Dynamical Systems in Cell Division Cycle, Winnerless Competition Models, and Tensor Approximations

A dissertation presented to
the faculty of
the College of Arts & Sciences of Ohio University

In partial fulfillment
of the requirements for the degree
Doctor of Philosophy

Xue Gong
April 2016
© 2016 Xue Gong. All Rights Reserved.
This dissertation titled
Dynamical Systems in Cell Division Cycle, Winnerless Competition Models, and Tensor
Approximations

by
XUE GONG

has been approved for
the Department of Mathematics
and the College of Arts & Sciences by

Todd R. Young
Professor of Mathematics

Robert Frank
Dean of the College of Arts & Sciences
Abstract

GONG, XUE, Ph.D., April 2016, Mathematics

Dynamical Systems in Cell Division Cycle, Winnerless Competition Models, and Tensor Approximations (144 pp.)

Director of Dissertation: Todd R. Young

This dissertation discusses the application of dynamical systems in three different fields. We study the cell cycle dynamics in a cell-cell coupling system, heteroclinic binding networks modeling sequential working memory and the tensor approximation problems using the gradient flow.

We consider the cell cycle coupling model for budding yeast (*Saccharomyces cerevisiae*), where the rates of progression for cells in a certain phase (*responsive region* $R$) of the cell division cycle are influenced by cells in another phase (*signaling region* $S$). We first study a model to mimic the situation where there is a delay between cells in $S$ producing the chemical agents and cells in $R$ experiencing the feedback from the agents by including a gap between the two regions. We analyze the dynamics of this system for the $k$-periodic solutions and compare the dynamics with the system without a gap.

Second, we study a model where two biologically motivated sources of noise are introduced into the cell cycle system and compare them with the model that has Gaussian white noise perturbations. In simulations, we explore how an ordered two-cluster periodic state of cells becomes disordered by the increase of noise and a uniform distribution emerges for large noise. As a corollary to the results, we can estimate the coupling strength of the cell cycle in yeast autonomous oscillation experiments where clustering is observed.

We also study high-dimensional dynamical systems with the existence of a heteroclinic network in the phase space under certain assumptions. This kind of dynamical systems can be used to describe the phenomenon of winnerless competition.
We study the heteroclinic network that is considered as a heteroclinic chain of heteroclinic cycles: Starting from each saddle point that is not in the last heteroclinic cycle, there are two heteroclinic orbits, one stays in its heteroclinic cycle and the other one goes to the next heteroclinic cycle; there is only one heteroclinic orbit starting from each saddle point in the last heteroclinic cycle. We prove under certain technical assumptions that for each collection of successive heteroclinic orbits inside this network, there is an open set of initial points such that any trajectory starting from each of them follows the prescribed collection of heteroclinic orbits and stays in a small neighborhood of it. We show also that the symbolic complexity function of the system restricted to this small neighborhood is a polynomial of degree $L - 1$, where $L$ is the number of heteroclinic cycles in the network.

Approximations of a high-dimensional tensor by sums of separable tensors can be used to avoid the curse of dimensionality in numerical computations. We study the approximation of a rank-2 symmetric tensor using a rank-1 tensor according to the gradient flow of the regularized error function. It had been previously known that at least one local minimum is attained in the symmetric set. We prove that the non-symmetric equilibria to the gradient flow are either saddle points or the maximum points. This implies that outside of the symmetric set there do not exist local minimum points. Furthermore, for these saddle points, we study the possibility of the so-called swamp behavior, where the trajectory stays in a neighborhood of the saddle point for a relatively long time before it leaves the neighborhood.
I dedicate this dissertation to my parents, Huiming Gong and Yajuan Lu, who made every effort to give me the best education.
ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Todd Young, for his invaluable help and guidance throughout my graduate study. I would not be able to complete this work without him. I am also grateful to my collaborators, Drs. Valentin Afraimovich, Erik Boczko, Richard Buckalew, Martin Mohlenkamp, Gregory Moses, Alexander Neiman, and Mikhail Rabinovich. They have generously helped me in several projects, which form the foundation of this dissertation. I would also like to thank Dr. Winfried Just who gave me many helpful suggestions and detailed comments to improve my work. I am thankful to all my professors and coworkers at Ohio University. I especially thank Drs. Sergiu Aizicovici, Winfried Just, Alexander Neiman, and Todd Young to be my committee members. I also thank Dr. Wei Lin who provided insight in the statistical aspects of my work.

Furthermore, I thank God my shepherd who has blessed me with the opportunities I have had in my life to achieve what I have accomplished today. I would like to thank my family, my fiancé, and his family for their support and love. I would also like to express my appreciation to the brothers and sisters in Christ who have encouraged me and prayed for me.

Finally, I would like to acknowledge the institutional support that my collaborators and I have received. The work in Chapter 2 was supported by the NIH-NIGMS grant R01GM090207; and the work in Chapter 4 is supported by the National Science Foundation under Grant No. 1418787.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>3</td>
</tr>
<tr>
<td>Dedication</td>
<td>5</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>6</td>
</tr>
<tr>
<td>List of Tables</td>
<td>9</td>
</tr>
<tr>
<td>List of Figures</td>
<td>10</td>
</tr>
<tr>
<td>List of Acronyms</td>
<td>13</td>
</tr>
<tr>
<td>1 Application of Dynamical Systems</td>
<td>14</td>
</tr>
<tr>
<td>2 Cell Cycle Coupling Models</td>
<td>16</td>
</tr>
<tr>
<td>2.1 Background</td>
<td>16</td>
</tr>
<tr>
<td>2.2 Responsive/signaling system with a gap</td>
<td>20</td>
</tr>
<tr>
<td>2.2.1 Dynamics of clusters via return maps</td>
<td>21</td>
</tr>
<tr>
<td>2.2.2 Dynamics for $k = 2$</td>
<td>25</td>
</tr>
<tr>
<td>2.2.3 Dynamics for $k = M + 1$ cyclic solutions</td>
<td>31</td>
</tr>
<tr>
<td>2.3 Noise-induced dispersion and breakup of 2-cluster stable solutions</td>
<td>38</td>
</tr>
<tr>
<td>2.3.1 Perturbed models and numerical integration</td>
<td>39</td>
</tr>
<tr>
<td>2.3.2 Statistical measures</td>
<td>44</td>
</tr>
<tr>
<td>2.3.3 Numerical simulation results</td>
<td>50</td>
</tr>
<tr>
<td>2.3.4 Yeast cultures have large coupling strengths</td>
<td>56</td>
</tr>
<tr>
<td>2.4 Conclusions</td>
<td>58</td>
</tr>
<tr>
<td>3 Feasibility of Heteroclinic Routes</td>
<td>60</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>60</td>
</tr>
<tr>
<td>3.2 Symbolic complexity of the subshift</td>
<td>65</td>
</tr>
<tr>
<td>3.3 The main theorem</td>
<td>67</td>
</tr>
<tr>
<td>3.4 Proof of the main theorem</td>
<td>70</td>
</tr>
<tr>
<td>3.4.1 Local maps</td>
<td>73</td>
</tr>
<tr>
<td>3.4.2 Global maps</td>
<td>77</td>
</tr>
<tr>
<td>3.4.3 Images and preimages</td>
<td>80</td>
</tr>
<tr>
<td>3.5 Discussion</td>
<td>82</td>
</tr>
<tr>
<td>3.5.1 An example of the system with the heteroclinic network $\Gamma$</td>
<td>82</td>
</tr>
<tr>
<td>3.5.2 A modified heteroclinic network $\hat{\Gamma}$</td>
<td>84</td>
</tr>
<tr>
<td>3.6 Summary</td>
<td>88</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table                                                                 | Page  
---                                                                 | ----  
3.1 Eigenvalues at the equilibrium point $Q^l_i$                       | 84    
3.2 Eigenvectors at the equilibrium point $Q^l_i$ corresponding to the eigenvalues in Table 3.1. | 84    
4.1 The existence of non-symmetric equilibrium points of the gradient flow in the form of $\alpha_1 = \cdots = \alpha_m \neq \alpha_{d+1} = \cdots = \alpha_d$ under different parameters $d, z$ and $m$. We see from this table that such equilibrium points do not exist when $d$ is even, $m$ is odd and $0 < z \leq 1$. When $d$ and $m$ are both even, such an equilibrium point might not exist for certain values of $z \in [-1, 0)$. For all other parameters values, at least one non-symmetric equilibrium point of the gradient flow exists. | 98    
4.2 Possibilities for the saddle value $\nu < 1$ according to different dimensions $d$. The functions $h_1$ and $h_4$ are given in Case 2 and Case 1 of Subsection 4.4.1 respectively. | 123   

List of Figures

Figure                               Page

2.1 An illustration of the map \( F \) with \( k = 3 \). Starting with \( x_1 = x_1(0) = 0 \), \( x_2 = x_2(0) \) and \( x_3 = x_3(0) \), we have \( F(x_2, x_3) = (x_1', x_2') \) where \( x_1' = x_1(t_1), x_2' = x_2(t_1) \) and \( t_1 \) is the shortest time such that \( x_3(t_1) = 1 \). Also, \( F(x_1', x_2') = (x_1'', x_2'') \) where \( x_3'' = x_3(t_2), x_1'' = x_1(t_2) \) and \( t_2 \) is the shortest time satisfying \( x_2(t_2) = 1 \). Similarly, \( F(x_2'', x_1'') = (x_2''' = x_2(t_3), x_3''' = x_3(t_3) \) and \( t_3 \) is the smallest time satisfying \( x_1(t_3) = 1 \). Because 0 and 1 coincide, we have \( x_1(t_3) = x_1(0) = 0 \). Therefore, at the Poincaré section \((x_1, x_2, x_3)|x_1 = 0\), the Poincaré map satisfies \( P = F^3 \) .................................................... 24

2.2 Return map \( F \) for (a) Case 1 when \( \alpha = 0.5, s = 0.2, r = 0.6 \), and \( \epsilon = 0.05 \), (b) Case 1 when \( \alpha = -0.5, s = 0.2, r = 0.6 \), and \( \epsilon = 0.05 \). (c) Case 2 when \( \alpha = 0.5, s = 0.4, r = 0.6 \), and \( \epsilon = 0.05 \) and (d) Case 2 when \( \alpha = -0.3, s = 0.4, r = 0.8 \), and \( \epsilon = 0.05 \) .................................................... 29

2.3 Regions of parameter space for the \( k \)-cyclic solutions with \( k = M + 1 \). Here the coupling strength is negative with \( \beta = -0.2 \) and the gap is \( \epsilon = 0.03 \). The widest diagonal band contains the parameters for \( k = 2 \) and is partitioned into six cases. Note that case VI disobeys the constraint \( \epsilon < r - s \). The rest of the diagonal bands contain parameters for \( k = 3, 4, \ldots \) and each of them is partitioned into five cases .................................................... 35

2.4 A snapshot of a two-cluster solution in the ensemble of \( n = 10^4 \) cells represented on a unit circle. The coupling function is \( f(I) = -0.5I \). The signaling region is \( S = [r_2, s] \) and the responsive region is \( R = [r_1, r_2] \) with \( r_1 = 0.3, r_2 = 0.65 \) and \( s = 0.95 \). Positions of the cells are indicated by grey circles and black triangles. Cells within clusters are spread because of Gaussian noise added to the system (the SDE model). This figure is published in [17] where we see that the previous \( R-S \) boundary \((0 = 1)\) in Section 2.2 is shifted to 0.65 .................................................... 40

2.5 Steady-state distributions of two clusters in the variable rate model of noise for the indicated values of \( \sigma \). The first cluster of cells is in black and the second one is in grey. The second cluster is plotted on top of the first cluster if there is an overlap between the two clusters. Insets show positions of cells on the circle with indicated signalling \((S = [r_2, s]) \) and responsive \((R = [r_1, r_2]) \) regions. When the two clusters overlap, only the first cluster in black is shown on the circle .................................................... 42

2.6 Time-dependent probability distributions \( p(t, x) \) of the ensemble with Gaussian white noise (SDE model) for the indicated values of noise level \( \sigma \). Logarithm of the probability distribution, \( \log[p(t, x)] \), is shown as filled contours .................................................... 49
2.7 Time-dependence of (a) the order parameter, $L_2(t)$, and of (b) the angular deviation, $D_c(t)$, for the indicated values of noise level $\sigma$ in the SDE model. For comparison, the corresponding dependencies $L_2(t)$ for the variable rate model are shown by grey lines on panel (a). The values of noise standard deviations are the same as those for the SDE model, except for the middle grey curve, where the noise standard deviation is 0.034. In both figures, a.u. stands for arbitrary unit.

2.8 Fraction of cells that have escaped from their initial clusters after 80 cycles for the three models. For noise levels below $\sigma \approx 0.01$ cells are effectively bound to their cluster by the asymptotic stability of the two-cluster solution.

2.9 Plots of (a) the order parameter $L_2$ and (b) the angular deviation of the first cluster $D_c$ of the steady state distribution, as a function of the noise level $\sigma$ in the three models.

2.10 Stationary statistical measures of the cell ensemble versus noise strength $\sigma$ for the three noise models. (a) Time averaged order parameter $\bar{L}_2$, vs noise standard deviation, $\sigma$, of the three models. (b) Time averaged angular deviation of the first cluster vs $\sigma$. Dotted line shows the maximum possible value of $\sqrt{2}$.

2.11 (a) Steady-state probability distributions $p(x)$ of the ensemble with Gaussian white noise (SDE model) versus the standard deviation $\sigma$ of the noise. Logarithm of steady state probability distribution $\log[p(x)]$ is shown as filled contours. (b) Time averaged normalized entropy $\bar{E}$ of the ensemble in three models versus noise SD.

2.12 Phase portraits of the SDE model projected onto the complex plane of the second Fourier harmonic. Black lines show steady state trajectories; grey lines show transients on both panels. Left: two-cluster ordered state, with a stable periodic orbit $\sigma = 0.03$. Right: stable equilibrium distribution at the disordered state $\sigma = 0.04$. The order-disorder transition occurs at $\sigma_c \approx 0.035$.

2.13 Time averaged order parameter $\bar{L}_2$ as a function of the coupling strength $\alpha$ and noise SD $\sigma$.

3.1 An illustration of the heteroclinic network when $L = 3$ and $N = 6$.

3.2 Scheme of the local behavior (a) and global behavior (b) of trajectories in a neighborhood of $\Gamma$.

3.3 The positive value of $y_1$ (point A) corresponds to the positive value of $x_{i+1}^l$. The negative value (point B) corresponds to the negative one. We only consider the non-negative values of the variables.

3.4 The projection of the domains of $T_{i1}$ and $T_{i2}$ onto the y-plane. For fixed $i = 1$ or 2, $D_1$ is the projection of the domain of $T_{i1}$ onto the y-plane and $D_2$ is the projection of the domain of $T_{i2}$ onto the y-plane. We see that $D_1$ and $D_2$ contain non-empty open sets and $D_1 \cap D_2 = \{0\}$. The figure looks the same for both $i = 1$ and $i = 2$.
3.5 Projections onto the y-plane of preimages of compositions of local and global maps. For fixed \(i = 1, 2, D_1\) in (a) is the projection of the preimage of \(T_{12}\) onto the y-plane and \(D_2\) in (b) is the projection of the preimage of \(T_{12}\) onto the y-plane as in Figure 3.4. In (a), \(D_{11} \subset D_1\) is the projection of the map \(\tilde{T}_{11} \circ T_{1l}^{gl} \circ T_{11}\) onto the y-plane and \(D_{12} \subset D_1\) is the projection of the map \(\tilde{T}_{12} \circ T_{1l}^{gl} \circ T_{12}\) onto the y-plane. In (b), \(D_{21} \subset D_2\) is the projection of the map \(\tilde{T}_{21} \circ T_2^{gl} \circ T_{12}\) onto the y-plane and \(D_{22} \subset D_2\) is the projection of the map \(\tilde{T}_{22} \circ T_2^{gl} \circ T_{12}\) onto the y-plane. We see that all sets contain nonempty open sets, \(D_{jk} \subset D_j\) and \(D_{jk} \cap D_{j'k} = \emptyset\) for \(j, k = 1, 2\). Both figures look the same for \(i = 1, 2\), but when \(i = 1\) these projections are projected from \(S_1^- (Q'_1)\) where \(\xi_1 = \delta\) and when \(i = 2\) these projections are projected from \(S_2^- (Q'_1)\) where \(\xi_2 = \delta\).

4.1 The three cases in the parameter space \(z \in [-1, 1] \setminus \{0\}\) and \(\phi \in (0, \pi/2]\) when \(d = 7\). The only region where \(\nu < 1\) is between the green and red curves. On the green curve \(\nu\) is undefined.

4.2 Saddle value \(\nu\) at each parameter pair \((-z)^{1/(d-2)}, \cos \phi\) for (a) \(d = 3\) and (b) \(d = 4\). The dark red region corresponds to where \(\nu\) is very large, the maximum of \(\nu\) in both figures is on the order of \(10^7\). The region where \(\nu < 1\) is under the black curves. We can see that in (a) the inequality \(\nu < 1\) holds in a small region where \(z > 0\) and \(\cos \phi\) is small as well as where \((-z)^{1/(d-2)} \in \left(\frac{1-\sin \phi}{\cos \phi}, 1\right]\). For (b), the inequality \(\nu < 1\) holds only when \((-z)^{1/(d-2)} \in \left(\frac{1-\sin \phi}{\cos \phi}, 1\right].\)

4.3 Saddle value \(\nu\) at each parameter pair \((-z)^{1/(d-2)}, \cos \phi\) for (a) \(d = 5\), (b) \(d = 6\), (c) \(d = 10\) and (d) \(d = 15\). The regions where \(\nu < 1\) are enclosed by black curves. (a) For \(d = 5\), \(\nu < 1\) occurs under all parameter values in Case 2. When \(z > 0\), \(\nu < 1\) cannot happen. (b) For \(d = 6\), only part of the parameter region \((\cos \phi \in (h_2(z), h_1(z)))\) in Case 2 yields \(\nu < 1\). (c) When \(d = 10\), the region \(\cos \phi \in (h_2(z), h_1(z))\) becomes smaller. (d) When \(d = 15\), \(\nu > 1\) in the region where \((-z)^{1/(d-2)} < 0\), so we did not plot this left half of the parameter space. We see only a narrow region \(\cos \phi \in (h_2(z), h_1(z))\) where \(\nu < 1\).

4.4 Values of \(\nu\) in the parameter space \(z \in [-1, 1] \setminus \{0\}\) and \(\phi \in (0, \pi/2]\) when (a) \(d = 5\), \(m = 2\) and (b) \(d = 9\), \(m = 4\). On the dark blue line at \(z = 0\), \(\nu\) is undefined. We see \(\nu > 1\) on the entire parameter space.

4.5 Values of \(\nu\) when \(d = 6\) and \(m = 3\) on the parameter space \(z \in [-1, 0)\) and \(\phi \in (0, \pi/2]\). We can see that there is no region where \(\nu < 1\).

4.6 Values of \(\nu\) for \(m = 2\) on the parameter space \(z \in [-1, 1] \setminus \{0\}\) and \(\phi \in (0, \pi/2]\) when (a) \(d = 4\) and (b) \(d = 6\). The dark blue region in \(z < 0\) is where the equilibrium point of the gradient flow does not exist. We see that there is no region in the parameter space where \(\nu < 1\). When \(d = 4\) there is a yellow region where \(\nu > 1\) and \(\nu\) is close to 1.
LIST OF ACRONYMS

CDC  Cell Division Cycle
YAO  Yeast Autonomous Oscillation
SDE  Stochastic Differential Equation
SD   Standard Deviation
K-S  Kolmogorov-Smirnov
WLC  Winnerless Competition
1 APPLICATION OF DYNAMICAL SYSTEMS

Dynamical systems theory is a powerful tool that has been used to study problems in various areas such as mathematics, biology, chemistry, physics, and finance. In this dissertation, we study the application of dynamical systems in both deterministic and stochastic models in yeast cell division cycles, in a heteroclinic network which can demonstrate a winnerless competition phenomenon, and to the existence of critical points in a specific tensor approximation problem.

Many of the phenomena that biologists are interested in are emergent properties of complex systems. Emergence in the theory of complex systems is a process whereby larger entities, patterns, and regularities arise through interactions among smaller or simpler entities that themselves do not exhibit such properties. One example of a complex system that shows interesting emergent behaviors is a large population of yeast cells (Saccharomyces cerevisiae) in a bioreactor. Biologists have long observed periodic oxygen consumption oscillations in yeast populations in a bioreactor and the periods of such oscillations are nearly integer divisors of the calculated period of a yeast cell cycle. Based on this observation, a cell cycle coupling model was proposed in [8, 43]. In this model, cells are considered to be coupled oscillators with the cell cycle of each cell as one oscillator. In Chapter 2, both the deterministic and stochastic aspects of this model are studied.

Winnerless competition is a phenomenon in a system where the prevalence among participants sequentially switches. For example, in ecological systems with predator-prey interactions, the population of one species might dominate the environment for a while then decreases, and another species would dominate in population for some time and then repeat this process. No species can dominate for an infinitely length of time and the switching between the temporary “winners” repeats endlessly. Similar patterns of activity have also been observed in collections of neurons in the human brain and in central
pattern generator in invertebrates. In Chapter 3, we study dynamical systems with a heteroclinic network under certain assumptions. Such a system can demonstrate the winnerless competition phenomenon. Some interesting mathematical results about such systems are proved in this chapter.

In numerical analysis, mathematicians apply the theory of dynamical systems to study problems such as the stability and the convergence rate of a numerical method. In this dissertation, we study problems in tensor approximations using dynamical systems. Numerical computations for functions of many variables may require huge computational cost and suffer from an effect called the *curse of dimensionality*. Therefore, it is essential to develop and understand methods to bypass this curse by representing multivariate functions as sums of separable functions. A related problem to approximating a high-dimensional function using sums of separable functions is approximating a high-dimensional tensor using sums of separable tensors. There are several algorithms for tensor approximations, but none of them works well for all approximations. One of the problems algorithms encounter is the *swamp* phenomenon. A swamp is an unpleasant behavior caused by the numerical approximation method. The swamp phenomenon occurs when the error function decreases extremely slowly for a great number of iterations before converging at a reasonable rate at the end. We apply dynamical systems theory to this problem by studying the gradient flow of the regularized error function in the approximations. In Chapter 4, we study the critical points for a specific tensor approximation problem and consider the possibility that the swamp phenomenon occurs.
2 Cell Cycle Coupling Models

The work in this chapter consists of two papers published in [16] and [17]. I am the first author of both of the papers. The models used in this chapter were introduced in [8, 43]. Section 2.1 is a combination of the introduction sections of papers [16] and [17]. I edited this section and added more details.

