{x}[n]=a\cos(\omega n+\theta)$, $y_1[n]=a'\cos(\omega n+\theta')$ and $y_2[n]=a''\cos(\omega n+\theta'')$. By mere application of the trigonometric formula $\cos(\beta)=-\cos(\beta+\pi)$ and $\cos(\alpha)+\cos(\beta)=2\cos(\frac{a+b}{2})\cos(\frac{a-b}{2})$, one easily finds that (38) is true if and only if

$2a' \cos(\frac{\omega-\pi}{2}) = a$, $\frac{\omega+\pi}{2} + \theta' = \theta$, $2a'' \cos(\frac{\omega-\pi}{2}) = a'$ and $\frac{\omega+\pi}{2} + \theta'' = \theta'$. This is equivalent to $a' = \frac{a}{2\cos(\frac{\omega-\pi}{2})} = \frac{-a}{2\sin(\omega/2)}$, $a'' = \frac{a}{4\sin^2(\omega/2)}$, $\theta' = \theta - \frac{\omega+\pi}{2}$ and $\theta'' = \theta - \omega - \pi$. The general result of (7.4) is simply obtained by linear combination of solutions.

Bibliography

[1] S.R.Norsworthy, R.Schreier, and G.C.Temes, eds., *Delta-sigma data converters: theory, design and simulation.* IEEE Press, 1996.

[2] R. M. Gray, "Oversampled Sigma-Delta Modulation," *IEEE Trans. Communications*, vol. COM-35, NO. 5, MAY 1987.

[3] R. Schreier, "G. Temes. Understanding Delta-Sigma Data Converters." ISBN 0-471-46585-2.

[4] R. Schreier, "Delta-sigma toolbox," available on: http://www.mathworks/matlabcentral/fileexchange.

[5] N.Rouche, P.Habets, and M.Laloy, "Stability theory by liapunov's direct method." Springer-Verlag, 1977.

[6] R. M. Gray, "Spectral analysis of quantization noise in a single loop sigma-delta modulator with dc input," *IEEE Trans. Commun.*, vol. 37, pp. 588-599, June 1989.

[7] Max W. Hauser, "Principles of Oversampling A/D Conversion," *J. Audio Eng. Soc.*, Vol. 39, No. 1/2, pp. 3-26, 1991 January/February.

[8] Feely, O. "Nonlinear dynamics of sigma-delta modulation," *Circuits and Systems*, Proceedings of the 34th Midwest Symposium on Volume, Issue, 14-17 May 1991, vol.2, pp. 760-763, 1991.

[9] S. Pinault and P. Lopresti, "On the behavior of the double-loop sigma delta modulator," *IEEE Trans. Circuits Systems II,* vol. 40, pp. 467-479, 1993.

[10] Delchamps, D.F., "Nonlinear dynamics of oversampling A-to-D converters," *Proceedings of the 32nd IEEE Conference on 15-17 Dec. Decision and Control,* Page(s):480 - 485 vol.1, 1993.

[11] S. Hein and A. Zakhor, "On the stability of sigma delta modulators," *IEEE Trans,* SP, vol. 41, pp.2322-2348, 1993.

[12] Hongmo Wang, "$\Sigma\Delta$ modulation from the perspective of nonlinear dynamics," *Circuits and Systems. ISCAS '92. Proceedings., IEEE International Symposium on,* vol.3, pp.:1296 - 1299,3-6 May 1992

[13] Hongmo Wang, "On the stability of third-order sigma-delta modulation," *ISCAS,* pp.1377-1380, 1993.

[14] N. T. Thao, "Deterministic Analysis of Oversampled A/D Conversion and Sigma-Delta Modulation, and Decoding Improvements using Consistent Estimates," 1993. PhD dissertation, Department of Electrical Engineering, Columbia University.

[15] P. Steiner, W. Yang, "Stability analysis of the second order $\Sigma\Delta$ modulator," *IEEE International Symposium, Circuits and Systems, ISCAS '94.* vol.5, pp.365 - 368, 30 May-2 June 1994.

[16] B. Zhang, M.V. Goodson, R. Schreier, "Invariant sets for general second-order lowpass delta-sigma modulators with DC inputs," *Proceedings of the ISCAS,* pp. 1-4, 1994.

[17] Orla Feely, "Theory of lowpass and bandpass sigma-delta modulation," *1995 The Institution of Electrical Engineers, Printed and published by the IEE, Savoy Placce, London WC2R 0BL, UK.*

[18] M.V. Goodson, B. Zhang, R. Schreier, "Proving stability of delta-sigma modulators using invariant sets," *Circuits and Systems. ISCAS '95., IEEE International Symposium on,* vol.1 , pp. 633 - 636, 28 April-3 May 1995

[19] O. Feely, D. Fitzgerald, "Bandpass sigma-delta modulation-an analysis from the perspective of nonlinear dynamics," *Circuits and Systems. ISCAS '96., IEEE International Symposium on,* vol.3, pp.146 - 149 ,12-15 May 1996

[20] R. Schreier, M.V. Goodson, B. Zhang, "An algorithm for computing convex positively invariant sets for delta-sigma modulators," *IEEE Trans. Circuits Systems II,* pp. 38-44, 1997.

[21] Ronan Farell and Orla Feely, "Bounding the Integrator Outputs of Second-Order Sigma-Delta Modulators," *IEEE Trans. Circuits and Systems II,* Vol. 45, NO. 6, pp.691-702, JUN 1998.

[22] Alexey Teplinsky, Orla Feely and Alan Rogers, "Phase-Jitter Dynamics of Digital Phase-Locked Loops," *IEEE Trans. Circuits and Systems I,* Vol. 46, NO. 5, pp.545-558, MAY 1999.

[23] Alexey Teplinsky and Orla Feely, "Phase-Jitter Dynamics of Digital Phase-Locked Loops: Part II," *IEEE Trans. Circuits and Systems I,* Vol. 47, NO. 4, pp.458-473, MAY 2000.

[24] S. Güntürk, "Harmonic analysis of two problems in signal quantization and compression," Oct. 2000. PhD dissertation, Program in Applied and Computational Mathematics, Princeton University. http://www.math.nyu.edu/gunturk/research.html.

[25] Ö. Yilmaz, "Stability analysis for several second-order sigma-delta methods of coarse quantization of banklimited functions," *Constr. Approx.*, vol 18, no. 4, pp. 599-623, 2002.

[26] N. T. Thao, "The tiling phenomenon in $\Sigma\Delta$ modulation," *IEEE Trans. Circuits and Systems I,* Vol.51, No.7, pp.1365-1378, July 2004.

[27] N. T. Thao, "Breaking the feedback loop of a class of $\Sigma\Delta$ A/D converters," *IEEE Trans. Signal Processing,* Vol. 52, no. 12, pp.3378-3393, Dec. 2004.

[28] C. S. Güntürk and N. T. Thao, "Ergodic dynamics in $\Sigma\Delta$ quantization: invariant tiles and spectral analysis of error," *Advances in Applied Math,* 34 (2005) 523-560.

[29] S. Zeng and N. T. Thao, "Trapping Sets of Second Order $\Sigma\Delta$ Modulators by Lyapunov Function Approach," *IEEE Trans. Circuits and Systems I,* March 2008 submitted.