Section 2.2 is extracted from [16], which is joint work with Drs. Richard Buckalew, Todd Young and Erik Boczko. I wrote part of Subsection 2.2.1. Subsections 2.2.2 and 2.2.3 are my work. All figures in Section 2.2 were made by me. The work in Section 2.3 is joint work with Drs. Gregory Moses, Alexander Neiman and Todd Young. I wrote part of Subsections 2.3.1, 2.3.2 and 2.3.3. I did the calculations in Subsection 2.3.2. Drs. Young and Neiman proposed the usage of circular statistics and Shannon entropy respectively. I did the numerical simulations and plotted Figures 2.5, 2.8 and 2.9. Dr. Neiman made the rest of the figures and Dr. Young did the work in Subsection 2.3.4. Section 2.4 is a combination of the conclusion sections of papers [16] and [17].

2.1 Background

The modeling of the mitotic cell division cycle (CDC) in large ensembles of budding yeast cells has been studied for a few years by several authors in [8, 43]. This research is motivated by experimental studies of a phenomenon in bioreactor yeast cultures, known as yeast autonomous oscillations (YAO) [23, 29, 34, 40]. In autonomous oscillations, the metabolic signatures of large, well-mixed yeast cultures spontaneously begin to oscillate regularly without the presence of outside forcing [12]. A possible explanation of the link between YAO and CDC was proposed in [8] where the authors introduced the idea of cell cycle clustering—subpopulations or cohorts of cells traversing the CDC in near temporal synchrony. Note that this type of clustering occurs in time, not in space. The existence of
clusters was confirmed in oscillating yeast cultures by experiments reported in [8] and [38] and also is supported by other experimental evidence [36].

In [8] and [43], the authors introduced a class of generic models of cell cycle dynamics with coupling. They assumed that cells in a certain phase of the CDC, which is called the signaling region, may influence progression of cells in another phase of the CDC, which is called the responsive region, by producing some chemical agents. The following details of the model were proposed.

Consider a collection of \( n \) cells. The progression of the \( i \)-th cell with respect to the cell cycle is denoted by \( c_i \in [0, 1) \equiv S^1 \), where 0 corresponds to the time of birth of a cell and 1 corresponds to the time of division. We let 0 and 1 coincide in the model. Define the following specific regions in the cell cycle:

- the signaling region \( S = [\epsilon, s) \);
- and the responsive region \( R = [r, 1) \), where \( 0 \leq \epsilon < s < r < 1 \).

In [8] and [43], the authors assumed that there was no gap between \( R \) and \( S \), i.e., \( \epsilon = 0 \). The model without a gap will be referred to as the “immediate model”. The dynamics of the \( i \)-th cell is governed by the following system with coupling:

\[
\frac{dc_i}{dt} = 1 + a(c_i, \bar{c}), \quad 1 \leq i \leq n,
\]

where \( c_i \) is the progression of the \( i \)-th cell within the cycle and \( \bar{c} = (c_1, \ldots, c_n)^T \) denotes the state of all the cells in the culture. The term \( a(c_i, \bar{c}) \) allows for cells in the signaling region \( S \) to effect the progression of cells in the responsive region \( R \). There are a wide range of functions that can be used as \( a(c_i, \bar{c}) \), but in this chapter we study a simpler form of model:

\[
\frac{dc_i}{dt} = \begin{cases} 1, & \text{if } c_i \notin R, \\ 1 + f(I), & \text{if } c_i \in R, \end{cases}
\]
where
\[ I(\bar{c}) = \frac{\#\{j : c_j \in S\}}{n} \]  
(2.3)

is the proportion of cells in the signaling region \( S \). Note that when the cell \( c_i \) reaches 1, it will continue its trajectory at 0. We say that the coupling strength is a positive coupling if \( f(I) > 0 \) and a negative coupling if \( f(I) < 0 \). We call this an RS-coupling system because \( S \) follows \( R \).

There were some important results of the “immediated model” that have been established in [43] and [11]:

1. For positive coupling systems, isolated clustered solutions are stable. By isolated clustered solutions, we mean that clusters of cells do not influence the progression of each other throughout the whole cell cycle period. There is a stable synchronized solution, i.e., all cells may eventually merge into one cluster. On the other hand, for negative coupling systems, isolated clusters are unstable. Clusters must interact with each other to reinforce coherence.

2. The stability does not depend on the form of the coupling function \( f \). It depends on the size of the signaling and responsive regions, \(|S|\) and \(|R|\), where \(|S| = s - \epsilon = s\) (because \( \epsilon = 0 \) for the immediate model) and \(|R| = 1 - r\). For any negative coupling strength and any parameter values \(|S|\) and \(|R|\), there seems to be at least one stable \( k \)-clustered solution \((k \geq 2)\). On the other hand, for positive coupling systems, the \( k \)-clustered solutions \((k \geq 2)\) are not asymptotically stable for any parameter values.

3. Define \( M = \lfloor(|S| + |R|)^{-1}\rfloor \), the maximum number of clusters that can simultaneously exist without interacting with each other. For \( k = M + 1 \) clusters, the \( k \)-cyclic periodic solutions are the solutions for which after a certain time each cluster moves to the position of the cluster ahead of it. The authors found that under some
restrictions on the parameters, the $k$-cyclic periodic solutions are unstable in positive coupling systems and stable in negative coupling systems.

To make the cell cycle model in (2.2) and (2.3) more realistic, we proposed the inclusion of a gap between the responsive and signaling regions with the gap of length $\epsilon > 0$. This can mimic a time delay between the production of chemical signaling agents and the exertion of the feedback by these chemical agents.

For the modeling assumptions to be quantitatively realistic, there needs to be some estimate of the size of the gap. One of the metabolites that is known to have effects on cell cycle progression and is thought to be the primary signaling agent in glycolytic oscillations is acetaldehyde [26]. The diffusion rate of acetaldehyde across a yeast cell membrane is estimated to be $d = 300 \text{ s}^{-1}$ [26]. This rate is quite large. Furthermore, the diffusion/convection in the well-mixed bioreactor is even faster. The decay rate for acetaldehyde in the same yeast bioreactor study was estimated to be $\tau = 7.3 \text{ min}^{-1} - 9.9 \text{ min}^{-1}$ [26]. In that experiment, the yeast were performing fast glycolytic oscillations with a period of 37 s, so the production rate of acetaldehyde would need to be on the same order of magnitude as the decay rate. This implies that the time scale of a delay is on the order of $3 \times 10^{-4}$ in the normalized, non-dimensional coordinates. Using the period 37 s as an upper bound of the time scale of acetaldehyde production and diffusion would give us an upper bound on the gap of 0.0025 in normalized, non-dimensional coordinates. It is not known with certainty that acetaldehyde is the only or even the dominant signaling agent in the oscillations where clustering has been observed. Thus at this point these estimates are only speculative. Fortunately, as we shall see, we are able to analytically treat a broader range of gaps.

Several studies of cell cycle dynamics of large ensembles of cells without coupling have concluded that diffusive or dispersive forces, by various mechanisms, lead to asymptotic stability of “steady-state” dynamics in which the age structure profile or
distribution of cells within the cell cycle becomes stationary (although cells are still progressing through the cycle) [13, 18].

Meanwhile, our models with a general class of coupling mechanisms yield clustering of cells, which can explain the experimental results in [8, 38] that clustering exists in a continuous yeast culture undergoing the phenomenon of autonomous oscillations.

Thus, there are two competing phenomena at work in ensembles of cell-cycle oscillators, coherence and dispersion. We study the relationship between these two tendencies in the cell cycle coupling model. We consider the two biological mechanisms for dispersion that are known in budding yeast, namely variable growth rate and asymmetric division [42] and contrast them with standard additive Gaussian perturbations in terms of the effect on solutions. We then study the effects of random perturbations on the two-cluster periodic solutions of the cell-cycle model.

2.2 Responsive/signaling system with a gap

The following is an RS-coupling system with a gap [16]:

**Definition 2.2.1.** Consider $n$ cells whose coordinates are given by $c_k \in [0, 1)$. The progression of each cell $c_k$ is governed by Equation (2.1). We call such a system an RS-coupling system with $\epsilon$ gap if:

(H1) $R$ is an interval that precedes another interval $S$ with a small gap of length $\epsilon$, i.e., the distance between the last endpoint of $R$ and the first endpoint of $S$ is $\epsilon$;

(H2) When there is no cell in $S$, the $i$-th cell in $R$ is not experiencing feedback and its rate of progression is a constant $1$. That is to say, $a(c_i, \bar{c}) = 0$ except when $c_i \in R$ and there are some $c_j$ in $S$;
(H3) The rate of progression of each cell is positive and bounded. So there exist positive rates $v_{\text{min}}$ and $v_{\text{max}}$ such that $0 < v_{\text{min}} \leq 1 + a(c_i, \bar{c}) \leq v_{\text{max}}$ for all $c_i$ and $I$. Note that we do not allow cell cycle arrest, which is a common biological state;

(H4) The coupling strength is monotone. It must increase in the absolute value as the number of cells in $S$ increases. Adding a cell to $S$ will increase the coupling strength $|a(c_i, \bar{c})|$ which each cell $c_i \in R$ are experiencing; and

(H5) $a(c_i, \bar{c})$ is a smooth function for $c_i$ in the interior of $R$ and each $c_j$ in the interior of $S$, $j \neq i$ and the one-sided derivatives exist at the boundaries of $R$ and $S$.

Note that in this chapter, we consider a specific form of the model given by Equations (2.2) and (2.3). The hypotheses (H1)-(H5) hold for the specific model. The coupling strength is governed by the value of the function $f(I)$. In the rest of this chapter, positive coupling means that $f(I)$ is positive for $c_i \in R$ when at least one $c_j$ is in $S$. We define negative coupling analogously.

It has been shown in [43] that hypotheses (H3) and (H5) (a technical assumption) are sufficient to imply existence and uniqueness of solutions of (2.2).

Recall that we specify the coordinates throughout this section as follows:

\[ S = [\epsilon, s] \quad \text{and} \quad R = [r, 1), \quad 0 < \epsilon < s < r < 1. \]

The final endpoint of $R$ is 1, which corresponds to 0, the initial endpoint of the $\epsilon$ gap $[0, \epsilon)$.

2.2.1 Dynamics of clusters via return maps

In the remainder of this section we consider the existence and stability of periodic “clustered” solutions. Because the cells are assumed to be identical and obey (2.2), cells that initially share the same coordinate will remain synchronized as time evolves. By a cluster, we formally mean a group of cells that have identical coordinates in the CDC.
Because clustered solutions will remain clustered, one may consider such solutions using the coordinates of the clusters rather than those of individual cells. The dimensions of the system can be greatly reduced by doing so. In a yeast bioreactor culture, \( n \) can be on the order of \( 10^{10} \), while we can study clustered solutions with a small number of clusters.

The equation of progression for \( k \) clusters consisting of an equal number of cells is:

\[
\frac{dx_j}{dt} = \begin{cases} 
1, & \text{if } x_j \notin R, \\
1 + f(I), & \text{if } x_j \in R, 
\end{cases} \quad 1 \leq j \leq k, \tag{2.4}
\]

where \( I \) is the fraction of clusters in the signaling region: \( I \equiv \#\{j : x_j \in S\}/k \). Note that the coupling function \( f(I) \) in Equation (2.4) may be different from that in Equation (2.2). In Equation (2.4) the cells in one cluster remains clustered at all times, so \( f \) depends on the fraction of clusters that are in the signaling region instead of the fraction of cells.

However, it is proved in [27] on page 24 that for the system of \( k \) clusters where each cluster has \( n/k \) cells the coupling functions \( f(I) \) in (2.4) and (2.2) are the same. So we can consider the clustered subspace \( \{(x_1, \ldots, x_k) : x_i \in [0, 1)\} \) instead of the full space in the rest of this section.

The extreme case of clustering is when all cells have identical coordinates; then we have a single cluster and the system (2.4) becomes a single equation. The system will be trivial since there will be only one equation and there will never be coupling strength acting on the cluster. The only solution is a periodic solution that progresses at velocity 1 around the circle (i.e., \( x_1(t) = x_1(0) + t \mod 1 \) for all \( t > 0 \)). We refer to this special clustered solution as synchronized.

Note in [8] that a synchronized solution is locally asymptotically stable for positive coupling and unstable for negative coupling in the general model (2.2). This is because when a group of cells that are nearly synchronized leaves \( R \) and enters \( S \), the first cell in the group exerts feedback on the trailing cells. For positive coupling, the trailing cells in \( R \)
will speed up. The cells will eventually approach synchronization. For negative coupling, cells in \( R \) will slow down instead, thus moving further from the synchronized state.

In contrast, with a gap between the responsive region \( R \) and the signalling region \( S \), the above situation will not happen. If a solution consists of all cells within \( \epsilon \) of each other, then when this group of cells crosses from \( R \) to \( S \), all of the cells will be out of \( R \) before the first one enters \( S \). Hence, no cells in the solution experience feedback and the following proposition is obtained in [16].

**Proposition 2.2.2.** For the most general RS-coupling with a gap (2.2) and Definition 2.2.1 with either positive or negative coupling, the synchronized periodic solution is (locally) neutrally stable.

It is natural to ask whether or not the inclusion of a gap changes the global dynamics of the system. As discussed in [43], we can assume that all coordinates \( x_j(0) \) of the \( k \) clusters are initially ordered as

\[
0 = x_1(0) \leq \cdots \leq x_j(0) \leq \cdots \leq x_k(0) < 1.
\]

This ordering on the circle is preserved under the dynamics and the coordinate of the first cluster \( x_1 \) must eventually reach 1, i.e., there exists \( t_R \) such that \( x_1(t_R) = 1 \). Thus, the set \( \{(x_1, \ldots, x_k) | x_1 = 0\} \) defines a Poincaré section for the dynamics and the mapping

\[
P : (x_2(0), x_3(0), \ldots, x_k(0)) \mapsto (x_2(t_R), x_3(t_R), \ldots, x_k(t_R))
\]

defines the corresponding return map.

Starting from \( t = 0 \), we compute the time \( t_1 \) that the last cluster \( x_k \) needs to reach 1 and the coordinates of the remaining clusters at this time. As in [43] we can define a map \( F \) by

\[
F : (x_2(0), x_3(0), \ldots, x_k(0)) \mapsto (x_1(t_1), x_2(t_1), \ldots, x_{k-1}(t_1)).
\]
Notice that \( x_1(t_1) = t_1 \) because \( x_1 \) cannot be experiencing the feedback exerted from other clusters during this time according to the position of the signaling and responsive regions.

If the clusters are identical to each other, then the map \( F \) is a root of the Poincaré map, i.e., \( F^k = P \) and it can be regarded as a continuous, piecewise smooth map of the \((k - 1)\)-dimensional simplex

\[
0 \leq x_2 \leq x_3 \leq \cdots \leq x_k \leq 1
\]

into itself. See Figure 2.1.

![Figure 2.1: An illustration of the map \( F \) with \( k = 3 \). Starting with \( x_1 = x_1(0) = 0 \), \( x_2 = x_2(0) \) and \( x_3 = x_3(0) \), we have \( F(x_2, x_3) = (x'_1, x'_2) \) where \( x'_1 = x_1(t_1) \), \( x'_2 = x_2(t_1) \) and \( t_1 \) is the shortest time such that \( x_3(t_1) = 1 \). Also, \( F(x'_1, x'_2) = (x''_3, x''_1) \) where \( x''_3 = x_3(t_2) \), \( x''_1 = x_1(t_2) \) and \( t_2 \) is the shortest time satisfying \( x_2(t_2) = 1 \). Similarly, \( F(x''_3, x''_1) = (x'''_2, x'''_3) \) where \( x'''_2 = x_2(t_3) \), \( x'''_3 = x_3(t_3) \) and \( t_3 \) is the smallest time satisfying \( x_1(t_3) = 1 \). Because 0 and 1 coincide, we have \( x_1(t_3) = x_1(0) = 0 \). Therefore, at the Poincaré section \( \{(x_1, x_2, x_3)|x_1 = 0\} \), the Poincaré map satisfies \( P = F^3 \).

We concentrate on studying special periodic clustered solutions that satisfy:

\[
x_j(d) = x_{j+1}(0) \quad \text{for all} \quad j = 1, \ldots, k - 1, \quad \text{and} \quad x_k(d) = x_1(0) \mod 1, \tag{2.5}
\]

where \( d \) is some positive real number. We will refer to such solutions as \( k\)-cyclic clustered solutions. If \( k \) is a divisor of \( n \), then a \( k\)-cyclic solution consisting of \( n/k \) cells in each cluster exists. This fact can be proved using the Brouwer fixed point theorem applied to \( F \) [43]. As pointed out in [11], clustered solutions become less important as \( k \) grows larger.

In the remainder of this section we will let \( k \) denote the number of clusters and will suppose that each cluster consists of \( n/k \) cells.
2.2.2 Dynamics for \(k = 2\)

In order to compute and analyze the map \(F\) for two equal clusters in the model (2.4), we start with \(x_1(0) = 0\) and \(x_2(0) = x_2\) and integrate until \(x_2(t)\) reaches 1. So the calculations of \(F\) are reduced to solving:

\[
\frac{dx_2(t)}{dt} = \begin{cases} 
1, & x_2 \notin R \\
1 + f(I), & x_2 \in R, x_1 \in S,
\end{cases}
\]

where \(I(\bar{x}) = 0\) or \(1/2\), because there can be at most one cluster in the signaling region \(S\) when another cluster is in the response region \(R\). Let \(\alpha = f(1/2)\). Note that in the definition of the map \(F\), \(t_1\) is defined as the time that the last cluster needs to reach 1. Therefore, throughout the calculations for \(k = 2\), we let \(t_1\) be the time that \(x_2\) reaches 1, i.e., \(x_2(t_1) = 1\). The map \(F\) satisfies \(F(x_2) = x_1(t_1)\). Because \(x_1\) starts at 0 and it cannot experience feedback from \(x_2\) during the time interval \([0, t_1]\), we have \(x_1(t_1) = t_1\). Thus the map \(F\) satisfies \(F(x_2) = t_1\). The calculation of the map \(F\) requires us to find the time \(t_1\) that \(x_2(t_1) = 1\).

We will see that the problem can be analyzed by dividing it into the following two cases:

- **Case 1**: \((1 + \alpha)(s - \epsilon) < 1 - r\), and
- **Case 2**: \((1 + \alpha)(s - \epsilon) \geq 1 - r\).

For Case 1 \(((1 + \alpha)(s - \epsilon) < 1 - r)\), there are five situations depending on the location of \(x_2\):

- **a)** \(0 \leq x_2 \leq r - s\). In this case, \(x_1\) leaves \(S\) before \(x_2\) enters \(R\), thus \(x_2\) is not submitted to any feedback and \(x_2(t) = x_2 + t\) for all \(t\) in \([0, 1 - x_2]\). Hence, \(F(x_2) = t_1 = 1 - x_2\).

- **b)** \(r - s < x_2 \leq r - \epsilon\). Because \(x_2\) enters the responsive region \([r, 1]\) after \(x_1\) gets into \([0, \epsilon]\) and before \(x_1\) leaves \([\epsilon, s]\), \(x_2\) is influenced by \(x_1\) for some part of the time...
when it is in the responsive region. We can obtain the following piecewise-defined function for the position of $x_2$ depending on time:

$$x_2(t) = \begin{cases} x_2 + t, & 0 \leq t \leq r - x_2 \\ r + (1 + \alpha)(t - r + x_2), & r - x_2 < t < s \\ r + (1 + \alpha)(s - r + x_2) + t - s, & s \leq t \leq t_1, \end{cases}$$

where $x_2(t_1) = 1$, so we find $t_1 = 1 - (1 + \alpha)x_2 + \alpha(r - s)$. Hence,

$$F(x_2) = 1 - (1 + \alpha)x_2 + \alpha(r - s).$$

c) $r - \epsilon < x_2 \leq 1 - (1 + \alpha)s + \alpha \epsilon$. In this case, $x_2$ is influenced by $x_1$ for the entire time when $x_1$ is in the signaling region. This is because $x_1$ arrives at $S$ after $x_2$ reaches $r$ and $x_1$ gets out of $S$ before $x_2$ reaches 1. We can obtain:

$$x_2(t) = \begin{cases} x_2 + t, & 0 \leq t \leq \epsilon \\ x_2 + \epsilon + (1 + \alpha)(t - \epsilon), & \epsilon < t \leq s \\ x_2 + t - s + \epsilon + (1 + \alpha)(s - \epsilon), & s < t \leq t_1 = 1 - x_2 - \alpha(s - \epsilon). \end{cases}$$

Then $F(x_2) = 1 - x_2 - \alpha(s - \epsilon)$.

d) $1 - (1 + \alpha)s + \alpha \epsilon < x_2 \leq 1 - \epsilon$. In this case, $x_2$ reaches $r$ before $x_1$ gets into $S$, and when $x_2$ reaches 1, $x_1$ is still in $S$. We can find the position of $x_2$:

$$x_2(t) = \begin{cases} x_2 + t, & 0 \leq t \leq \epsilon \\ x_2 + \epsilon + (1 + \alpha)(t - \epsilon), & \epsilon < t \leq t_1. \end{cases}$$

We can find that $t_1 = \frac{1-x_2 + \alpha \epsilon}{1+\alpha}$ by solving $x_2(t_1) = 1$. Thus $F(x_2) = \frac{1-x_2 + \alpha \epsilon}{1+\alpha}$.

e) $1 - \epsilon < x_2 \leq 1$. In this case, $x_1$ does not affect $x_2$ because $x_2$ reaches 1 before $x_1$ enters $S$. So $x_2(t) = x_2 + t$ for all $t > 0$ and before $x_2$ reaches 1. So solve $x_2(t_1) = 1$ for $t_1$, we have $t_1 = 1 - x_2$. Thus, $F(x_2) = 1 - x_2$. 
Combining these five sub-domains, we can obtain the map $F$ for Case 1:

\[
F(x_2) = \begin{cases} 
1 - x_2, & 0 \leq x_2 \leq r - s \\
1 - (1 + \alpha)x_2 + \alpha(r - s), & r - s < x_2 \leq r - \epsilon \\
1 - x_2 - \alpha(s - \epsilon), & r - \epsilon < x_2 \leq 1 - (1 + \alpha)s + \alpha \epsilon \\
\frac{1-x_2+\alpha \epsilon}{1+\alpha}, & 1 - (1 + \alpha)s + \alpha \epsilon < x_2 \leq 1 - \epsilon \\
1 - x_2, & 1 - \epsilon < x_2 \leq 1.
\end{cases}
\] (2.6)

For Case 2 \((1 + \alpha)(s - \epsilon) \geq 1 - r\), similar to the calculations above, we consider five subcases:

a) \(0 \leq x_2 \leq r - s\). It is the same as Case 1(a). We have \(F(x_2) = t_1 = 1 - x_2\).

b) \(r - s < x_2 \leq \frac{1+\alpha r}{1+\alpha} - s\). In this case, \(x_1\) reaches \(\epsilon\) first, then \(x_2\) enters \(R\), and \(x_1\) reaches \(s\) before \(x_2\) reaches 1. So \(x_2\) is partially influenced by \(x_1\). This yields the position of \(x_2\) expressed by the function:

\[
x_2(t) = \begin{cases} 
x_2 + t, & 0 \leq t \leq r - x_2 \\
r + (1 + \alpha)(t - r + x_2), & r - x_2 < t \leq s \\
r + (1 + \alpha)(s - r + x_2) + t - s, & s < t \leq t_1.
\end{cases}
\]

Hence, we can derive \(F(x_2) = t_1 = 1 - (1 + \alpha)x_2 + \alpha(r - s)\) by solving \(x_2(t_1) = 1\).

c) \(\frac{1+\alpha r}{1+\alpha} - s < x_2 \leq r - \epsilon\). Note that \(\frac{1+\alpha r}{1+\alpha} - s < r - s\) is implied by this case. Here \(x_2\) is influenced by \(x_1\) for the entire time when it is in the responsive region because \(x_2\) enters \(R\) after \(x_1\) gets into \(S\) and when \(x_2\) reaches 1, \(x_1\) is still in \(S\). We obtain the following piecewise-defined function for the position of \(x_2\):

\[
x_2(t) = \begin{cases} 
x_2 + t, & 0 \leq t \leq r - x_2 \\
r + (1 + \alpha)(t - r + x_2), & r - x_2 < t \leq t_1,
\end{cases}
\]

where \(t_1 = 1 - x_2 - \frac{\alpha}{1+\alpha} (1 - r)\). Hence, \(F(x_2) = 1 - x_2 - \frac{\alpha}{1+\alpha} (1 - r)\).
d) \( r - \epsilon < x_2 \leq 1 - \epsilon \). For this case \( x_2 \) enters \( R \) before \( x_1 \) reaches \( \epsilon \), and so it is influenced by \( x_1 \) for some part of the time when it is in the responsive region until it reaches 1.

\[
x_2(t) = \begin{cases} 
    x_2 + t, & 0 \leq t \leq \epsilon \\
    \epsilon + x_2 + (1 + \alpha)(t - \epsilon), & \epsilon < t \leq t_1 = \frac{1-x_2+\alpha \epsilon}{1+\alpha}
\end{cases}
\]

Thus \( F(x_2) = \frac{1-x_2+\alpha \epsilon}{1+\alpha} \).

e) \( 1 - \epsilon < x_2 \leq 1 \). This is the same as Case 1(e) that \( x_1 \) does not affect \( x_2 \). We obtain:

\[
x_2(t) = x_2 + t \text{ for all } t \in [0, 1 - x_2]. \text{ Thus } F(x_2) = 1 - x_2.
\]

We combine the five subcases and arrive at the map \( F \) for Case 2:

\[
F(x_2) = \begin{cases} 
    1 - x_2, & 0 \leq x_2 \leq r - s \\
    1 - (1 + \alpha)x_2 + \alpha(r - s), & r - s < x_2 \leq \frac{1+\alpha r}{1+\alpha} - s \\
    1 - x_2 - \frac{\alpha}{1+\alpha}(1 - r), & \frac{1+\alpha r}{1+\alpha} - s < x_2 \leq r - \epsilon \\
    \frac{1-x_2+\alpha \epsilon}{1+\alpha}, & r - \epsilon < x_2 \leq 1 - \epsilon \\
    1 - x_2, & 1 - \epsilon < x_2 \leq 1.
\end{cases}
\)

The graph of \( F \) coincides with the anti-diagonal \( y = 1 - x_2 \) for \( 0 \leq x_2 \leq r - s \) and \( 1 - \epsilon \leq x_2 \leq 1 \) (Figures 2.2). For positive coupling \( \alpha = f(\frac{1}{2}) > 0 \), it is strictly less than \( 1 - x_2 \) when \( r - s < x_2 < 1 - \epsilon \) (Figures 2.2(a) and (c)); while for negative coupling \( \alpha < 0 \), it is strictly greater than \( 1 - x_2 \) if \( r - s < x_2 < 1 - \epsilon \) (Figures 2.2(b) and (d)).

In the following discussion, we consider the restriction that \( \epsilon \) is relatively small, more precisely, \( \epsilon < r - s \). This is because when considering the unit circle with parameters \( 0 < \epsilon < s < r < 1 \) and 1 is identical to 1, between the signaling region \( S = [\epsilon, s) \) and the responsive region \( R = [r, 1) \), there are two intervals on the unit circle \( [0, \epsilon) \) and \( [s, r) \). Both of them are the gaps between \( S \) and \( R \). If \( \epsilon > r - s \), then the length of the gap from 0 to \( \epsilon \) is longer than that of the gap from \( s \) to \( r \). So \( [s, r) \) becomes the smaller gap between \( S \) and \( R \). When \( \epsilon > r - s \), we can change the time variable \( t \) to be negative and let the time go
Figure 2.2: Return map $F$ for (a) Case 1 when $\alpha = 0.5, s = 0.2, r = 0.6$, and $\epsilon = 0.05$, (b) Case 1 when $\alpha = -0.5, s = 0.2, r = 0.6$, and $\epsilon = 0.05$, (c) Case 2 when $\alpha = 0.5, s = 0.4, r = 0.6$, and $\epsilon = 0.05$ and (d) Case 2 when $\alpha = -0.3, s = 0.4, r = 0.8$, and $\epsilon = 0.05$.

“backwards” to obtain the similar results as when $\epsilon < r - s$. Therefore, we only consider $\epsilon < r - s$ in this section.

For Case 1 $((1 + \alpha)(s - \epsilon) < 1 - r)$, we can see that

$$1 - (1 + \alpha)s + \alpha \epsilon = 1 - (1 + \alpha)(s - \epsilon) - \epsilon > s.$$ 

Therefore, the graph of $F$ cannot intersect with the line $\{y = x_2\}$ at the line segments IV and V, or the points C and D. This is illustrated in Figures 2.2(a) and (b).
Similarly, for Case 2 \(((1 + \alpha)(s - \epsilon) \geq 1 - r)\), we can find that the graph of \(F\) and the diagonal cannot intersect at C, IV, D or V (see Figures 2.2(c) and (d)), because
\[
(1 + \alpha)(r - \epsilon) > (1 + \alpha)s \geq 1 - r + (1 + \alpha)\epsilon, \quad \text{i.e.,} \quad r - \epsilon > \frac{1 - r}{1 + \alpha} + \epsilon.
\]

Considering the situations above, we conclude the following results:

1. If \(r - s > \frac{1}{2}\), the intersection of the graph of \(F\) and the diagonal \(y = x_2\) lies in the first line segment \((0 < x_2 < r - s)\), denoted by I. Each point in the interval \(x_2 \in [1 - r + s, r - s]\) is part of a 2-periodic orbit \(F(x_2) = 1 - x_2\), and \(F^2(x_2) = x_2\), except for the point \(x_2 = \frac{1}{2}\) which is fixed. Hence, the return map \(P = F^2\) has an interval of neutrally stable fixed points centered around \(x_2 = \frac{1}{2}\).

2. If \(r - s = \frac{1}{2}\), the intersection, denoted by A is at \((\frac{1}{2}, \frac{1}{2})\). It is a unique fixed point that is stable for negative \(\alpha\) and unstable for positive \(\alpha\).

3. If \(r - s < \frac{1}{2}\), then \(\epsilon < \frac{1}{2}\), thus \(1 - \epsilon > \frac{1}{2}\). There are three possibilities depending on where the diagonal line \(y = x_2\) intersects the graph of \(F\). Also here, we need to consider Case 1 and Case 2 separately.

   (a) If the diagonal intersects the second line segment (II), the intersection lies in the interval \(x_2 \in (r - s, r - \epsilon)\) for Case 1 or \(x_2 \in (r - s, \frac{1 + \alpha r}{1 + \alpha} - s)\) for Case 2. Hence, there is a unique fixed point. It is stable if \(\alpha < 0\) and unstable if \(\alpha > 0\).

   (b) If the diagonal hits the boundary between segments II and III of \(F\) (point B as in the figures), there is a unique fixed point at \((r - \epsilon, 1 - r + \epsilon - \alpha(s + \epsilon))\) for Case 1 or at \((\frac{1 + \alpha r}{1 + \alpha} - s, s)\) for Case 2, which is stable for negative \(\alpha\) and unstable for positive \(\alpha\).

   (c) If the diagonal intersects the third segment (III) of the graph of \(F\), then the slope is -1. It follows that there is an interval of neutral period-two points.
This interval can be found to be \((r - \epsilon, 1 - (1 + \alpha)s + \alpha \epsilon)\) for Case 1 or \((\frac{1 + \alpha s}{1 + \alpha} - s, r - \epsilon)\) for Case 2. The edge of the interval (point C) is stable for negative \(\alpha\) and unstable for positive \(\alpha\).

Note that in all cases, the 2-cyclic solution corresponds to the unique fixed point of \(P = F^2\). According to the above results, we find four distinct types of dynamics; two for positive coupling and two for negative coupling.

- **Positive coupling:**
  
  (a) The 2-cyclic solution is unstable when the intersection is at II, A or B.
  
  (b) The 2-cyclic solution belongs to an interval of fixed points of period-two solutions when the intersection is at I or III. This interval is repelling.

- **Negative coupling:**
  
  (a) The 2-cyclic solution is stable when the intersection is at II, A or B.
  
  (b) The 2-cyclic solution belongs to an interval of fixed points of period two solutions when the intersection is at I or III. This interval is attracting.

These behaviors are exactly the same as for the model with no gaps in [43]. Note that in the parameter regions where the fixed point of the return map is either stable or unstable, the slopes of \(F\) are exactly the same as for the model without gaps (see [43]). This follows that adding a gap does not alter the strength of the stability or the instability of the 2-cyclic solutions.

### 2.2.3 Dynamics for \(k = M + 1\) cyclic solutions

We consider the case when \(\epsilon\) is small, i.e., \(\epsilon < r - s\), as in the former subsection. The following definition will play a large role in the analysis in this subsection.
Definition 2.2.3. Define

\[ M \equiv \lfloor (|R| + |S| + \epsilon)^{-1} \rfloor \equiv \lfloor (1 - r + s)^{-1} \rfloor, \]  
(2.8)

i.e., \( M \) is the maximum number of isolated clusters that can simultaneously exist.

We consider the cyclic solutions consisting of \( k = M + 1 \) clusters. For \( k = M + 1 \), it can be shown that the cyclic \( k \)-cluster solution has initial conditions:

\[ x_1 = 0, x_2 = d, x_3 = 2d, \ldots, x_k = (k - 1)d. \]

During each time interval \( d \), each cluster will move to the position of the cluster ahead of it.

For \( k = M + 1 \) and for all solutions in a neighborhood of the cyclic solution above, at most one cluster can be in the signaling region \( S \) at a time. Thus, the feedback experienced by any cluster in \( R \) is either 0 or \( f(1/k) \). Hereafter, when studying systems with \( k = M + 1 \) clusters, we will use the abbreviation:

\[ \beta \equiv f(1/k). \]

The evolution of a solution can be described in terms of the order of events of the solution. Clusters of cells progress through the cell cycle at rates specified by the Equation (2.4). These rates remain constant until a cluster reaches one of the milestones \( \epsilon \), \( s \), \( r \), or 1. Given the next such event and the index \( j \) of the cluster for which it occurs, we can calculate the time elapsed until the event occurs by finding out the initial position of cluster \( j \) and its corresponding rate during that time. We will use \( x_j \mapsto \mu \) to denote the event that cluster \( j \) reaches the milestone \( \mu \). Then for the order of events, for example, as in Case I below the notation \( x_1 \mapsto \epsilon, x_k \mapsto r \) means that first \( x_1 \) reaches \( \epsilon \), then \( x_k \) reaches \( r \).

We discuss the following possibilities of the orders of events for the \( k \)-cyclic solutions. In the following calculations for the \( k \)-cyclic solutions, we let \( t_1 \) be the time that the last cluster \( x_k \) needs to reach 1. So \( t_1 \) satisfies \( x_k(t_1) = 1 \). The map \( F \) in this situation is defined as \( F(x_2, \ldots, x_k) = (x_1(t_1), \ldots, x_{k-1}(t_1)) \) and the return map \( P = F^k \).
Case I: \( x_2 = d > s \) and \( x_1 \mapsto \epsilon, x_k \mapsto r, x_1 \mapsto s, x_k \mapsto 1 \).

By calculating the position of \( x_k \) in each step, we have

\[
x_k(t) = \begin{cases} 
(k-1)d + t, & 0 \leq t \leq r - (k-1)d \\
 r + (1+\beta)(t-r + (k-1)d), & r - (k-1)d < t \leq s \\
 r + (1+\beta)(s-r + (k-1)d) + t - s, & s < t \leq t_1.
\end{cases}
\]

Here, \( x_k(t_1) = 1 \). So \( t_1 = 1 + \beta(r - s) - (1+\beta)(k-1)d \) where \( \beta = f(1/k) \). At time \( t_1 \), \( x_1 \) reaches \( x_2 = d \) by definition of the cyclic solutions. Since \( x_1 \) has velocity 1, \( t_1 = d \). Thus,

\[
1 + \beta(r - s) - (1+\beta)(k-1)d = d.
\]

One can solve for \( d \):

\[
d = \frac{1 + \beta(r - s)}{k + \beta(k-1)}. \tag{2.9}
\]

This case can only happen when \( d > s \) and \( x_k = (k-1)d < r - \epsilon \). Using (2.9) in these two inequalities gives conditions on \( r \) and \( s \) required by this order of events:

\[
s < \frac{1+\beta r}{k(1+\beta)} \quad \text{and} \quad r > \frac{k-1}{k} [1 - \beta(s - \epsilon)] + \epsilon. \tag{2.10}
\]

Case II: \( x_2 = d > s \) and \( x_k \mapsto r, x_1 \mapsto \epsilon, x_1 \mapsto s, x_k \mapsto 1 \).

Similar to Case I, we may calculate:

\[
d = \frac{1 - \beta(s - \epsilon)}{k}. \tag{2.11}
\]

The above \( d \) also needs to satisfy \( d > s \) and \( r > x_k > r - \epsilon \), so we obtain:

\[
s < \frac{1 + \beta \epsilon}{k + \beta} \quad \text{and} \quad \frac{k-1}{k} [1 - \beta(s - \epsilon)] < r < \frac{k-1}{k} [1 - \beta(s - \epsilon)] + \epsilon. \tag{2.12}
\]

Case III: \( x_2 = d > s \) and \( x_1 \mapsto \epsilon, x_1 \mapsto s, x_k \mapsto 1 \).

This order of events can happen only when \( x_k > r \). We can find that \( d \) is the same as that in (2.11), \( d = (1 - \beta(s - \epsilon))/k \), and the order of events in this case implies the inequalities:

\[
s < \frac{1 + \beta \epsilon}{k + \beta} \quad \text{and} \quad r < \frac{k-1}{k} [1 - \beta(s - \epsilon)]. \tag{2.13}
\]
Case IV: $x_2 = d < s$ and $x_1 \mapsto \epsilon, x_2 \mapsto s, x_k \mapsto r, x_k \mapsto 1$.

After calculating the final position of $x_1$, we find that

$$d = \frac{1 + \beta r}{k(1 + \beta)}.$$  \hfill (2.14)

This case occurs under the conditions that $d < s - \epsilon$ and $x_k < r$. Using (2.14) with the order of events gives that

$$s > \frac{1 + \beta r}{k(1 + \beta)} + \epsilon.$$  \hfill (2.15)

Case V: $s - \epsilon < d < s$ and $x_2 \mapsto s, x_1 \mapsto \epsilon, x_k \mapsto r, x_k \mapsto 1$.

After a similar calculation, we find the same $d$ as in (2.14); that is $d = \frac{1 + \beta r}{k(1 + \beta)}$.

Applying this to $s - \epsilon < d < s$ and $x_k < r - \epsilon$, we obtain the following necessary and sufficient conditions for Case V to occur:

$$\frac{1 + \beta r}{k(1 + \beta)} < s < \frac{1 + \beta r}{k(1 + \beta)} + \epsilon \quad \text{and} \quad r > \frac{ek(\beta + 1) + k - 1}{k + \beta}.$$  \hfill (2.16)

Case VI: $x_2 \mapsto s, x_k \mapsto r, x_1 \mapsto \epsilon, x_k \mapsto 1$.

This case occurs if, and only if, $d < s$ and $r > x_k > r - \epsilon$. Solving for $d$, we find that

$$d = \frac{1 + \epsilon \beta}{k + \beta}.$$  \hfill (2.17)

We substitute (2.17) into the above inequalities to obtain:

$$s > \frac{1 + \epsilon \beta}{k + \beta} \quad \text{and} \quad \frac{ek(\beta + 1) + k - 1}{k + \beta} - \epsilon < r < \frac{ek(\beta + 1) + k - 1}{k + \beta}.$$  \hfill (2.18)

The regions described by the inequalities in (2.10) to (2.18) are illustrated in Figure 2.3. We can observe that cases I to VI exhaust the parameter set in $(r, s)$ and there is never more than one cluster in the signaling region when $R$ is non-empty. Thus the dynamics of the system are determined by the coupling strength $\beta = f(\frac{1}{k})$.

Now we calculate the map $F$. We note that $F$ is affine in a neighborhood of the cyclic solution provided that the parameters are in the interior of one of the cases. Thus, we have
Figure 2.3: Regions of parameter space for the $k$-cyclic solutions with $k = M + 1$. Here the coupling strength is negative with $\beta = -0.2$ and the gap is $\epsilon = 0.03$. The widest diagonal band contains the parameters for $k = 2$ and is partitioned into six cases. Note that case VI disobeys the constraint $\epsilon < r - s$. The rest of the diagonal bands contain parameters for $k = 3, 4, \ldots$ and each of them is partitioned into five cases.

$F : \tilde{x} \mapsto A\tilde{x} + \tilde{b}$ where $\tilde{x} = (x_2, \ldots, x_k)^T$ where $A$ is a $(k - 1) \times (k - 1)$ matrix. To find the stability of the $k$-cyclic solutions, the linear part of the system, i.e., the matrix $A$ needs to be analyzed.

In Case I, we can calculate the time for $x_k$ to reach 1, that is $t_1 = 1 + \beta(r - s) - (1 + \beta)x_k$. Since each of the clusters moves for time $t_1$, we can find $x_j(t_1) = x_j + t_1 = x_j + 1 + \beta(r - s) - (1 + \beta)x_k$ for $j = 1, \ldots, k - 1$. Given that $x_1 = 0$ and $x_k(t_1) = 1$, we have $F(x_1) = 1 + \beta(r - s) - (1 + \beta)x_k$ and $F(x_k) = 1$. Thus we can obtain the
matrix $A$ by taking the partial derivatives of the map $F((x_2, \ldots, x_k)^T)$:

$$A = \begin{bmatrix}
0 & 0 & 0 & \cdots & 0 & -(1 + \beta) \\
1 & 0 & 0 & \cdots & 0 & -(1 + \beta) \\
0 & 1 & 0 & \cdots & 0 & -(1 + \beta) \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & -(1 + \beta) \\
0 & 0 & 0 & \cdots & 1 & -(1 + \beta)
\end{bmatrix}. \quad (2.19)$$

In [43], it was proved that all the eigenvalues of $A$ defined by (2.19) lie outside of the unit disc when $\beta > 0$ and lie on the interior of the unit disc when $\beta < 0$. Thus, the fixed point of the map $F$ in case I is stable for negative coupling and unstable for positive coupling, so is the $k$-cyclic solution.

In Case II and Case III, following the same procedures as above, we can calculate that $t_1 = 1 - \beta(s - \epsilon) - x_k$. Thus $x_j(t_1) = x_j + 1 - \beta(s - \epsilon) - x_k$ for $j = 2, \ldots, k - 1$, $x_1(t_1) = 1 - \beta(s - \epsilon) - x_k$, and $x_k(t_1) = 1$. Therefore, the linear part of the map at the fixed point is represented by the matrix:

$$A = \begin{bmatrix}
0 & 0 & 0 & \cdots & 0 & -1 \\
1 & 0 & 0 & \cdots & 0 & -1 \\
0 & 1 & 0 & \cdots & 0 & -1 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & -1 \\
0 & 0 & 0 & \cdots & 1 & -1
\end{bmatrix}. \quad (2.20)$$

In Case IV and Case V, we find that $t_1 = r + \frac{1 - r}{1 + \beta} - x_k$. Here the coefficient for the $x_k$ term is still $-1$, so the linear part of the map is the same as the matrix $A$ given by (2.20) in Case II and Case III. This matrix $A$ is known as the companion matrix of the polynomial $p(u) = u^{k-1} - u^{k-2} - \cdots - u - 1$ [20]. We know that all the eigenvalues of this matrix $A$,
which are the same as the zeros of the polynomial \( p \), have modulus 1. Thus the \( k \)-cyclic solution is neutrally stable in all the cases from II to V.

For Case VI, the time for \( x_k \) to arrive at 1 is \( t_1 = \frac{1+\varepsilon \beta}{1+\beta} - \frac{x_1}{1+\beta} \). Note that this case only occurs for \( M = 1 \) or \( k = 2 \) and \( \varepsilon > r - s \). For \( M > 1 \) and \( \varepsilon < r - s \), this region is outside the band \( k = M + 1 \). The matrix \( A \) can be derived as the following:

\[
A = \begin{bmatrix}
0 & 0 & 0 & \ldots & 0 & -\frac{1}{1+\beta} \\
1 & 0 & 0 & \ldots & 0 & -\frac{1}{1+\beta} \\
0 & 1 & 0 & \ldots & 0 & -\frac{1}{1+\beta} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 0 & -\frac{1}{1+\beta} \\
0 & 0 & 0 & \ldots & 1 & -\frac{1}{1+\beta}
\end{bmatrix}.
\]

We can use a change of variable and then compare with the result for (2.19) to study the stability of the fixed point of this map \( F \).

Let \( \frac{1}{1+\beta} = 1 + \gamma \), then \( \beta = \frac{1}{1+\gamma} - 1 \) and \( A \) is changed to

\[
A = \begin{bmatrix}
0 & 0 & 0 & \ldots & 0 & -(1+\gamma) \\
1 & 0 & 0 & \ldots & 0 & -(1+\gamma) \\
0 & 1 & 0 & \ldots & 0 & -(1+\gamma) \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 0 & -(1+\gamma) \\
0 & 0 & 0 & \ldots & 1 & -(1+\gamma)
\end{bmatrix}.
\]

We already know that all the eigenvalues \(|\lambda| > 1\) (the fixed point is unstable) if \( \gamma > 0 \), and \(|\lambda| < 1\) (the fixed point is stable) if \( \gamma < 0 \). When \( \gamma > 0 \), by the relation between \( \beta \) and \( \gamma \), we can see that \( \beta < 0 \). Otherwise, if \(-1 < \gamma < 0\), we have \( \beta > 0 \). So we can conclude that the fixed point is stable for positive coupling (\( \beta > 0 \)) and unstable for negative coupling (\( \beta < 0 \)) in this case.

The dynamics for the \( k = M + 1 \)-cyclic solutions can be summarized as follows:
(1) When \( s < \frac{1+\beta_\epsilon}{k(1+\beta)} \) and \( r > k^{-1} \left[ 1 - \beta(s - \epsilon) \right] + \epsilon \), the \( k \)-cyclic solution is stable for negative coupling and unstable for positive coupling.

(2) For \( k = 2 \) and \( \epsilon > r - s \), when \( s > \frac{1+\beta_\epsilon}{k\beta} \) and \( \frac{\epsilon(k+1)+k-1}{k\beta} - \epsilon < r < \frac{\epsilon(k+1)+k-1}{k\beta} \), the \( k \)-cyclic solution is stable for positive coupling and unstable for negative coupling.

(3) Otherwise, the \( k = M + 1 \)-cyclic solution is neutrally stable.

The behavior in (2) occurs only on a small part of parameter space with \( \epsilon > r - s \), i.e., Case VI in Figure 2.3.

The behavior in (1) and (3) are entirely consistent with those for the model without a gap [43]. Further, as noted above, the matrix \( A \) in both the stable and unstable cases is exactly the same as that for the model without a gap. Thus, the gap does not alter the strength of the stability or instability; the linearization of the map \( F \) is exactly the same.

By comparing regions in the parameter space for which the \( k = M + 1 \)-cyclic solutions are stable for the gap model and the immediate model, we note that the two regions are the same except that for the gap model we need to exclude a strip of width \( \epsilon \). This is because the behavior in (2) does not exist for the immediate model, which is represented by this strip of width \( \epsilon \).

2.3 Noise-induced dispersion and breakup of 2-cluster stable solutions

In this section, we will study the effect of random perturbation on the RS-coupling model of a large collection of cells. We set the number of cells to be \( n \). In bioreactor experiments \( n \) is on the order of \( 10^{10} \). In our simulations in this section, we use \( n = 10^4 \) unless otherwise indicated.

The responsive region \( R \) is possibly part of the G1 phase of the cell cycle and the S region is postulated to be the budded portion of the cell cycle, i.e. the literal S-phase, the G2 phase and part of the M-phase. From experiments on autonomous oscillation with the
yeast strain CEN.PK, the authors of [38] found that the G1 phase occupies about 65% of the total cell cycle. In our simulations below we will set the $R$-$S$ boundary to be at 0.65, with $R = (0.3, 0.65)$ and $S = (0.65, 0.95)$. This gives parameters that are in the interior of the set in the parameter space for which a two-cluster solution is stable for the model (2.2) with negative coupling as calculated in [43]. The signaling and responsive regions $R$ and $S$ are illustrated in Figure 2.4. Note that in the previous section, the cell cycle is conceptualized as the unit circle $S^1 \equiv [0, 1)$ where 0 is the $R$-$S$ boundary. In this section, we let the $R$-$S$ boundary to be at 0.65 instead. This implies that we need to shift the coordinate by 0.65 in the clockwise direction. This shift does not affect our results on the stability of the two-cluster solutions provided that the sizes of $S$ and $R$ remain the same. Therefore, under negative coupling and certain sizes of $S$ and $R$, there exist asymptotically stable two-cluster solutions. In our simulations, we use $|S| = 0.3$ and $|R| = 0.35$.

Probing the biological details of the cell cycle is a topic of intense and fruitful study and many of the mechanisms involved in regulating the cell cycle within a cell are well-understood (see [28]). Our mathematical model does not take into account the details of these mechanisms directly, but because of its universal nature, it can shed light on the biology when paired with experimental data.

2.3.1 Perturbed models and numerical integration

In this subsection, we consider only two-cluster periodic stable solutions. We use three types of random dispersive models to study how the two-cluster periodic solutions eventually break up under large noise.

Physical variation and epigenetic differences among the isogenic cells may produce small variations in growth rate. In [5], several scenarios of this are discussed. One may model this by adding Gaussian white noise, which is described by a system of coupled stochastic differential equations (SDEs).
Figure 2.4: A snapshot of a two-cluster solution in the ensemble of \( n = 10^4 \) cells represented on a unit circle. The coupling function is \( f(I) = -0.5I \). The signaling region is \( S = (r_1, s) \) and the responsive region is \( R = (r_1, r_2) \) with \( r_1 = 0.3 \), \( r_2 = 0.65 \) and \( s = 0.95 \). Positions of the cells are indicated by grey circles and black triangles. Cells within clusters are spread because of Gaussian noise added to the system (the SDE model). This figure is published in [17] where we see that the previous \( R-S \) boundary \((0 = 1)\) in Section 2.2 is shifted to 0.65.

The SDE model reads,

\[
\frac{dx_k}{dt} = a(x_k, I) + \sigma N_k(t) = \begin{cases} 
1, & \text{if } x_k \notin R, \\
1 + f(I), & \text{if } x_k \in R,
\end{cases} + \sigma N_k(t), 
\]  

(2.22)

where \( N_k(t) \) are zero-mean Gaussian white noise sources, uncorrelated among different cells. So the covariance between the two random variables satisfy

\[
\text{Cov}(N_k(t), N_{k'}(t + \tau)) = \delta(k, k')\delta(\tau).
\]

We use the Euler-Maruyama method to numerically integrate (2.22):

\[
x_k^{j+1} = x_k^j + h a(x_k^j, I_j) + \sigma \sqrt{h}N_k^j,
\]
where each $N^j_k$ is a normally distributed pseudo-random number and $h$ is the integration time step.

The SDE model lacks strict biological basis. For instance, at extreme values of $N_k(t)$ the cell might grow at an unrealistic rate, which can cause this cell to skip one cell cycle or even “grow backwards”. However, SDE models are convenient because of the well-developed methods for studying them.

In another model we will consider a more biologically reasonable random perturbation representing the unavoidable small inhomogeneity among cells. When each cell begins its cycle (at division) it will be randomly and independently assigned a rate of progression that is a small and bounded perturbation from the mean rate of the ensemble. We call this the variable rate model. This model is governed by the following equation.

\[
\frac{dx_k}{dt} = \begin{cases} 
1 + U_k(t), & \text{if } x_k \notin R \\
1 + U_k(t) + f(I), & \text{if } x_k \in R,
\end{cases}
\tag{2.23}
\]

where $U_k(t)$ is a uniformly distributed random variable, $U_k(t) \in [-\sqrt{3}\sigma, \sqrt{3}\sigma]$, with standard deviation (SD) $\sigma$. The variables $\{U_k(t), k = 1, \ldots, n\}$ are taken to be independent and identically distributed and satisfy $\text{Cov}(U_k, U_{k'}) = \delta_{kk'}\sigma^2$. For the $k$-th cell, variable $U_k(t)$ remains constant within each cycle and is updated when the $k$-th cell reaches its division at $1 \sim 0$.

We integrate (2.23) numerically by the Euler method,

\[x_k^{j+1} = x_k^j + h \left[ a(x_k^j, I_j) + U_k^j \right],\]

where $U_k^j$ is updated when $x_k$ crosses 1. Figure 2.5 shows the distributions of the cells in the variable rate model for several values $\sigma$ of the noise SD.

Another source of the noise comes from the fact that budding yeast divides asymmetrically resulting in distinguishable mother and daughter cells. Because daughter
cells are smaller than their mothers right after division, they have a longer time to maturity and thus a longer cell cycle. These differences disperse the cell cycle synchrony of an initially synchronous population [37, 42] and complete population synchrony is generally impossible to maintain in the laboratory [10, 22, 41]. We refer to this mechanism of population dispersion as asymmetric division. Versions of this mechanism were studied in [13, 18] where, in the absence of coupling, it was shown to lead to a globally attracting
steady state (in the sense of a time-independent distribution of cells although these cells are progressing through the cell cycle).

To model asymmetric division we use Equation (2.2) for progression through the cell cycle, but we adjust the length of the cell cycle by applying a Bernoulli random variable to each cell when it divides. When a cell arrives at 1, the Bernoulli random variable will determine if the cell begins the next cycle at either 0 or $-2\sigma$, with equal probability. Thus each cell after division will have an equal probability of being either a mother or daughter cell. Here we choose $-2\sigma$ instead of $2\sigma$ for the sake of convenience in analyzing simulation results. Because in the simulations which will be explained later, we use measurements such as the center and the dispersion of cells in the first cluster and we study the distribution of cells that lie within the first half of the cell cycle, i.e., $[0, 0.5]$. If we let some cells start its cell cycle at $2\sigma$, then when looking at the distribution of cells, we will observe a sudden decrease in the number of cells right after 0 and the number of cells remains the same in $(0, 2\sigma)$ with a sudden increase at $2\sigma$. In this situation, the distribution of cells on $[0, 0.5]$ cannot provide us satisfactory statistical measurements (the density function of this distribution is not smooth). Therefore, it is mathematically reasonable to set the beginning of the cell cycle at a position on the unit circle that is not inside $[0, 0.5]$. For the same reason we do not use $-\sigma$ and $\sigma$ as the beginning of the cell cycles for daughter cells and mother cells.

Let $B_k \sim 2\sigma B(1, 0.5)$ be independent, identically distributed Bernoulli random variables, we apply $B_k$ on the cell $x_k$ when it divides, (i.e., the first time $x_k$ crosses 1 within each cycle). The progression of the $k$-th cell in each cycle governed by Equation (2.2) can be numerically integrated as

$$x_{k}^{l+1} = \begin{cases} 
    x_k^l + ha(x_k^l, I_j) - 1 - B_k, & x_k^l + ha(x_k^l, I_j) > 1 \\
    x_k^l + ha(x_k^l, I_j), & \text{otherwise}.
\end{cases}$$
Note that in contrast to the previous two models, the Bernoulli noise, $B_k$, is applied once within each cycle rather than at every step. Although all the three noise mechanisms have the same SD which is denoted by $\sigma$, the Bernoulli noise is a discrete random variable while Gaussian noise and uniform noise are continuous. Hence we expect different effective noise strengths between them.

### 2.3.2 Statistical measures

Throughout the simulations, we use a negative and linear coupling function, $f(I) = -0.5I$ for simplicity. We consider a collection of $n = 10^4$ cells. Initially, cells are distributed in two even clusters at $x_k(0) = \frac{1}{4}$ for $k = 1, \ldots, \frac{n}{2}$ and $x_k(0) = \frac{51}{70}$ for $k = \frac{n}{2} + 1, \ldots, n$. For convenience, we let the initial two clusters be denoted as $\tilde{x}_1(0) = \frac{1}{4}$ and $\tilde{x}_2(0) = \frac{51}{70}$. We pick these values because they are the 2-cluster periodic solution of the deterministic system. This can be shown as follows:

In the deterministic system (2.4) with $f(I) = -0.5I$, $R = [0.3, 0.65)$, and $S = [0.65, 0.95)$, we start with initial conditions $\tilde{x}_1(0) = \frac{1}{4}$ and $\tilde{x}_2(0) = \frac{51}{70}$. After time $t_1 = \frac{1}{20}$,

$$\tilde{x}_1(t_1) = \tilde{x}_1(0) + t_1 = \frac{1}{4} + \frac{1}{20} = 0.3.$$  

This implies that $\tilde{x}_1$ is going to enter the responsive region and at this time,

$$\tilde{x}_2(t_1) = \tilde{x}_2(0) + t_1 = \frac{51}{70} + \frac{1}{20} = \frac{109}{140},$$

$\tilde{x}_2$ is still in the signaling region.

The first cluster $\tilde{x}_1$ is affected by the feedback from the second cluster for a duration of $t_2 = 0.95 - \frac{109}{140} = \frac{6}{35}$ because at time $t = t_1 + t_2$ the second cluster $\tilde{x}_2$ is getting out of the signaling region, i.e., $\tilde{x}_2(t_1 + t_2) = 0.95$. We have

$$\tilde{x}_1(t_1 + t_2) = \tilde{x}_1(t_1) + (1 + f(\frac{1}{2}))t_2 = 0.3 + (1 - \frac{1}{4})\frac{6}{35} = \frac{3}{7}.$$
For $t_3 = 0.3$, we see that no clusters are experiencing feedback during the time interval $(t_1 + t_2, t_1 + t_2 + t_3)$, so

$$\tilde{x}_1(t_1 + t_2 + t_3) = \tilde{x}_1(t_1 + t_2) + t_3 = \frac{3}{7} + 0.3 = \frac{51}{70} = \tilde{x}_2(0).$$

Meanwhile,

$$\tilde{x}_2(t_1 + t_2 + t_3) = \tilde{x}_2(t_1 + t_2) + t_3 = 0.95 + 0.3 = 0.25 \text{ (mod 1)} = \tilde{x}_1(0).$$

Therefore, for a period $T = 2(t_1 + t_2 + t_3) = \frac{73}{70}$, we found that $\tilde{x}_1(T/2) = \tilde{x}_2(0)$ and $\tilde{x}_2(T/2) = \tilde{x}_1(0)$. Furthermore, $\tilde{x}_1(T) = \tilde{x}_1(0)$ and $\tilde{x}_2(T) = \tilde{x}_2(0)$. This implies that it is a 2-cluster periodic solution. We know this is also a stable solution because the parameter values we use are in the set of parameter values where 2-periodic solutions are stable under negative coupling [43].

The collective dynamics of the cell ensemble are conveniently described on a unit circle, so that the probability density function of the ensemble $p(t, x)$ is periodic, i.e., $p(t, x) = p(t, x + 1)$. The initial condition we use is a delta-peaked distribution with equal weighted peaks at $x = 1/4$ and $x = 51/70$. We will refer to this distribution of two clusters as delta-peaked clusters.

The dispersion of the first cluster under noise can be measured by using the angular deviation in the context of circular statistics. We apply the definitions of the mean resultant vector, the angular deviation and the circular mean in [14] to define the following measurements. The mean resultant vector of the first cluster is introduced as,

$$V_1(t) = \frac{2}{n} \left( \frac{1}{n} \sum_{k=1}^{n/2} \cos(2\pi x_k(t)), \frac{1}{n} \sum_{k=1}^{n/2} \sin(2\pi x_k(t)) \right).$$

For convenience, we will write this mean resultant vector $V_1(t)$ as:

$$V_1(t) = \langle \exp[2\pi i x_k(t)] \rangle_1,$$
where the averaging, \( \langle \cdot \rangle_1 \), is taken only over the cells which are initially in the first cluster.

Then, the angular deviation of the first cluster is

\[
D_c(t) = \sqrt{2[1 - |V_1(t)|]} = \sqrt{2[1 - |\langle \exp[4\pi i x_k(t)] \rangle_1|]}.
\] (2.24)

If cells form a delta-peaked cluster, then \( |V_1(t)| = 1 \) and \( D_c(t) = 0 \). Thus \( D_c(t) = 0 \) implies no dispersion of cells in the cluster. If cells are evenly distributed on the circle, then \( V_1(t) = 0 \) and \( D_c(t) \) attains its maximum \( \sqrt{2} \). Therefore, this angular deviation can be used to measure the dispersion of cells in the first cluster. Note that \( D_c(t) = \sqrt{2} \) does not necessarily imply the uniform distribution. For example, if cells form delta-peaked clusters with the number of clusters to be an even number \( 2m \) and each cluster is \( 1/2m \) distance away from its neighbors on the unit circle, then \( V_1(t) \) is also 0 and \( D_c(t) = \sqrt{2} \). But in our simulations this distribution of delta-peaked \( 2m \)-clusters cannot appear because we only consider the first cluster of cells. The first cluster of cells initially are at \( 1/4 \) and cells in this cluster progress according to the SDE model, the variable rate model or the asymmetric division model.

The circular mean of the first cluster can indicate the center of the first cluster. It is defined as follows:

\[
\bar{x}_1(t) = \frac{1}{2\pi} \text{Arg}(V_1(t))
\]

\[
= \begin{cases} 
\frac{1}{2\pi} \tan^{-1} \left( \frac{\sum_{k=1}^{n/2} \sin(2\pi x_k(t))}{\sum_{k=1}^{n/2} \cos(2\pi x_k(t))} \right), & \text{if } \sum_{k=1}^{n/2} \sin(2\pi x_k(t)) > 0, \sum_{k=1}^{n/2} \cos(2\pi x_k(t)) > 0 \\
\frac{1}{2\pi} \tan^{-1} \left( \frac{\sum_{k=1}^{n/2} \sin(2\pi x_k(t))}{\sum_{k=1}^{n/2} \cos(2\pi x_k(t))} \right) + \pi, & \text{if } \sum_{k=1}^{n/2} \cos(2\pi x_k(t)) < 0 \\
\frac{1}{2\pi} \tan^{-1} \left( \frac{\sum_{k=1}^{n/2} \sin(2\pi x_k(t))}{\sum_{k=1}^{n/2} \cos(2\pi x_k(t))} \right) + 2\pi, & \text{if } \sum_{k=1}^{n/2} \sin(2\pi x_k(t)) < 0, \sum_{k=1}^{n/2} \cos(2\pi x_k(t)) > 0.
\end{cases}
\]

The second Fourier harmonic of the distribution of cells on the unit circle is

\[
V_2(t) = \langle \exp[4\pi i x_k(t)] \rangle,
\] (2.25)

where in Equation (2.25) the averaging, \( \langle \cdot \rangle \), is taken over the entire ensemble, i.e.,

\[
\langle \cdot \rangle \equiv (1/n) \sum_{k=1}^{n} [\cdot].
\]

The magnitude of the second harmonic,

\[
L_2(t) = |V_2(t)| = |\langle \exp[4\pi i x_k(t)] \rangle|,
\] (2.26)
is called the order parameter. If the two delta-peaked clusters are at the opposite phases of each other, for example, $\tilde{x}_1(t) = 1/4$ and $\tilde{x}_2(t) = 3/4$, then

$$4\pi \tilde{x}_1(t) = 4\pi \tilde{x}_2(t) = \pi \pmod{2\pi},$$

so $L_2(t) = |\langle \exp[\pi i] \rangle| = 1$. As cells in the clusters disperse under the noise, $L_2$ decreases. So we can use this order parameter $L_2(t)$ to measure the phase coherence of the two-cluster solution. For the deterministic case the probability density is the sum of two delta functions in our models, $p(x) = 0.5[\delta(x - \tilde{x}_1) + \delta(x - \tilde{x}_2)]$ where $\tilde{x}_1 = 1/4$ and $\tilde{x}_2 = 51/70$, we can see that $L_2$ attains its maximal value near 1. In this case $L_2$ is not exactly 1 because the phase difference of the two clusters is not exactly $\pi$. If the distribution of cells is uniform on the unit circle (a completely disordered state), then the order parameter $L_2$ approaches its minimum 0.

In addition to circular statistical measures defined above, we have characterized the shape of the steady state probability distribution $p(x)$ using a normalized entropy, which is denoted as $E$. Numerically, we used $M = 100$ partitions to estimate $p(x)$ and its Shannon entropy,

$$H = -\sum_{m=1}^{M} p_m \log p_m.$$  

In the absence of noise, the two-cluster solution is characterized by a double delta-peaked distribution and the entropy takes its minimal value, $H_{\text{min}} = \log 2$ for symmetrical clusters with $n/2$ cells in each. In the completely disordered state the distribution $p(x)$ is uniform with the maximal possible entropy $H_{\text{max}} = \log M$. The normalized entropy, defined as

$$E = \frac{H - H_{\text{min}}}{H_{\text{max}} - H_{\text{min}}} = -\left( \frac{\sum_{m=1}^{M} p_m \log p_m + \log 2}{\log(M/2)} \right),$$

(2.27)

varies between 0 (noiseless two symmetrical clusters) and 1 (a disordered state where cells are distributed uniformly).

In the simulations, we integrate the three noise-induced models respectively for a fixed noise level, $\sigma$, from the initial condition (a double-peaked distribution) until the
distribution of cells $p(x)$ stabilizes at a steady state. We can make sure the final
distribution is at a steady state using two methods. The first method is a two-sample
Kolmogorov-Smirnov (K-S) test to compare the distributions at two different time $t$ but
the circular mean values of the first cluster are the same. The K-S test can show whether
the two distributions at the different time are the same at a given significant level. In
particular, snapshots of the ensemble are taken at time points when the circular mean of
the first cluster is $\bar{x}_1(t) = 1/4$. We compare the ensemble distribution after traversing the
circle 80 times with the ensemble distribution after traversing the circle 100 times. We
found that the K-S test accepted the hypothesis that the two distributions are the same at
significance level 0.05. So we conclude that the ensemble distribution settles down to its
steady state after 80 cycles. Representative examples of steady-state distributions for
individual cluster are shown in Figure 2.5. For weak noise ($\sigma \in [0.01, 0.02]$) the ensemble
converges in less than 20 cycles to a state that according to the K-S test is
indistinguishable for the steady state. For larger noise ($\sigma > 0.025$) the convergence rate is
slower. It takes about 80 cycles to reach the steady state. Figure 2.6 shows that the
ensemble distributions taken at a fixed position on the cycle evolves towards steady states
for the SDE model. Weak noise ($\sigma = 0.01$) results in spread of cells within each cluster
with extremely rare transitions of cells between clusters. Clusters still exist for an
intermediate noise level ($\sigma = 0.02$), although cells migrate between clusters. In the middle
picture in Figure 2.6, we see non-zero probability between the two clusters in the
time-dependent probability map. This shows that cells are migrating between the two
clusters at this intermediate noise level. For $\sigma > 0.04$ the clusters are destroyed by the
noise, so that the ensemble quickly approaches a steady state with a uniform distribution.

Another method we use to determine the final distribution is convergent to a steady
state relies on the convergence of the circular measures, $L_2(t)$ and $D_c(t)$, introduced above.
As we will show in the next subsection, for all the three models, the relaxation time for the
Figure 2.6: Time-dependent probability distributions $p(t,x)$ of the ensemble with Gaussian white noise (SDE model) for the indicated values of noise level $\sigma$. Logarithm of the probability distribution, $\log[p(t,x)]$, is shown as filled contours.

ensemble to reach its steady state is less than 100 cycles, except near a critical noise level at which clusters are destroyed completely. Because of periodicity of the two-cluster solution and finite size fluctuations, the ensemble-averaged measures $L_2$, $D_c$ and $E$ are time-dependent even in the steady state. We thus performed additional time averaging of $L_2$, $D_c$ and $E$ over $T = 100$ after the ensemble reached its steady state, resulting in time-averaged quantities, $\overline{L}_2$, $\overline{D}_c$ and $\overline{E}$. 
2.3.3 Numerical simulation results

In this subsection, we will summarize the results from numerical simulations of the three noise-induced models. The transient dynamics are illustrated in Figure 2.7. Both the order parameter \( L_2(t) \) and angular deviation of an individual cluster \( D_c(t) \) show characteristics approaching their steady states which are similar for all three models of noise. For small \( (\sigma < 0.03) \) and large \( (\sigma > 0.04) \) noise the order parameter and angular deviation settle to steady states within 100 cycles. However, for a critical noise strength \( (\sigma \approx 0.035 \) for the SDE, \( \sigma \approx 0.034 \) for the uniform and \( \sigma \approx 0.04 \) for the Bernoulli noise model) the relaxation process slows down dramatically and shows larger fluctuations in the steady state, indicating occurrence of a phase transition (bifurcation) from the ordered state with two clusters to the completely incoherent disordered state of the ensemble.

![Figure 2.7](image)

Figure 2.7: Time-dependence of (a) the order parameter, \( L_2(t) \), and of (b) the angular deviation, \( D_c(t) \), for the indicated values of noise level \( \sigma \) in the SDE model. For comparison, the corresponding dependencies \( L_2(t) \) for the variable rate model are shown by grey lines on panel (a). The values of noise standard deviations are the same as those for the SDE model, except for the middle grey curve, where the noise standard deviation is 0.034. In both figures, a.u. stands for arbitrary unit.

For the SDE model, cell migration between clusters is expected for any non-zero noise intensity with exponentially long waiting time. In contrast, for bounded noise
models (variable rate and asymmetric division) there is a critical noise intensity, $\sigma_h$, at which a “hard” bifurcation occurs [19]. For $\sigma < \sigma_h$ cells are stuck within clusters, while for $\sigma > \sigma_h$ cells can migrate between the clusters. Figure 2.8 shows the fraction of migrated cells after the ensemble has evolved over 80 cycles, plotted against noise level. As expected, the fraction of migrated cells grows with $\sigma$ for all three models. When the two-cluster structure is completely destroyed, cells that were in different clusters are mixed together and become uniformly distributed on the unit circle. Therefore, we expect half of the cells migrate out of the cluster they initially belong to, in the following sense:

In simulations, when the circular mean of the first cluster $\bar{x}_1(t^*)$ is 0.25 after 80 cycles at some time $t^*$, the fraction of cells that have escaped from their initial clusters is determined by

$$\frac{\#\{k : 1 \leq k \leq n/2, x_k(t^*) \in (0.5, 1)\} + \#\{k : n/2 + 1 \leq k \leq n, x_k(t^*) \in (0, 0.5)\}}{n}.$$  

As the SD of the noise increases to some critical value, we can see the fraction of cells that have escaped from their initial clusters is saturating to 1/2. This indicates the disordered state of the ensemble where the two-cluster structure is destroyed. Analytical calculations presented in Appendix B of [17] show that for the parameter values used, the hard bifurcation occurs at $\sigma_h \leq 1.4 \times 10^{-5}$, which is far below the noise level where migration of cells can be possibly observed for a reasonable integration time. We thus conclude that the hard bifurcation plays no essential role in these examples of biologically motivated noise.

Figure 2.9(a) shows the order parameter $L_2$ of the steady state distribution decreases as the noise level increases until the noise level $\sigma$ reaches a critical value $\sigma_c$, at which $L_2$ vanishes. Figure 2.9(b) shows the spread of cells in the first cluster, measured by the angular deviation $D_c$, increases as the noise level increases. $D_c$ reaches its maximum $\sqrt{2}$ at the critical value of $\sigma_c$ and remains at its maximum. As mentioned in the previous
Figure 2.8: Fraction of cells that have escaped from their initial clusters after 80 cycles for the three models. For noise levels below $\sigma \approx 0.01$ cells are effectively bound to their cluster by the asymptotic stability of the two-cluster solution.

subsection, these time averaged quantities are used at certain noise levels due to the fact that $L_2$ and $D_c$ are time-dependent because of the periodicity of the two-cluster solution and finite size fluctuations.

From both of the Figures 2.9 and 2.10, we see that the three noise models show qualitatively identical dependencies versus the standard deviation $\sigma$ of the noise. Quantitatively, however, the asymmetric division model shows consistently lower sensitivity to noise perturbations. We note that while in the SDE and variable rate models noise is applied throughout the whole cycle, the asymmetric division model is perturbed by Bernoulli noise once per cycle only. Hence an effective noise strength of the SDE and variable rate model is larger than that of the asymmetric division model. Furthermore, because the distribution of cells in asymmetric division model always has a higher density in the interval $[1 - 2\sigma, 1]$ in phase space the order parameter saturates to a non-zero value for large $\sigma$ as can be seen in Figure 2.10(a).

In addition, the dependencies $\overline{D}_c(\sigma)$ in Figure 2.10(b) show three distinct responses to noise perturbations. For weak noise, $\sigma < 0.015$, the angular deviation increases almost
linearly with $\sigma$, indicating a near linear response of the ensemble to noise perturbations. The near linear response is broken for intermediate noise when $D_c$ increases nonlinearly with progressively larger slopes. For strong noise $\sigma > \sigma_c$, we see that $D_c$ saturates to its maximum value. This is further illustrated in Figure 2.10(c) by local slopes of $D_c$ versus $\sigma$ dependencies, which clearly shows the regions of linear and nonlinear responses. We note that the local slope is a measure of sensitivity of the ensemble to noise perturbations. Thus the results show that this sensitivity is maximal near the transition to the disordered state.

In the disordered state, i.e., when $\sigma > \sigma_c$, the ensemble is insensitive to random perturbations, as indicated by the zero slope of $D_c(\sigma)$.

From Figure 2.11(a), we also see that with the increase of noise level the clusters exist until $\sigma$ reaches its critical value $\sigma_c$, above which the ensemble has a disordered state where clusters are indistinguishable. This change of shape of the steady state distribution is well reflected by the normalized entropy $E$, which increases from 0 (no noise) to its maximal value when the ensemble distribution becomes uniform. This is shown in

Figure 2.9: Plots of (a) the order parameter $L_2$ and (b) the angular deviation of the first cluster $D_c$ of the steady state distribution, as a function of the noise level $\sigma$ in the three models.
Figure 2.10: Stationary statistical measures of the cell ensemble versus noise strength $\sigma$ for the three noise models. (a) Time averaged order parameter, $\overline{L}_2$, vs noise standard deviation, $\sigma$, of the three models. (b) Time averaged angular deviation of the first cluster vs $\sigma$. Dotted line shows the maximum possible value of $\sqrt{2}$. (c) Local slopes of angular deviations estimated from the curves in panel (b).

Figure 2.11(b). We can also see from this figure that the three models show qualitatively the same behavior with respect to the distribution of the ensemble.

At the critical value $\sigma_c$ of noise SD, the ensemble demonstrates features of an order-disorder phase transition, such as slowdown of transient relaxation to the steady state and large fluctuations (see Figure 2.7). The transition of the distribution of the solution has the characteristics of an Andronov-Hopf bifurcation, as indicated by
Figure 2.11: (a) Steady-state probability distributions $p(x)$ of the ensemble with Gaussian white noise (SDE model) versus the standard deviation $\sigma$ of the noise. Logarithm of steady state probability distribution $\log[p(x)]$ is shown as filled contours. (b) Time averaged normalized entropy $\overline{E}$ of the ensemble in three models versus noise SD.

Figure 2.12. For $\sigma < \sigma_c$ the model possesses a stable limit cycle that bifurcates to a stable focus for $\sigma > \sigma_c$. Here the large noise effectively stabilizes the uniform solution, which otherwise is unstable. Further, in Figure 2.10(a) we observe that the amplitude of the periodic orbit has a square root type dependency on $\sigma$, as in a Hopf bifurcation. We note here that this transition is consistent with order-disorder transitions in the Kuramoto model of globally coupled stochastic phase oscillators [30]. There the rotation of the original variables has been discarded by the transformation to phase differences, so the linearization there has zero imaginary part at the bifurcation value.

All the above simulations use the coupling function $f(I) = -0.5I$ and we found that the critical noise level $\sigma_c \approx 0.035$. Below this critical noise level, the two-cluster structure of cells still exists. Because there is conclusive experimental evidence of clustering in a continuous yeast culture undergoing the phenomenon of autonomous oscillations [38, 43], we want to study the relationship between the critical noise strength and the coupling
Figure 2.12: Phase portraits of the SDE model projected onto the complex plane of the second Fourier harmonic. Black lines show steady state trajectories; grey lines show transients on both panels. Left: two-cluster ordered state, with a stable periodic orbit $\sigma = 0.03$. Right: stable equilibrium distribution at the disordered state $\sigma = 0.04$. The order-disorder transition occurs at $\sigma_c \approx 0.035$.

strength. Figure 2.13 shows that a stronger coupling requires stronger noise to destroy the two-cluster ordered state. The critical noise level $\sigma_c$ is indeed a function of the coupling strength $\alpha$, which is shown by the white curve in the figure. This white curve corresponds to a bifurcation line separating ordered and disordered states in the $(\alpha, \sigma)$ parameter plane. The region below the white curve represents cells in the two-cluster ordered state, possessing a limit cycle. While the region above the white curve represents cells in disordered states, which is a stable uniform solution. This curve is also consistent with phase transitions observed in globally coupled stochastic phase oscillator systems [30].

2.3.4 Yeast cultures have large coupling strengths

In budding yeast, a mother cell carries a scar from the bud and the daughter does not. Replicative age can be determined from the number of scars [31]. A natural phase space for a model stratified by replicative age is to represent the cell cycle of the daughters by $I_0 = [0, 1]$ and successive generations by $I_m = [\alpha_m, 1]$. In other words, the phase space is: $I_0 \cup I_1 \cup \cdots \cup I_n$ where $n$ is the highest generation considered. In a steady state bioreactor,
the fraction of cells in each generation decreases roughly geometrically with age, so $n$ does not need to be large for accurate modeling.

Denote $\tau_m = |I_m| = 1 - a_m$. The age distribution can be measured and is directly related to the relative lengths of the cell cycles $\tau_m$ [6]. For one autonomous oscillation experiment in which clustering was detected, the authors of [6] determined the following population fractions as a function of generation for the first four generations: 0.55, 0.24, 0.11, 0.05. The corresponding $\tau_m$’s are calculated to be: 1.0, 0.93, 0.86, 0.82. So the authors find:

$$a_0 = 0, \quad a_1 = .07, \quad a_2 = .14, \quad a_3 = .18.$$  

In [6, 38] it was shown that simulations using these proportions can accurately reproduce the behavior seen in the YAO experiment [8, 38]. We can calculate the standard deviation of the data $a_i$ with the corresponding weights as the standard deviation of the generational noise in the experiment. We denote this as

$$\sigma_g = 0.0365.$$  

This quantity is normalized to a time unit of one cell cycle in the experiment.
In Figure 2.13, we see that for $\sigma \approx 0.036$ a coupling coefficient of $\alpha > 0.6$ is needed for the coupling strength to overcome the dispersive action of the asymmetric division and produce detectable clustering. Consider that the units here are all normalized in terms of one unperturbed cell cycle. We can conclude that for clustering to exist in the autonomous oscillation experiment, as observed, the strength of coupling must be a rather large effect.

In a two-cluster state, as the clusters enter the region where they experience coupling, they need to have their rate of progression decreased by a normalized factor of at least 0.3, or, 30%. Such a decrease cannot be considered to be a small perturbation.

### 2.4 Conclusions

In this chapter, we have studied both the deterministic and the stochastic aspects of the cell cycle coupling models.

In Section 2.2, the inclusion of a gap in the cell cycle system is introduced. It is a reasonable biological modeling assumption on a small time delay between the production of and the action to the signaling mechanism which causes coupling in cell cycles. We have found that inclusion of the gap does not complicate the model beyond amenability to rigorous analysis.

We have fully studied the dynamics of this system when the cells are arranged into two cell cycle clusters. We have shown that the dynamics in this subspace are almost the same as for the model without a gap. The only difference is that the $\epsilon$-gap introduces a corresponding $\epsilon$ interval of neutral solutions near the synchronous solution. Because this solution has no basin of attraction beyond itself, it is not likely to be realized in real systems. We also considered the stability of cyclic periodic solutions with $k = M + 1$ clusters; just enough clusters to guarantee that interactions occur between consecutive clusters. For these cases we again found that the dynamics mirror those of the system with no gap.
In Section 2.3, we studied the effects of three different types of random perturbation on the stable 2-cluster periodic solutions and observed how the 2-cluster structure is destroyed as noise strength increases to a critical values. Two types of the dispersive mechanisms are biologically motivated and we compared the results for these two type of noise with the results for the standard Gaussian white noise. We see that the differences between biologically plausible bounded noise and Gaussian noise mechanisms are insignificant after proper scaling of the noise strength.

Meanwhile, for all three forms of applied noise perturbations the dispersion of clusters occurs in three distinct phases. First, for small noise there is a linear relationship between the noise level and the dispersion of the clusters measured by the circular angular deviation. In this phase cells do not migrate from their initial clusters on the time scales considered. In the second phase cells begin to migrate more frequently between clusters and the slope of the dispersion versus noise increases. Finally, for noise level above a certain threshold the clustering structure is destroyed and the cells assume a nearly uniform distribution. This phase transition between disorder and order exhibits many characteristics of a Hopf bifurcation.

Finally, we conclude that the clustering observed in autonomous oscillation experiments cannot be due to small coupling effects. When we compare experimental data with the transition from a clustered to the uniform distribution in Figure 2.11, we can see that the experimentally established level of dispersion due to asymmetric division alone is enough to require a coupling strength at least strong enough to produce a 30% relative decrease in rate of progression when a cell enters the responsive region. Although the model is minimal, it is both universal and normalized to the length of the unperturbed cell cycle. Thus all available evidence indicates that this estimate of relative effects of noise and coupling is reliable. One possible explanation for a coupling strength of this magnitude is that the mechanism actually involves one or more check points in the cycle.
3 Feasibility of Heteroclinic Routes

The material in this chapter is based on joint work published in [1] with Drs. Valentin Afraimovich and Mikhail Rabinovich. I am the second author of this paper. The particular heteroclinic network we studied in [1] and the main theorem were proposed by Drs. Afraimovich and Rabinovich. Here, I present a more general version of this theorem. I wrote Sections 3.1, 3.2, 3.5, 3.6, and part of Sections 3.3 and 3.4. In particular, I proved Lemma 3.2.1, Lemma 3.4.1 and Lemma 3.5.1. Lemma 3.4.2 and details in Subsections 3.4.2 and 3.4.3 were completed by Dr. Afraimovich and me together. I made all the figures in this chapter.

3.1 Introduction

In this chapter, we study continuous dynamical systems with a heteroclinic network under certain assumptions. We consider a continuous dynamical system in the following form

\[ \frac{dX}{dt} = F(X), \quad X \in \mathbb{R}^{LN}_+ \quad \text{and} \quad t \in \mathbb{R}, \]  

(3.1)

where

1. the state space $\mathbb{R}^{LN}_+$ is the non-negative hyperoctant of the $LN$-dimensional space $\mathbb{R}^{LN}$ where $L \geq 2$ and $N \geq 3$ are positive integers. We use the notation $X = (\ldots, x_l^i, \ldots)$ for the elements of the state space $\mathbb{R}^{LN}_+$.

2. the function $F : \mathbb{R}^{LN}_+ \to \mathbb{R}^{LN}_+$ is continuously differentiable and $F(0) = 0$.

More specifically, we focus on systems as above satisfying the following assumptions:

A1: The system (1) has $LN$ saddle equilibrium points where each saddle point lies on a coordinate axis of $\mathbb{R}^{LN}_+$. Denote the saddles by $Q^i_l$ where $i = 1, \ldots, N$ and $l = 1, \ldots, L$. We assume that for each $i$ and $l$ the saddle point
\[ Q'_i = (x'_1, \ldots, x'_l, \ldots, x'_N) \in \mathbb{R}^{LN}_{++} \] where \( x'_j \neq 0 \), and \( x'^m_j = 0 \) for all \((j, m) \neq (i, l)\). Thus, each \( Q'_i \) is on a coordinate axis, which we refer to as the \( x'_j \)-axis.

**A2:** The non-negative hyperoctant \( \mathbb{R}^{LN}_{++} \) is invariant. The coordinate hyperplanes \( \{x'_i = 0\} \) for all \( i = 1, \ldots, N \) and \( l = 1, \ldots, L \) are also invariant. This implies that for any \( i \) and \( l \), the \( x'_i \)-axis is invariant. We can also see that for any \( j \neq i \) and \( m \neq l \), the \( (x'_j, x'^m) \)-planes are also invariant, i.e., the plane where \( x'^m_k = 0 \) if \((k, n) \neq (i, l)\) and \((k, n) \neq (j, m)\) is invariant. Moreover, the subspace containing the saddles \( Q'_1, Q'_2, \ldots, Q'_N \) and the origin \( 0 \) is invariant for each \( l \). Denote such an invariant subspace by \( \Sigma_l \) for each \( l = 1, \ldots, L \).

**A3:** For each of the saddle points \( Q'_i \) where \( i = 1, \ldots, N \) and \( l = 1, \ldots, L \), denote the derivative of \( F \) at \( Q'_i \) as \( DF(Q'_i) \), and denote the eigenvalues and eigenvectors of \( DF(Q'_i) \) as \( \lambda^m_{ij} \) and \( v^m_{ij} \), where \( j = 1, \ldots, N \) and \( m = 1, \ldots, L \). We assume that all the eigenvalues are real, and denote the eigenvectors as \( v^m_{ij} = [\ldots, x'^m_k, \ldots]^T \). The eigenvalues and eigenvectors satisfy:

(a) The eigenvalue \( \lambda^m_{i+1,i} > 0 \) for all \( i = 1, \ldots, N \) where \( N + 1 \equiv 1 \) and \( l = 1, \ldots, L \).

(b) The eigenvalue \( \lambda^m_{i,i+1} > 0 \) for all \( i = 1, \ldots, N \) and \( l = 1, \ldots, L - 1 \).

(c) When \( l = L \), the eigenvalue \( \lambda^m_{i,1} < 0 \) for all \( i = 1, \ldots, N \).

(d) The remaining eigenvalues \( \lambda^m_{ij} < 0 \) where \((j, m) \neq (i + 1, l)\) and \((j, m) \neq (i, l + 1)\), for all \( i, j = 1, \ldots, N \) where \( N + 1 \equiv 1 \), and \( l, m = 1, \ldots, L \).

(e) The eigenvector \( v^m_{ij} \) follows the direction of the \( x'_j \)-axis. Recall that \( Q'_i \) is on this axis. This eigenvector can be written as \( v^m_{ij} = \{x'_j = 1, \text{ and } x'^m_k = 0 \text{ if } (k, n) \neq (i, l)\} \).

(f) when \((j, m) \neq (i, l)\), there is a constant \( c_j^m > 0 \) such that the eigenvector

\[ v^m_{ij} = \{x'_j = 1, x'^m_j = c_j^m, \text{ and } x'^m_k = 0 \text{ if } (k, n) \notin (i, l, (j, m))\} \].
From this particular form of the eigenvectors, we can see that when \((j, m) \neq (i, l)\), the direction of the eigenvector \(v_{ij}^{lm}\) is transversal to the \(x_i^l\)-axis, and parallel to the \((x_i^l, x_m^m)\)-plane. Our assumptions imply that \(DF(Q_l^i)\) has a full set of eigenvectors for all \(i = 1, \ldots, N\) and \(l = 1, \ldots, L\).

A4: There is a heteroclinic orbit, denoted as \(P_{l1}^i\), connecting \(Q_l^i\) to \(Q_{l+1}^i\) for \(i = 1, \ldots, N\) and \(l = 1, \ldots, L\). We let \(N + 1 \equiv 1\) so that there is a heteroclinic orbit \(P_{N1}^i\) that connects the saddle \(Q_N^i\) to the saddle \(Q_1^i\).

A heteroclinic orbit is a path in the phase space which joins two different equilibrium points. For any \(X_0 \in P_{l1}^i\), let \(X(t, X_0)\) be the solution of (3.1) and \(X(0) = X_0\), then

\[
\lim_{t \to \infty} X(t, X_0) = Q_{i+1}^l, \quad \text{and} \quad \lim_{t \to -\infty} X(t, X_0) = Q_l^i.
\]

A5: There is also a heteroclinic orbit connecting the saddle \(Q_l^i\) to the saddle \(Q_{i+1}^l\) for \(i = 1, \ldots, N\) and \(l = 1, \ldots, L - 1\). Denote this heteroclinic orbit as \(P_{i2}^i\).

We can see from Assumptions A3-A5 that at each saddle point \(Q_l^i\) (where \(i = 1, \ldots, N\), and \(l = 1, \ldots, L - 1\)), there is a 2-dimensional unstable manifold and an \((LN - 2)\)-dimensional stable manifold. At the saddle point \(Q_i^l\) (where \(i = 1, \ldots, N\)), there is a 1-dimensional unstable manifold and an \((LN - 1)\)-dimensional stable manifold.

A6: The heteroclinic orbit that starts at the saddle \(Q_{i-1}^l\) and goes to \(Q_i^l\) follows the leading direction on the stable manifold of \(Q_i^l\) in the subspace \(\Sigma_i\) for all \(i = 1, \ldots, N\) and \(l = 2, \ldots, L\). We assume that \(\max\{\lambda_{ij}^l : \lambda_{ij}^l < 0\} = \lambda_{N-1}^l\) so that the leading direction on the stable manifold is \(\tilde{v}_{l}^{N-1}\) which is transversal to the \(x_i^l\)-axis and parallel to the \((x_{i-1}^l, x_i^l)\)-plane.
In the invariant plane $\Sigma$, there are saddles $Q^1_l, Q^2_l, \ldots, Q^N_l$ by Assumption A2. Under Assumptions A4-A6, we know that there exists a heteroclinic cycle, which is a collection of saddle points and the connecting heteroclinic orbits. For each $l$, the heteroclinic cycle can be written as

$$\Gamma_l = \bigcup_{i=1}^{N} (Q^l_i \cup P^l_{i1}) \quad \text{for} \quad l = 1, \ldots, L.$$ 

For each $l = 1, \ldots, L - 1$, by Assumption A5, there are heteroclinic orbits $P^l_{i2}$ joining the saddles $Q^l_i$ and $Q^{l+1}_i$ between the heteroclinic cycles $\Gamma_l$ and $\Gamma_{l+1}$ where $i = 1, \ldots, N$. A heteroclinic network consists of all the heteroclinic cycles $\Gamma_l$ and the heteroclinic orbits $P^l_{i2}$ joining them. Denote this heteroclinic network as

$$\Gamma = \bigcup_{l=1}^{L-1} \left( \bigcup_{i=1}^{N} P^l_{i2} \right) \bigcup \Gamma_L.$$ 

Figure 3.1 is an illustration of the heteroclinic network with $N = 6$ and $L = 3$.

![Figure 3.1](image)

Figure 3.1: An illustration of the heteroclinic network when $L = 3$ and $N = 6$.

We define a directed graph $G = (V, E)$ that is identified with the heteroclinic network. In this graph $G$, the set of vertices is $V = \{Q^l_i\}_{i=1}^{N}, {l=1, \ldots, L}$, and the set of directed edges $E$ is chosen in such a way that there is an edge starting at $Q^l_i$ and ending at $Q^m_j$ if and only if
there is a heteroclinic orbit connecting $Q^l_i$ to $Q^m_j$. Let us enumerate the vertices by numbers $1, 2, \ldots, p$ where $p = LN$. The adjacency matrix of the graph is defined as $M = \{a_{sk}, a_{sk} \in \{0, 1\}\}$, where $a_{sk} = 1$ if and only if there exists an edge starting at the vertex $s$ and ending at the vertex $k$.

Define an alphabet $A = \{\omega_0, \omega_2, \ldots, \omega_{LN-1}\}$ as the symbols of vertices in the graph $G$. The set of words is denoted as $A^* = \cup_{n \geq 0} A^n$ and the set of infinite words is denoted as $A^N$.

We can define a one-sided subshift $(\Omega_M, \sigma)$ where $\sigma$ is the shift operator and

$$\Omega_M = \{ (\omega_0, \omega_1, \ldots) \in A^N : \forall i \geq 0, a_{\omega_i\omega_{i+1}} = 1 \text{ in the matrix } M \}.$$ 

That is to say, for any $(\omega_0, \omega_1, \ldots) \in \Omega_M$, there is a directed edge in the graph from the vertex $\omega_i$ to the vertex $\omega_{i+1}$. So $(\omega_0, \omega_1, \ldots)$ is a path in the graph consisting of an infinite sequence of edges that connect vertex $\omega_0$ to vertex $\omega_1$, vertex $\omega_1$ to vertex $\omega_2$, and so on.

The subshift $(\Omega_M, \sigma)$ can be defined in terms of the following set of forbidden words:

$$W = \{ (\omega_i\omega_j) : (\omega_i\omega_j) \in A^*, a_{\omega_i\omega_j} = 0 \text{ in the matrix } M \}.$$ 

In other words, $W$ is the set of all pairs of vertices which are not connected by directed edges. Because the number of vertices in the graph $G$ is finite, the set $W$ is also finite. Thus $(\Omega_M, \sigma)$ is a subshift of finite type.

Any non-empty cylinder $[\omega_0, \omega_1, \ldots, \omega_k]$ of $(\Omega_M, \sigma)$ is the set of all infinite words starting with $\omega_0, \omega_1, \ldots, \omega_k$. Let us label the saddle points in the heteroclinic network corresponding to $\omega_0, \omega_1, \ldots, \omega_k$ as $Q_{\omega_0}, Q_{\omega_1}, \ldots, Q_{\omega_k}$. Then the path $\omega_0, \omega_1, \ldots, \omega_k$ corresponds to a sequence of heteroclinic orbits that starts from the saddle $Q_{\omega_0}$ to the saddle $Q_{\omega_1}$, follows a heteroclinic orbit to the saddle $Q_{\omega_2}$, and so on until it reaches the saddle $Q_{\omega_k}$. We call such a sequence of heteroclinic orbits as a heteroclinic route and denote it by $\Gamma([\omega_0, \omega_1, \ldots, \omega_k])$. 


3.2 Symbolic complexity of the subshift

In the previous section we have introduced heteroclinic routes that are sequences of heteroclinic orbits in the heteroclinic network that connects the saddle points according to the symbolic description represented by cylinders of the subshift of finite type \((\Omega_M, \sigma)\). It is natural to ask how many different heteroclinic routes there are, i.e., how many different non-empty cylinders can be realized by the system.

The symbolic complexity of the subshift of finite type \((\Omega_M, \sigma)\) is defined as

\[
C_n = \#\{ \text{all nonempty cylinders } [\omega_0, \omega_1, \ldots, \omega_{n-1}] \text{ of length } n. \}
\]

It is known (see, for instance, page 73 in [2]) that \(C_n = uM^{n-1}u^T\), where \(u = (1, 1, \ldots, 1)\) is the NL-row vector with all coordinates equal one and \(M\) is the adjacency matrix defined in the previous section. We prove the following lemma.

**Lemma 3.2.1.** For the subshift of finite type \((\Omega_M, \sigma)\) derived by the heteroclinic network \(\Gamma\), the symbolic complexity function \(C_n = P_{L-1}(n)\) is a polynomial of degree \(L - 1\), where \(L\) is the number heteroclinic cycles in the heteroclinic network.

To prove this lemma, we need to calculate \(uM^{n-1}u^T\) when \(n\) is large. Note that the adjacency matrix \(M\) can be partitioned into an \(L \times L\) block matrix:

\[
M = \begin{pmatrix}
B & I & 0 & \cdots & 0 \\
0 & B & I & 0 & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & B & I & \vdots \\
0 & \cdots & \cdots & 0 & B
\end{pmatrix},
\]

where \(B\) and \(I\) denote block matrices.
where $I$ is an $N \times N$ identity matrix, $0$ is an $N \times N$ zero matrix and $B$ is an $N \times N$ permutation matrix of the following form:

$$B = \begin{pmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & 0 & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & 0 & 1 \\
1 & 0 & \cdots & \cdots & 0
\end{pmatrix}.$$  

We can find that for $n \geq L$,

$$M^{n-1} = \begin{pmatrix}
B^{n-1} & (n-1)B^{n-2} & (n-1)B^{n-3} & \cdots & (n-1)B^{n-L} \\
0 & B^{n-1} & (n-1)B^{n-2} & \cdots & (n-1)LB^{n-L+1} \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & \cdots & B^{n-1}
\end{pmatrix}. \quad (3.2)$$

Therefore, for $u_N = (1, 1, \ldots, 1)$, the $N$-row vector of all coordinates $1$, we obtain

$$uM^{n-1}u^T = LnB^{n-1}u_N^T + (L-1)\begin{pmatrix}
(n-1) \\
1
\end{pmatrix}u_NB^{n-2}u_N^T + \cdots + 2\begin{pmatrix}
(n-1) \\
L-2
\end{pmatrix}u_NB^{n-L+1}u_N^T + \begin{pmatrix}
(n-1) \\
L-1
\end{pmatrix}u_NB^{n-L}u_N^T. \quad (3.3)$$

Note here that an iteration of a permutation matrix is still a permutation matrix. So $u_NB^{k}u_N^T = N$ for any integer $k$ when $B$ is an $N \times N$ permutation matrix. We can see that the highest degree of $n$ in the polynomial is determined by the term $\begin{pmatrix}
(n-1) \\
L-1
\end{pmatrix}$. Therefore, the highest degree of $n$ is $L - 1$. 

We also see from here that the topological entropy

\[ h = \lim_{n \to \infty} \ln C_n/n = 0. \]

This implies that the symbolic system is not chaotic.

### 3.3 The main theorem

In this section, we state a theorem about the behavior of some particular trajectories of the system (3.1) satisfying Assumptions A1-A6. Lemma 3.2.1 shows the number of heteroclinic routes can be realized in this system when the number of switchings is \( n \) (i.e., a representative point on a trajectory started in a neighborhood of the beginning of a heteroclinic orbit meets small pairwise disjoint neighborhoods of some saddle equilibrium points \( n \) times). We found that the number of such heteroclinic routes is a polynomial of \( n \) of degree \( L - 1 \), where \( L \) is the number of heteroclinic cycles in the heteroclinic network \( \Gamma \). We can also study the behavior of the trajectories in system (3.1) that are in a neighborhood of such heteroclinic routes.

We make the following assumptions in order to prove the theorem:

**A7:** For each \( l = 1, \ldots, L \), in the subspace \( \Sigma_l \) that contains the saddles \( \{Q^l_i\}_{i=1}^N \), by Assumptions A5-A6 and A3, the saddle value of each \( Q^l_i \) is

\[ \nu^l_i = \frac{\lambda^l_{i+1}}{\lambda^l_{i-1}}. \]

We assume that \( \nu^l_i > 1 \) for all \( i \) and \( l \). Such a saddle point \( Q^l_i \) is called a *dissipative saddle*.

**A8:** For the saddles \( Q^l_i \) where \( i = 1, \ldots, N \) and \( l = 1, \ldots, L - 1 \), we assume that

\[ \lambda^l_{i+1} > \lambda^l_{i+1}. \]

By Assumptions A3-A5, this implies that the unstable direction within the subspace \( \Sigma_i \) is stronger than the unstable direction between the subspaces \( \Sigma_i \) and \( \Sigma_{i+1} \).
A9: The system (3.1) is $C^1$-linearizable in a neighborhood of each $Q'_l$, $i = 1, \ldots, N$.

To study the behavior of the trajectories following a heteroclinic route $\Gamma((\omega_0, \omega_1, \ldots, \omega_k))$ in the phase space which is defined according to the non-empty cylinder $[\omega_0, \omega_1, \ldots, \omega_k]$ (admissible words) of the subshift of finite type $(\Omega_M, \sigma)$, we prove the following theorem. Recall that in Section 3.1 the corresponding saddle points on the heteroclinic route $\Gamma((\omega_0, \omega_1, \ldots, \omega_k))$ are denoted by $Q_{\omega_0}, Q_{\omega_1}, \ldots, Q_{\omega_k}$.

**Theorem 3.3.1.** Under Assumptions A1-A9, for any non-empty cylinder $[\omega_0, \omega_1, \ldots, \omega_k]$ in the subshift of finite type $(\Omega_M, \sigma)$, and for any neighborhood $V$ of the heteroclinic route $\Gamma((\omega_0, \omega_1, \ldots, \omega_k))$ and any neighborhoods $O_\epsilon(Q_{\omega_i}) \subset V$ of saddle points $Q_{\omega_i}$ where $i = 0, 1, \ldots, k$, there exists an open set $U \subset O_\epsilon(Q_{\omega_0})$ such that for each initial point $X_0 \in U$ the corresponding trajectory $X(t, X_0) \in V$ for $0 \leq t \leq T$. This trajectory $X(t, X_0)$ visits the neighborhoods $O_\epsilon(Q_{\omega_i})$ in the corresponding order and $X(T, X_0) \in O_\epsilon(Q_{\omega_k})$; i.e., the trajectory $X(t, X_0)$ is shadowing the heteroclinic route $\Gamma((\omega_0, \omega_1, \ldots, \omega_k))$.

The shadowing property in Theorem 3.3.1 means that given heteroclinic route $\Gamma((\omega_0, \omega_1, \ldots, \omega_k))$ and any small neighborhoods $O_\epsilon(Q_{\omega_i})$ of the saddles, there is a smaller neighborhood $U$ of the first saddle $Q_{\omega_0}$, such that trajectories starting with initial points in $U$ stay close to the fixed heteroclinic route $\Gamma((\omega_0, \omega_1, \ldots, \omega_k))$ and visit the fixed neighborhoods of the remaining saddle points $Q_{\omega_1}, \ldots, Q_{\omega_k}$ in the corresponding order.

The proof of this theorem is given in Section 3.4. We first describe the idea of the proof as follows.

Given a non-empty cylinder $[\omega_0, \ldots, \omega_k]$, denote the heteroclinic orbits in $\Gamma([\omega_0, \ldots, \omega_k])$ as $P_{\omega_i\omega_{i+1}}$ for $i = 0, \ldots, k - 1$, where $P_{\omega_i\omega_{i+1}}$ is the heteroclinic orbit that starts at $Q_{\omega_i}$ and ends at $Q_{\omega_{i+1}}$. Because of the assumptions A1-A6, each saddle $Q_{\omega_i}$ (labeling $Q''_j$, $m \neq 1, L$) has two heteroclinic orbits, say $H_1^+$ and $H_2^+$, coming to it (one of
them is $P_{\omega_{i-1}\omega_i}$) and two heteroclinic orbits, say $H_1^+$ and $H_2^+$ (one of them is $P_{\omega_i\omega_{i+1}}$) coming out of it, see Figure 3.2(a). If $Q_{\omega_i} = Q_j^L$ for some $j$, then we deal with the only heteroclinic trajectory $P_{\omega_{i-1}\omega_i}$ joining $Q_{\omega_{i-1}}$ and $Q_{\omega_i}$ where $Q_{\omega_{i-1}} = Q_{j-1}^L$, and two heteroclinic orbits coming out of $Q_{\omega_i}$. If $Q_{\omega_i} = Q_j^L$ for some $j$, then we deal with the only heteroclinic orbit coming out of $Q_{\omega_i}$, say $P_{\omega_i\omega_{i+1}}$ where $Q_{\omega_{i+1}} = Q_{j+1}^L$, and two heteroclinic orbits ending at $Q_{\omega_i}$.

![Figure 3.2: Scheme of the local behavior (a) and global behavior (b) of trajectories in a neighborhood of $\Gamma$.](image)

We endow each heteroclinic orbit with a section transversal to the flow in a neighborhood of each saddle $Q_{\omega_i}$, $H_k^-$ with $S_k^-$ and $H_k^+$ with $S_k^+$, $k = 1, 2$, see Figure 3.2(a). Then we show in Subsection 3.4.1 that for each pair $(S_k^-, S_p^+)$, $k, p \in \{1, 2\}$ there is an open set of initial points in an $\epsilon$-neighborhood of $H_k^- \cap S_k^-$ that is mapped by a local map along the trajectories into an open set in a neighborhood of $S_p^+ \cap H_p^+$. Then points in this open set move near the heteroclinic orbit $P_{\omega_i\omega_{i+1}}$ to a section $S_k^-$ in a neighborhood of $Q_{\omega_{i+1}}$, see Figure 3.2(b). Based on this observation, we can define global maps from $S_k^+(Q_{\omega_i})$ to $S_k^-(Q_{\omega_{i+1}})$. We show in Subsection 3.4.2 that this global map behaves qualitatively as a translation. Thus, we can repeat our local consideration again around $Q_{\omega_{i+1}}$ and repeat the above procedures. Finally, we obtain an open set of initial points in the $\epsilon$-neighborhood of
such that the trajectory going through each of them shadows the heteroclinic route \( \Gamma([\omega_0, \ldots, \omega_k]) \).

### 3.4 Proof of the main theorem

We consider the linearization of the system (3.1) at each saddle points \( Q^l_i \). By Hartman-Grobman Theorem, we know that the local behavior of (3.1) is topologically equivalent to the local behavior of the linearization of (3.1). Because of Assumption A9 that the system is \( C^1 \)-linearizable at \( Q^l_i \), there exists a continuously differentiable one-to-one map of a small neighborhood of \( Q^l_i \) onto an open set \( \tilde{U}(0) \) containing the origin \( 0 \in \mathbb{R}^{LN} \) which transforms (3.1) into its linearization. Let us write this linearization of (3.1) at \( Q^l_i \) as

\[
\frac{dX}{dt} = DF(Q^l_i)X,
\]

(3.4)

where \( DF(Q^l_i) \) is the derivative of \( F \) at \( Q^l_i \).

Because of Assumption A3, the matrix \( DF(Q^l_i) \) has a full set of eigenvectors. Thus, it can be diagonalized even in the case where some of its eigenvalues are of multiplicity greater than 1. Denote the matrix of the eigenvectors of \( DF(Q^l_i) \) as \( \tilde{V}^l_i \), then we apply a similarity transformation to diagonalize the matrix \( DF(Q^l_i) \) in the following way:

\[
(\tilde{V}^l_i)^{-1}DF(Q^l_i)(\tilde{V}^l_i) = \tilde{\Lambda}^l_i,
\]

where \( \tilde{\Lambda}^l_i \) is a diagonal matrix with the corresponding eigenvalues of \( DF(Q^l_i) \) on its diagonal. Let us apply a change of variables that \( \eta = (\tilde{V}^l_i)^{-1}X \), then (3.4) can be rewritten as

\[
\frac{d\eta}{dt} = \tilde{\Lambda}^l_i\eta.
\]

(3.5)

Denote the new coordinates as \( \eta = (y_1, y_2, \xi_1, \ldots, \xi_{LN-2}) \), where \( y_1, y_2 \) are the new coordinate axes corresponding to the eigenvectors related to the positive eigenvalues and \( \{\xi_i\}_{i=1,\ldots,LN-2} \) are those related to the negative eigenvalues. Let \( y = (y_1, y_2) \) and
\( \xi = (\xi_1, \ldots, \xi_{LN-2}) \in \mathbb{R}^{LN-2} \). Thus, under the assumptions A1-A7, the system (3.1) in a neighborhood of \( Q'_l \) can be rewritten as

\[
\begin{align*}
\dot{y}_1 &= \gamma_1 y_1, \\
\dot{y}_2 &= \gamma_2 y_2, \\
\dot{\xi} &= \Lambda \xi,
\end{align*}
\]

where \( 0 < \gamma_1 = \lambda_{ii}^{ll+1} \), and \( 0 < \gamma_2 = \lambda_{ii}^{ll} \), and \( \Lambda \) is a diagonal matrix with the negative eigenvalues on its diagonal. According to Assumption A8, we know that \( 0 < \gamma_2 < \gamma_1 \). We also know that

\[
||e^{\Lambda t}|| \leq e^{\mu t}, \quad \text{for some } \mu < 0 \text{ and all } t \geq 0.
\]

Without loss of generality we assume that \( \xi_1 \) is the coordinate corresponding to the leading direction of the intersection of the stable manifold of \( Q'_l \) with the subspace \( \Sigma_l \) (see Assumption A6). The coordinate \( \xi_2 \) corresponds to the eigenvector related to the eigenvalue \( \lambda_{ii}^{ll-1} < 0 \), i.e., it corresponds to the direction of the heteroclinic orbit joining to \( Q'_l \) from \( Q'_{l-1} \) in the \((x_{li}, x_{li+1})\)-plane.

By Assumption A3, at each saddle point \( Q'_l \), one eigenvector \( \vec{v}_{ii}^{ll} \) coincides with the \( x_{li} \)-axis, and another eigenvector \( \vec{v}_{ii+1}^{ll} \) is parallel to the \((x_{li}, x_{li+1})\)-plane. The new coordinate \( y_1 \) corresponds to the direction of this eigenvector \( \vec{v}_{ii+1}^{ll} \). A piece of the phase portrait on the \((x_{li}, x_{li+1})\)-plane looks as in Figure 3.3. We can see that positive values of \( y_1 \) correspond to positive values of \( x_{li+1}^{l} \), and negative values of \( y_1 \) correspond to negative values of \( x_{li+1}^{l} \).

We also know that \( y_1 = 0 \) corresponds to \( x_{li+1}^{l} = 0 \). The same consideration can be done on the \((x_{li}, x_{li+1})\)-plane, where 0, positive or negative values of \( y_2 \) correspond to values of \( x_{li+1}^{l} \) with the same property under the change of coordinates. Note that in our system (3.1), we only consider that the variables \( x_{li} \) are non-negative. Therefore, \( y_1 \geq 0 \) and \( y_2 \geq 0 \). The same consideration can also be done on the other \((x_{li}, x_{mj})\)-planes where \((j, m) \notin \{(i, l), (i + 1, l), (i, l + 1)\}\). Therefore, at each \( Q'_l \), except for one coordinate that
coincides with the $x_i^l$-axis, all the other coordinates $y_i$ and $\xi_i$ are nonnegative and they correspond to non-negative values of $x_m^i$ (where $(j, m) \neq (i, l)$). Denote the coordinate that coincides with the $x_i^l$-axis as $\xi_3$ for each $Q_i^l$.

The intersections of a neighborhood of $Q_i^l$ with the heteroclinic orbit belonging to the heteroclinic route starting at $Q_i^l$ have equations:

$$\xi = 0, \quad y_2 = 0$$

for a local segment of the orbit in the heteroclinic cycle $\Gamma_i$, and

$$\xi = 0, \quad y_1 = 0$$

for a segment of the orbit joining $Q_i^l$ and $Q_i^{l+1}$. Furthermore, the local segments of the heteroclinic orbits ending at $Q_i^l$ have the equations

$$y_1 = y_2 = 0, \quad \xi_j = 0 \text{ where } j \neq 1$$

for the orbit in the heteroclinic cycle $\Gamma_i$ and

$$y_1 = y_2 = 0, \quad \xi_j = 0 \text{ where } j \neq 2$$

for the orbit joining $Q_i^{l-1}$ and $Q_i^l$. 

Figure 3.3: The positive value of $y_1$ (point A) corresponds to the positive value of $x_i^l$. The negative value (point B) corresponds to the negative one. We only consider the non-negative values of the variables.
To prove Theorem 3.3.1, i.e., to find trajectories that shadow the heteroclinic route, we endow the heteroclinic route with a sequence of sections transversal to the flow and we study images (and preimages) of regions on these sections with respect to the corresponding maps. Thus, at each $Q^l_i$, we introduce

$$
S^-_1 = \{(y, \xi) : \xi_1 = \delta, |(\xi_2, \ldots, \xi_{LN-2})| \leq \epsilon, y_1 \leq \epsilon, y_2 \leq \epsilon\},
$$

$$
S^-_2 = \{(y, \xi) : \xi_2 = \delta, |(\xi_1, \xi_3, \ldots, \xi_{LN-2})| \leq \epsilon, y_1 \leq \epsilon, y_2 \leq \epsilon\},
$$

$$
S^+_1 = \{(y, \xi) : y_1 = \delta, y_2 \leq \epsilon, |\xi| \leq \epsilon\},
$$

$$
S^+_2 = \{(y, \xi) : y_2 = \delta, y_1 \leq \epsilon, |\xi| \leq \epsilon\}.
$$

where $\epsilon$ is a small parameter which is, in fact, the size (diameter) of the section. We will write $S^-_1(Q^l_i)$, $S^-_2(Q^l_i)$, $S^+_1(Q^l_i)$ and $S^+_2(Q^l_i)$ where it is necessary. Note that because the eigenspace and any subspaces of it are invariant, the invariance of the coordinate axes, planes and hyperplanes in Assumption A2 are preserved. Also, as we stated earlier, all the new coordinates $y_i$ and $\xi_j$ are nonnegative except one coordinate $\xi_3$ that coincides with the $x_l^i$-axis.

### 3.4.1 Local maps

Denote the local map in a neighborhood of a saddle point $Q^l_m$ as $T_{ij} : S^-_i \rightarrow S^+_j$ where $i, j \in \{1, 2\}$. This local map $T_{ij}$ maps points in $S^-_i$ to points in $S^+_j$ in a small neighborhood of the saddle point $Q^l_m$ near the heteroclinic orbits $H^-_i$ and $H^+_j$ as shown in Figure 3.2(a). We can construct the local maps in the following way.

For the map $T_{i1} : S^-_i \rightarrow S^+_1$, $i = 1, 2$, we solve the system (3.6) to obtain that

$$
y_1(t) = y_{10}e^{\gamma_1 t},
$$

$$
y_2(t) = y_{20}e^{\gamma_2 t},
$$

$$
\xi(t) = \xi_0 e^{\Lambda t},
$$

where $(y_{10}, y_{20}, \xi_0) \in S^-_i$ is the initial condition. We can find the transition time $t_i$ for the trajectory starts at $(y_{10}, y_{20}, \xi_0)$ to arrive at the section $S^+_1$ under map $T_{i1}$ from the
restriction that \( y_1(t_1) = \delta \):

\[
t_1 = 1/\gamma_1 \ln(\delta/y_{10}).
\]

So, we can write the equations for \( T_{i1} \):

\[
y_2(t_1) = y_{20} (\delta/y_{10})^{y_2/y_1},
\]

\[
\xi(t_1) = \xi_0 e^{\Lambda t_1}.
\]

(3.8)

Since \( t_1 \to \infty \) as \( y_{10} \to 0 \), we know that \( t_1 \gg 1 \) if \( \varepsilon \ll 1 \). Under this condition, the inequality \(|\xi(t_1)| \leq \varepsilon\) holds. So we can omit the consideration of the behavior of the stable coordinates \( \xi \) and concentrate on the \( y \)-coordinates.

To construct \( T_{i2} : S^-_1 \to S^+_2 \), similarly, we find the transition time \( t_2 \) for the trajectory starting at \((y_{10}, y_{20}, \xi_0)\) to arrive at the section \( S^+_2 \) under map \( T_{i2} \) with the restriction that \( y_2(t_2) = \delta \):

\[
t_2 = 1/\gamma_2 \ln(\delta/y_{20}).
\]

So we obtain the equations for \( T_{i2} \):

\[
y_1(t_2) = y_{10} (\delta/y_{20})^{y_1/y_2},
\]

\[
\xi(t_2) = \xi_0 e^{\Lambda t_2}.
\]

(3.9)

Considering the local maps \( T_{1j} : S^-_1 \to S^+_j \) where \( j = 1, 2 \), now we want to study the preimages of these maps in the section \( S^-_1 \). Because we concentrate on the \( y \)-coordinates, we study the projection of the domain of \( T_{1j} \) onto the \( y \)-plane which is denote by \( D_j \) where \( j = 1, 2 \). Similarly, for the local maps \( T_{2j} : S^-_2 \to S^+_j \) where \( j = 1, 2 \), denote the projection of the domain of \( T_{2j} \) onto the \( y \)-plane as \( \tilde{D}_j \) for \( j = 1, 2 \). We prove the following lemma.

**Lemma 3.4.1.** The projections of the domains of \( T_{11} \) and \( T_{12} \) onto the \( y \)-plane contain nonempty open sets, and they intersect only at the point \( y_1 = y_2 = 0 \). The same property is true for the projections of the domains of \( T_{21} \) and \( T_{22} \) onto the \( y \)-plane.
This lemma implies that $D_1 \cap D_2 = \{0\}$ and $\tilde{D}_1 \cap \tilde{D}_2 = \{0\}$ on the $y$-plane. The proof is based on the equations of the local maps in (3.8) and (3.9).

For the local map $T_{11}$ that maps $(y_{10}, y_{20}, \xi_0) \in S^-_1$ to $(\delta, y_2(t_1), \xi(t_1)) \in S^+_1$ as defined in (3.8), we need that $y_2(t_1) \leq \varepsilon$, which is equivalent to

$$y_{20} \leq \delta^{-\gamma_2/\gamma_1} \varepsilon (y_{10})^{\gamma_2/\gamma_1}.$$  \hspace{1cm} (3.10)

This inequality (3.10) gives us the region $D_1$.

We need to make sure that $D_1$ is inside the square $y_{10} \leq \varepsilon$ and $y_{20} \leq \varepsilon$. This can be easily done because if $y_{10} = \varepsilon$, then the right-hand side of (3.10) is $\delta^{-\gamma_2/\gamma_1} \varepsilon^{1+\gamma_2/\gamma_1}$. This implies that $y_{20}$ is less than $\varepsilon$ if $\varepsilon \ll 1$ because $\gamma_2/\gamma_1 > 0$. By Assumption A8, we also know that $0 < \gamma_2/\gamma_1 < 1$. We plot an illustration of $D_1$ on the $y$-plane in Figure 3.4.

![Figure 3.4](image)

**Figure 3.4:** The projection of the domains of $T_{11}$ and $T_{12}$ onto the $y$-plane. For fixed $i = 1$ or 2, $D_1$ is the projection of the domain of $T_{1i}$ onto the $y$-plane and $D_2$ is the projection of the domain of $T_{12}$ onto the $y$-plane. We see that $D_1$ and $D_2$ contain non-empty open sets and $D_1 \cap D_2 = \{0\}$. The figure looks the same for both $i = 1$ and $i = 2$.

Similarly, for the local map $T_{12}$ that maps $(y_{10}, y_{20}, \xi_0) \in S^-_1$ to $(y_1(t_2), \delta, \xi(t_2)) \in S^+_2$ as defined in (3.9), we need $y_1(t_2) \leq \varepsilon$. This implies the following inequality:

$$y_{20} \geq \delta e^{-\gamma_2/\gamma_1} (y_{10})^{\gamma_2/\gamma_1}.$$  \hspace{1cm} (3.11)
Note that \( 0 < \gamma_2 / \gamma_1 < 1 \), so the projection of the domain of \( T_{12} \) onto the \( y \)-plane looks as region \( D_2 \) in Figure 3.4 where
\[
\varepsilon_1 = \delta^{\gamma_1 / \gamma_2} \varepsilon^{1+\gamma_1 / \gamma_2} \leq \varepsilon \quad \text{if} \quad \varepsilon \ll 1.
\]

In order for \( D_1 \cap D_2 = \{0\} \), we need \( \delta \varepsilon^{-\gamma_2 / \gamma_1} > \delta^{-\gamma_2 / \gamma_1} \varepsilon \), which is true if \( \varepsilon \ll 1 \). Therefore, Lemma 3.4.1 is proved for \( D_1 \) and \( D_2 \).

The same calculations can be done for the local maps \( T_{21} \) and \( T_{22} \) to obtain the regions \( \tilde{D}_1 \) and \( \tilde{D}_2 \). Let us remark that \( \tilde{D}_1 \) and \( \tilde{D}_2 \) look the same as \( D_1 \) and \( D_2 \) respectively in Figure 3.4 because the equations for the corresponding local maps in (3.8) and (3.9) are the same. Therefore, we proved Lemma 3.4.1.

Now we consider the images of the local maps. Denote the image of the maps \( T_{ij} \) where \( i, j \in \{1, 2\} \) as \( R(T_{ij}) \). We prove the following lemma.

**Lemma 3.4.2.** For \( i = 1, 2 \), for small values of \( \varepsilon \), there exists \( \alpha > 0 \) such that the image of the local map \( R(T_{ii}) \) contains the rectangle \( R_1^\alpha = \{ (\xi, y_2) : |\xi| \leq \alpha, 0 < y_2 \leq \varepsilon \} \). Similarly, for \( i = 1, 2 \), the image \( R(T_{ij}) \) contains the rectangle \( R_2^\alpha = \{ (\xi, y_1) : |\xi| \leq \alpha, \xi_1 > 0, 0 < y_1 \leq \varepsilon \} \).

Consider, for instance, \( T_{12} : S_1^- \to S_2^+ \). We have
\[
\begin{align*}
y_1(t) &= y_{10} e^{\gamma_1 t}, \\
y_2(t) &= y_{20} e^{\gamma_2 t}, \\
\xi_1(t) &= \xi_{10} e^{\lambda_1 t}, \\
\xi_2(t) &= \xi_{20} e^{\lambda_2 t},
\end{align*}
\]
where \( \lambda_1 < 0 \) is the eigenvalue which is denoted as \( \lambda^\ll_{ii-1} \) in Assumption A3, which corresponding to the eigenvector whose direction is the \( \xi_1 \)-axis, and \( \lambda_2 < 0 \) is the eigenvalue which is denoted as \( \lambda^\ll_{ii-1} \) in Assumption A3, which corresponds to the eigenvector whose is the \( \xi_2 \)-axis.
We already knew in the calculation of $T_{i2}$ that the transition time from $S^+_1$ to $S^+_2$ is
\[ t_2 = 1/\gamma_2 \ln (\delta/\gamma_{20}), \]
then we can find that
\[ \xi_1(t_2) = \delta e^{\lambda_1 t_2} = \delta (\delta/\gamma_{20})^{\lambda_1/\gamma_2}. \]
Therefore,
\[ y_{20} = \delta^{\gamma_2/\lambda_1 + 1} [\xi_1(t_2)]^{1-\gamma_2/\gamma_1}. \]
We know that $-\gamma_2/\lambda_1 > 0$ by Assumption A5. We let $y_{20} \leq \varepsilon$ in the above equation and solve for $\xi_1(t_2)$ to obtain that
\[ 0 < \xi_1(t_2) \leq \varepsilon^{1-\gamma_2/\gamma_1} \delta^{1+\gamma_1/\gamma_2} =: \alpha_1. \]
We already know that for $T_{i2}$, $y_1(t_2) \leq \varepsilon$. Now for $k \neq 1$,
\[ \xi_k(t_2) = \xi_{k0} e^{\lambda_k t_2} = \xi_{k0} \delta^{\lambda_k/\gamma_2} y_{20}^{-\lambda_k/\gamma_2}. \]
Because $|\xi_{k0}| \leq \varepsilon$ and $|y_{20}| \leq \varepsilon$, we can find the upper bound on $|\xi_k|$:
\[ |\xi_k| \leq \varepsilon \delta^{\lambda_k/\gamma_2} \varepsilon^{-\lambda_k/\gamma_2} =: \alpha_k. \]
Thus, every point $(y_1, y_2, \xi)$ where $0 < \xi_1 \leq \alpha_1, |\xi_k| \leq \alpha_k(k \neq 1), 0 \leq y_1 \leq \varepsilon$ and $y_2 = \delta$ is the image of a point on $S^-_1$ under the map $T_{i2}$. Let us define
\[ \alpha = \min_{k=1,\ldots,LN-2} \{\alpha_k\}. \]
So we obtain the desired result in Lemma 3.4.2. The proof is the same for other local maps $T_{ij}$, where $i, j \in \{1, 2\}$.

### 3.4.2 Global maps

Denote the global map along the trajectories close to the heteroclinic orbit joining $Q'_i$ and $Q'_{i+1}$ as:
\[ T^g_i : S^+_1(Q'_i) \to S^-_1(Q'_{i+1}). \]
This global map $T_{gl}^1$ maps points in the section $S_1^+$ in a small neighborhood of $Q_{l_i}^i$ to points in the section $S_1^-$ in a small neighborhood of $Q_{l_{i+1}}^i$ near the heteroclinic orbit $P_{l_{i+1}}^i$ under the flow of system (3.1). It is well-defined because of the finite transition time from a small neighborhood of $Q_{l_i}^i$ to a small neighborhood of $Q_{l_{i+1}}^i$ along the trajectory near the corresponding heteroclinic orbit. We also know that it is a diffeomorphism because it is defined by solutions of (3.1), and the solutions are unique and depend differentially on initial conditions.

Similarly, the global map that maps points in the section $S_2^+$ in a small neighborhood of $Q_{l_i}^i$ to points in the section $S_2^-$ in a small neighborhood of $Q_{l_{i+1}}^i$ near the heteroclinic orbit $P_{l_{i+2}}^i$ under the flow of system (3.1) is well-defined and a diffeomorphism as well. Denote this global map as:

$$T_{gl}^2 : S_2^+(Q_{l_i}^i) \rightarrow S_2^-(Q_{l_{i+1}}^i).$$

To write the formulas for the global maps in the $(y, \xi)$-coordinates, one needs to use formulas for the sections $S_1^+(Q_{l_i}^i), S_1^-(Q_{l_{i+1}}^i), S_2^+(Q_{l_{i-1}}^i), S_2^-(Q_{l_{i+1}}^i), S_2^+(Q_{l_{i+2}}^i)$ in the original coordinates $\{x_{l_i}^j\}$ to integrate the system (3.1) over finite transition time, and then to present the images of the global maps $T_{gl}^1$ and $T_{gl}^2$ in the new $(y, \xi)$-coordinates. However, we do not need them. We claim that these $(y, \xi)$-formulas determine diffeomorphisms, which follows easily from the facts that changes of variables we used are diffeomorphisms (Assumption A9) and the global maps in the original $x_{l_i}^j$-coordinates are also diffeomorphisms. The main fact to be checked is that the points with non-negative values of the $y$-coordinates are mapped into those with the same property. This is true because:

(i) According to Assumption A2, for the system (3.13) the non-negative hyperoctant is an invariant set. Furthermore, each coordinate plane $(x_{l_i}^j, x_{l_{i+2}}^m)$-plane is invariant.
(ii) The invariance of the hyperoctant and coordinate planes are preserved under the new coordinate system \((y, \xi)\) at each \(Q'_l\) because the eigenspace and any subspaces of it are invariant.

(iii) According to Figure 3.3 and our discussion above Figure 3.3, the non-negative values of \(y_1\) correspond to non-negative values of \(x'_{i+1}\), and the non-negative values of \(y_2\) correspond to non-negative values of \(x'_{i+1}\). We also know the non-negative values of \(\xi_i\) where \(i \neq 3\) correspond to non-negative values of \(x'_m\) for some \((j, m) \neq (i, l)\). Only the coordinate \(\xi_3\) which corresponds to the \(x'_l\)-axis can be either negative, 0 or positive.

Therefore, we know that the maps \(T^{gl}_1\) and \(T^{gl}_2\) send points with non-negative values of the \((y_1, y_2, \xi_1, \xi_2, \xi_4, \ldots, \xi_{LN-2})\)-coordinates in a neighborhood of \(Q'_l\) into those with non-negative values of the \((x'_i, x'_{i+1}, \ldots, x'_m, \ldots)\)-coordinates where \((j, m) \neq (i, l)\) in neighborhoods of \(Q'_{i+1}\) and \(Q'^{i+1}_l\) correspondingly. In a small neighborhood of \(Q'_{i+1}\), those non-negative values of \(x'_m\) where \((j, m) \neq (i, l)\) correspond to the non-negative values of the new coordinates \((y_1, y_2, \xi_2, \xi_3, \ldots, \xi_{LN-2})\). Furthermore, according to the construction of the sections \(S^+_1(Q'_l)\) and \(S^-_1(Q'^{i+1}_l)\) transversal to the heteroclinic orbit \(P'_{i+1}\), we know that \(T^{gl}_1\) sends points on \(y_1 = \delta\) to \(\xi_1 = \delta\). Therefore, the global map sends points in the rectangle \(R^1_\alpha \subset S^+_1(Q'_l)\) to the points in a set with non-empty interior in \(S^-_1(Q'^{i+1}_l)\). Similarly, in a small neighborhood of \(Q'^{i+1}_l\), those non-negative values of \(x'_m\) where \((j, m) \neq (i, l)\) correspond to the non-negative values of the new coordinates \((y_1, y_2, \xi_1, \xi_3, \ldots, \xi_{LN-2})\), and the global map \(T^{gl}_2\) sends points on \(y_2 = \delta\) to \(\xi_2 = \delta\). Therefore, \(T^{gl}_2\) sends points in the rectangle \(R^2_\alpha \subset S^+_2(Q'_l)\) to points in a set with non-empty interior in \(S^-_2(Q'^{i+1}_l)\).

Hence, assuming that the diameters of \(S^+_1(Q'^{i+1}_l)\) and \(S^-_2(Q'^{i+1}_l)\) are small enough, we obtain that the sets

\[
R^1_\alpha \cap (T^{gl}_1)^{-1}(S^-_1(Q'^{i+1}_l)) \subset S^+_1(Q'_l) \quad \text{and} \quad R^2_\alpha \cap (T^{gl}_2)^{-1}(S^-_2(Q'^{i+1}_l)) \subset S^+_2(Q'_l) \tag{3.12}
\]
both contain non-empty interiors.

### 3.4.3 Images and preimages

It follows from (3.12) and Lemma 3.4.2 that the projection onto the $y$-plane of the set

$$(T_j)^{-1}\left(R_u \cap (T_1^{-1}(S_1^-(Q_{i+1}^j}))\right)$$

belongs to $D_j \subset S_1^-(Q_{i}^j)$ if $j = 1$, and it belongs to $\tilde{D}_j \subset S_2^-(Q_{i}^j)$ if $j = 2$. Denote this projection as $\hat{D}_j$ if $j = 1$, and denote it as $\hat{\tilde{D}}_j$ if $j = 2$. Similarly, the projection of the set

$$(T_2)^{-1}\left(R_u \cap (T_2^{-1}(S_2^-(Q_{i+1}^{i+1})))\right)$$

onto the $y$-plane, are denoted as $\hat{D}_2 \subset D_2$ if $j = 1$, and as $\hat{\tilde{D}}_2 \subset \tilde{D}_2$ if $j = 2$.

Therefore, a trajectory going through an initial point $(y_0, \xi_0)$ on $S_j^-(Q_{i}^j)$, $j = 1, 2$, with small values of the $\xi_0$-coordinates and $y_0 \in \hat{D}_1 \cup \hat{\tilde{D}}_1$ will intersect $S_1^+(Q_{i}^j)$ and then will stay near the heteroclinic orbit joining $Q_{i}^j$ and $Q_{i+1}^j$ following the flow, to the intersection with $S_1^-(Q_{i+1}^{i+1})$. Similarly, the trajectory going through $(y_0, \xi_0)$ on $S_j^+(Q_{i}^j)$ where $j = 1, 2$ and with small $|\xi_0|$ and $y_0 \in \hat{D}_2 \cup \hat{\tilde{D}}_2$ will intersect $S_2^+(Q_{i}^j)$ and will shadow the heteroclinic orbit joining $Q_{i}^j$ and $Q_{i+1}^j$ to the intersection with $S_2^-(Q_{i+1}^{i+1})$.

The same consideration can be performed in a small neighborhood of $Q_{i+1}^j$ and $Q_{i}^{i+1}$. For instance, for $Q_{i+1}^{i+1}$ we again have four local maps

$$\tilde{T}_{kj} : S_k^-(Q_{i+1}^j) \to S_m^+(Q_{i+1}^{i+1}), \text{ where } k, j = 1, 2.$$ 

To distinguish the local maps $T_{kj}$ in a small neighborhood of $Q_{i}^j$, we use the notation $\tilde{T}_{kj}$ for the local maps in a small neighborhood of $Q_{i+1}^j$. Denote the domains of the local maps $\tilde{T}_{kj}$ as dom$\tilde{T}_{kj}$ and the images of $\tilde{T}_{kj}$ as $R(\tilde{T}_{kj})$.

By Lemma 3.4.1, projections of dom$\tilde{T}_{1j} \subset S_1^-(Q_{i+1}^j)$ onto the $y$-plane look as $D_j$ for $j = 1, 2$ in Figure 3.4; projections of dom$\tilde{T}_{2j} \subset S_2^-(Q_{i+1}^j)$ onto the $y$-plane look as $\tilde{D}_j$ for $j = 1, 2$ which look the same as $D_j$ in Figure 3.4.
We first consider $k = 1$. Since $\text{dom} \tilde{T}_{1j} \subset S^-_1(Q^i_{i+1})$, the map $(T_{gl}^{11})^{-1}$ is well-defined on it. Thus on $S^-_1(Q^i_j)$ we have two regions $(T_{gl}^{11})^{-1}[((T_{gl}^{11})^{-1}\text{dom} \tilde{T}_{1j}) \cap R^1_{i_0})]$ and $(T_{gl}^{11})^{-1}[((T_{gl}^{11})^{-1}\text{dom} \tilde{T}_{12}) \cap R^1_{j_0})]$. Denote by $D_{11}$ and $D_{12}$ their projections onto the $y$-plane. Then on $S^-_1(Q^i_j)$ we have two regions $(T_{gl}^{11})^{-1}[((T_{gl}^{11})^{-1}\text{dom} \tilde{T}_{11}) \cap R^1_{\alpha_1})]$ and $(T_{gl}^{11})^{-1}[((T_{gl}^{11})^{-1}\text{dom} \tilde{T}_{12}) \cap R^1_{\alpha_2})]$. Denote by $D_{11}$ and $D_{12}$ their projections onto the $y$-plane.

For $k = 2$, we can also obtain the regions $(T_{gl}^{22})^{-1}[((T_{gl}^{22})^{-1}\text{dom} \tilde{T}_{2j}) \cap R^2_{i_0})]$ on $S^-_2(Q^i_j)$ for $j = 1, 2$ and denote their projections onto the $y$-plane as $D_{2j}$, which look as in Figure 3.5(b).

For $k = 2$, we also can obtain the regions $(T_{gl}^{22})^{-1}[((T_{gl}^{22})^{-1}\text{dom} \tilde{T}_{2j}) \cap R^2_{i_0})]$ on $S^-_2(Q^i_j)$ for $j = 1, 2$ and denote their projections onto the $y$-plane as $D_{2j}$, which look similarly as $D_{1j}$ for $i, j = 1, 2$.

Figure 3.5: Projections onto the $y$-plane of preimages of compositions of local and global maps. For fixed $i = 1, 2$, $D_1$ in (a) is the projection of the preimage of $T_{11}$ onto the $y$-plane and $D_2$ in (b) is the projection of the preimage of $T_{12}$ onto the $y$-plane as in Figure 3.4. In (a), $D_{11} \subset D_1$ is the projection of the map $\tilde{T}_{11} \circ T_{gl}^{11} \circ T_{11}$ onto the $y$-plane and $D_{12} \subset D_1$ is the projection of the map $\tilde{T}_{12} \circ T_{gl}^{11} \circ T_{11}$ onto the $y$-plane. In (b), $D_{21} \subset D_2$ is the projection of the map $\tilde{T}_{21} \circ T_{gl}^{11} \circ T_{21}$ onto the $y$-plane and $D_{22} \subset D_2$ is the projection of the map $\tilde{T}_{22} \circ T_{gl}^{11} \circ T_{21}$ onto the $y$-plane. We see that all sets contain nonempty open sets, $D_{jk} \subset D_j$ and $D_{j_1} \cap D_{j_2} = \{0\}$ for $j, k = 1, 2$. Both figures look the same for $i = 1, 2$, but when $i = 1$ these projections are projected from $S^-_1(Q^i_j)$ where $\xi_1 = \delta$ and when $i = 2$ these projections are projected from $S^-_2(Q^i_j)$ where $\xi_2 = \delta$. 
Therefore, if \((y_0, \xi_0)\) on \(S_1^-(Q_l^j) \cup S_2^-(Q_l^j)\) has small \(|\xi_0|\) and \(y_0 \in D_{\alpha\beta} \cup \tilde{D}_{\alpha\beta}\), then the corresponding trajectory intersects \(S_\alpha^+(Q_l^j)\) first, then follows the trajectory near the corresponding heteroclinic orbit and intersects \(S_\alpha^-(Q_{l+1}^i)\) if \(\alpha = 1\) or \(S_\alpha^-(Q_{l+1}^i)\) if \(\alpha = 2\). After this, the trajectory comes to a neighborhood of the heteroclinic orbit joining:

(i) the saddles \(Q_{l+1}^i\) and \(Q_{l+2}^i\) if \(\alpha = \beta = 1\);
(ii) the saddles \(Q_{l+1}^i\) and \(Q_{l+1}^{i+1}\) if \(\alpha = 1\) and \(\beta = 2\);
(iii) the saddles \(Q_{l+1}^{i+1}\) and \(Q_{l+1}^{i+1}\) if \(\alpha = 2\) and \(\beta = 1\);
(iv) the saddles \(Q_{l+1}^{i+1}\) and \(Q_{l+2}^i\) if \(\alpha = \beta = 2\).

Since these initial points are in the sets that have non-empty interiors

\[
U_{\alpha\beta} = \{(y_0, \xi_0) : (y_0, \xi_0) \in S_1^-(Q_l^j) \cup S_2^-(Q_l^j), |\xi_0| \leq \varepsilon, y_0 \in D_{\alpha\beta} \cup \tilde{D}_{\alpha\beta}\},
\]

we in fact proved Theorem 3.3.1 for cylinders \([\omega_0, \omega_1, \omega_2]\) where \(\omega_0\) corresponds to \(Q_l^j\); the symbol \(\omega_1\) corresponds to \(Q_{l+1}^i\) or \(Q_{l+1}^{i+1}\), and \(\omega_2\) corresponds \(Q_{l+2}^i\), \(Q_{l+1}^{i+1}\), or \(Q_{l+2}^i\).

Exactly in the same way, one can prove Theorem 3.3.1 for an arbitrary cylinder \([\omega_0, \omega_1, \ldots, \omega_k]\).

### 3.5 Discussion

#### 3.5.1 An example of the system with the heteroclinic network \(\Gamma\)

We proved Theorem 3.3.1 for a general dynamical system (3.1) with a heteroclinic network satisfying Assumptions A1-A9. For such a dynamical system, one can expect that a trajectory that enters a small neighborhood of a saddle point might stay in this neighborhood for a duration of time, then leave this neighborhood following a heteroclinic orbit, and go into a neighborhood of another saddle point. Thus, the phenomenon of winnerless competition (WLC) might be observed in such a system. In ecological systems
with predator-prey interactions, WLC is the phenomenon where one species, whose population is described by one of the equations, dominates for a while, then decreases in population and another species dominates the environment; this process repeats endlessly and no species can dominate for an infinitely long time. The transient dynamics in WLC sequentially switch among the temporary “winners.” Similar patterns of activity have also been observed in collections of neurons in the human brain [3, 4, 33] and in central pattern generator in invertebrates [35].

In this subsection, we give an example of the system that a heteroclinic network might exist in its phase space. This system is a high-dimensional ODE system in the form of generalized Lotka-Volterra equations:

\[
\frac{dx_i^l}{dt} = x_i^l \left( \sigma_i^l - \sum_{j=1}^{N} \rho_{ij}^l x_j^l - \sum_{m=1}^{L} \sum_{j=1}^{N} \xi_{ij}^{lm} x_j^m \right) \quad \text{for} \quad i = 1, \ldots, N; l = 1, \ldots, L, \tag{3.13}
\]

where \(X = (\ldots, x_i^l, \ldots) \in \mathbb{R}_+^{LN}\). This system has dimension \(LN\) with all variables \(x_i^l \geq 0\) and all parameters \(\sigma_i^l \geq 0, \rho_{ij}^l \geq 0\) and \(\xi_{ij}^{lm} \geq 0\). We assume \(\rho_{ii}^l = 1\) and \(\xi_{ii}^{ll} = 0\) for any \(i\) and \(l\). Authors of [3, 4] have shown that it is common to observe the WLC phenomenon in a high-dimensional system of generalized Lokta-Volterra equations.

For this system (3.13), we can find some of the equilibrium points, which are

\(Q_i^l = \{x_i^l = \sigma_i^l, x_j^m = 0 \text{ if } j \neq i, \text{ or } m \neq l\}\). We see that Assumption A1 is satisfied. Because of the particular form of (3.13), we can also see that Assumption A2 is satisfied.

For each \(Q_i^l\) where \(i = 1, \ldots, N\) and \(l = 1, \ldots, L\), the eigenvalues and the corresponding eigenvectors are given in Tables 3.1 and 3.2. We can see from these tables that Assumption A3 is satisfied. The authors in [32] proposed this model and studied the existence of the heteroclinic network \(\Gamma\). For the eigenvalues in Table 3.1, it is generally believed that there is an open set of parameters where Assumptions A4, A5, A7 and A8 hold. But further study is needed to show whether Assumption A6 holds under the conditions A1-A8.
Table 3.1: Eigenvalues at the equilibrium point \( Q_l^i \).

<table>
<thead>
<tr>
<th>Eigenvalues</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_{ij}^l = \sigma_j^l - \sigma_j^l(\xi_{ji}^l + \rho_{ji}^l) )</td>
<td>( 1 \leq j \leq N, j \neq i, i + 1 )</td>
</tr>
<tr>
<td>( \lambda_{ii}^l = -\sigma_i^l )</td>
<td></td>
</tr>
<tr>
<td>( \lambda_{i+1}^l = \sigma_{i+1}^l - \sigma_{i+1}^l(\xi_{i+1i}^l + \rho_{i+1i}^l) )</td>
<td></td>
</tr>
<tr>
<td>( \lambda_{m}^l = \sigma_m^l - \sigma_m^l(\xi_{ji}^m + \rho_{ji}^m) )</td>
<td>( 1 \leq m \leq L, 1 \leq j \leq N, m \neq l, (j, m) \neq (i, l + 1) )</td>
</tr>
</tbody>
</table>

Table 3.2: Eigenvectors at the equilibrium point \( Q_l^i \) corresponding to the eigenvalues in Table 3.1.

<table>
<thead>
<tr>
<th>Eigenvectors</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \vec{v}<em>{ij}^l = {x_i^l = \xi</em>{ij}^l + \rho_{ij}^l, x_j^l = \xi_{ji}^l + \rho_{ji}^l - 1 - \sigma_j^l/\sigma_i^l, \text{others } = 0} )</td>
<td></td>
</tr>
<tr>
<td>( \vec{v}_{ii}^l = {x_i^l = 1, \text{others } = 0} )</td>
<td></td>
</tr>
<tr>
<td>( \vec{v}<em>{i+1}^l = {x_i^l = \xi</em>{i+1i}^l + \rho_{i+1i}^l, x_{i+1}^l = \xi_{i+1i}^l + \rho_{i+1i}^l - 1 - \sigma_{i+1i}/\sigma_i^l, \text{others } = 0} )</td>
<td></td>
</tr>
<tr>
<td>( \vec{v}<em>{i+1}^l = {x_i^l = \xi</em>{i+1i}^l, x_{i+1}^l = \xi_{i+1i}^l - 1 - \sigma_{i+1i}/\sigma_i^l, \text{others } = 0} )</td>
<td></td>
</tr>
<tr>
<td>( \vec{v}<em>{ij}^m = {x_i^l = \xi</em>{ij}^m, x_j^l = \xi_{ji}^m - 1 - \sigma_m^l/\sigma_j^l, \text{others } = 0} )</td>
<td></td>
</tr>
</tbody>
</table>

3.5.2 A modified heteroclinic network \( \tilde{\Gamma} \)

For the system (3.1) with the heteroclinic network, we can modify the assumptions on the heteroclinic network \( \Gamma \) so that the last heteroclinic cycle \( \Gamma_L \) is connected to the first heteroclinic cycle \( \Gamma_1 \) by heteroclinic orbits. In this case, the network in the phase space is a heteroclinic cycle of heteroclinic cycles. Denote such a heteroclinic network by \( \tilde{\Gamma} \) consisting of heteroclinic cycles \( \Gamma_i \) and the cyclic heteroclinic orbits connecting them. In order for \( \tilde{\Gamma} \) to exist, we modify the assumption A5 to A5’ as follows.

A5’: There is a heteroclinic orbit connecting the saddle \( Q_l^i \) to the saddle \( Q_l^{i+1} \) for

\[ i = 1, \ldots, N \text{ and } l = 1, \ldots, L \text{ where } L + 1 \equiv 1. \]

Then the eigenvalues in Assumption A3(c) is changed to \( \lambda_{ii}^{L+1} > 0 \) for all \( i = 1, \ldots, N \).
We can write such a heteroclinic network satisfying A1-A4, A5' and A6 as

\[
\tilde{\Gamma} = \bigcup_{i=1}^{L} \bigg( \Gamma_i \bigcup_{j=1}^{N} P_{i,j}^l \bigg).
\]

For such a heteroclinic network \(\tilde{\Gamma}\), we can find a directed graph \(\tilde{G}\) which is identified with \(\tilde{\Gamma}\). Using the same labels of vertices of the graph as in Section 3.1, we can obtain the adjacency matrix \(\tilde{M}\) corresponding to the graph \(\tilde{G}\). A new subshift of finite type \((\Omega_{\tilde{M}}, \sigma)\) associated with \(\tilde{M}\) can be defined analogously as \((\Omega_M, \sigma)\) in Section 3.1. We can study the number of possible heteroclinic routes in this subshift in a similar way as Lemma 3.2.1. In fact, we can prove the following lemma:

**Lemma 3.5.1.** For the subshift of finite type \((\Omega_{\tilde{M}}, \sigma)\) associated with the heteroclinic network \(\tilde{\Gamma}\), the symbolic complexity function is \(C_n = LN2^{n-1}\) where \(L\) is the number of heteroclinic cycles and \(N\) is the number of saddles in every heteroclinic cycle.

The proof of Lemma 3.5.1 is similar to that of Lemma 3.2.1. The transition matrix \(\tilde{M}\) is an \(L \times L\) block matrix in the following form:

\[
\tilde{M} = \begin{pmatrix}
B & I & 0 & \cdots & 0 \\
0 & B & I & 0 & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & B & I & \vdots \\
I & \cdots & \cdots & 0 & B
\end{pmatrix},
\]

where \(I, B\) and \(0\) are the same matrices as in Section 3.3.
We can calculate that

\[
\tilde{M}^{L-1}
\]

\[
= \begin{pmatrix}
B^{-1} & \begin{pmatrix} L - 1 \\ 1 \end{pmatrix} & B^{L-2} & \begin{pmatrix} L - 1 \\ 2 \end{pmatrix} & B^{L-3} & \ldots & \begin{pmatrix} L - 1 \\ L - 2 \end{pmatrix} & B & I \\
I & B^{-1} & \begin{pmatrix} L - 1 \\ 1 \end{pmatrix} & B^{L-2} & \begin{pmatrix} L - 1 \\ L - 3 \end{pmatrix} & \ldots & \begin{pmatrix} L - 1 \\ L - 2 \end{pmatrix} & B & I \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\begin{pmatrix} L - 1 \\ L - 2 \end{pmatrix} & \begin{pmatrix} L - 1 \\ 1 \end{pmatrix} & B^{n-2} & \begin{pmatrix} L - 1 \\ 2 \end{pmatrix} & B^{L-3} & \ldots & \ldots & I & B^{L-1}
\end{pmatrix}
\]

For \( n > L \), \( \tilde{M}^{n-1} \) is more complicated to calculate. We can see from above that each row of \( \tilde{M}^{L-1} \) contains the same block matrices but they are shifted by one block in each row. Because of the structure of \( \tilde{M} \), this property is true for \( \tilde{M}^{n-1} \) with all \( n \). Thus, we can use only the first row of \( \tilde{M}^{n-1} \) to calculate \( u \tilde{M}^{n-1} u^T \). Denote the first row of the matrix \( \tilde{M}^{n-1} \) as \( a_{n-1} \), which is actually an \( N \times NL \) matrix containing the corresponding block matrices, then we can obtain that

\[
u \tilde{M}^{n-1} u^T = L u_N a_{n-1} u_N^T,
\]

(3.14)

where \( u_N = (1, 1, \ldots, 1) \) is a row vector with length \( N \).
Because $a_{n-1} = a_{n-2}\bar{M}$ for all $n$, we can find that

$$a_L = \begin{pmatrix} B^{L-1}, & \begin{pmatrix} L - 1 \\ 1 \end{pmatrix} B^{L-2}, & \ldots, & \begin{pmatrix} L - 1 \\ j \end{pmatrix} B^{L-j-1}, & \ldots, & I \end{pmatrix} \begin{pmatrix} B & I & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ I & \cdots & 0 & B \end{pmatrix}$$

$$= \begin{pmatrix} B^L + I, & \begin{pmatrix} L - 1 \\ 1 \end{pmatrix} B^{L-1} + B^{L-1}, & \ldots, & \begin{pmatrix} L - 1 \\ j \end{pmatrix} B^{L-j} + \begin{pmatrix} L - 1 \\ j - 1 \end{pmatrix} B^{L-j}, & \ldots, & B + \begin{pmatrix} L - 1 \\ L - 2 \end{pmatrix} B \end{pmatrix}. $$

Note that for the binomial coefficients, the following identity is true for any integer $m$ and $1 \leq j \leq m - 1$:

$$\begin{pmatrix} m - 1 \\ j \end{pmatrix} + \begin{pmatrix} m - 1 \\ j - 1 \end{pmatrix} = \begin{pmatrix} m \\ j \end{pmatrix}.$$

Then we have

$$a_L = \begin{pmatrix} B^L + I, & \begin{pmatrix} L \\ 1 \end{pmatrix} B^{L-1}, & \ldots, & \begin{pmatrix} L \\ j \end{pmatrix} B^{L-j}, & \ldots, & \begin{pmatrix} L \\ L - 1 \end{pmatrix} B \end{pmatrix}.$$

Because $B$ is the permutation matrix, $u_N B^k u_N^T = N$ for any $k$, we can obtain that

$$u_N a_L u_N^T = N + \begin{pmatrix} L \\ 1 \end{pmatrix} N + \cdots + \begin{pmatrix} L \\ j \end{pmatrix} N + \cdots + \begin{pmatrix} L \\ L - 1 \end{pmatrix} N$$

$$= N \begin{pmatrix} L \\ 0 \end{pmatrix} + \begin{pmatrix} L \\ 1 \end{pmatrix} + \cdots + \begin{pmatrix} L \\ L \end{pmatrix}$$

$$= N2^L. $$

Similarly,

$$a_{L+1} = \begin{pmatrix} L + 1 \\ L \end{pmatrix} B + B^{L+1}, \; I + \begin{pmatrix} L + 1 \\ 1 \end{pmatrix} B^L, \; \ldots, \; \begin{pmatrix} L + 1 \\ j \end{pmatrix} B^{L-j+1}, \; \ldots, \; \begin{pmatrix} L + 1 \\ L - 1 \end{pmatrix} B^2, \; \ldots, $$
and we have
\[
\begin{align*}
\mathbf{X}_n^a & = \mathbf{N} \left[ \begin{array}{ccc} L+1 & L+1 & \cdots & L+1 \\ 0 & 1 & & \\ & 1 & & \\ & & & \end{array} \right] + \cdots + \left[ \begin{array}{ccc} L+1 & L+1 & \cdots & L+1 \\ 0 & 1 & & \\ & 1 & & \\ & & & \end{array} \right] \\
& = \mathbf{N}^2 \mathbf{L}^{L+1}.
\end{align*}
\]

Therefore, even though it is complicated to obtain the exact expression for \( a_{n-1} \) for \( n > L \), we find that \( a_{n-1} \) contains only the terms
\[
\left[ \begin{array}{ccc} n-1 & B^{n-1} & \cdots & B \end{array} \right], \quad \left[ \begin{array}{ccc} n-1 & B^{n-2} & \cdots & 1 \end{array} \right], \quad \left[ \begin{array}{ccc} n-1 & B & \cdots & 1 \end{array} \right], \quad \left[ \begin{array}{ccc} n-1 & B & \cdots & 1 \end{array} \right].
\]

Therefore, \( \mathbf{X}_n^a = \mathbf{N}^2 \mathbf{L}^{n-1} \). By Equation (3.14), we can conclude that
\[
\mathbf{X}_n^a = \mathbf{N}^2 \mathbf{L}^{n-1}.
\]

So we proved Lemma 3.5.1 and showed that the symbolic complexity function is
\[
C_n = \mathbf{L} \mathbf{N}^2 \mathbf{L}^{n-1}.
\]

From this result, we see that the topological entropy is
\[
h = \lim_{n \to \infty} \frac{\ln \mathbf{C}_n}{n} = \lim_{n \to \infty} \frac{\ln \mathbf{L} \mathbf{N}^2 + (n-1) \ln 2}{n} = \ln 2.
\]

Thus, chaotic behavior might exist in this subshift of finite type \((\Omega, \sigma)\). The corresponding heteroclinic network \( \Gamma \) could be a subject for future studies.

### 3.6 Summary

In this chapter, we studied dynamical systems with a heteroclinic network satisfying Assumptions A1-A9. This heteroclinic network in the phase space is a heteroclinic chain of heteroclinic cycles. Starting from each saddle point that is not in the last heteroclinic cycle, there are two heteroclinic orbits, one stays in its heteroclinic cycle and the other one goes to the next heteroclinic cycle; there is only one heteroclinic orbit starting from each saddle point in the last heteroclinic cycle.
We identified with such a heteroclinic network a directed graph and constructed a subshift of finite type based on whether there is a heteroclinic orbit (directed edge) between the two saddles (vertices). We showed that the symbolic complexity of such a subshift is a polynomial with degree $L - 1$, where $L$ is the number of heteroclinic cycles in the network (Lemma 3.2.1). We also showed that given a sequence of the heteroclinic orbits that can be realized in this heteroclinic network, there is an open set of initial conditions in every neighborhood of the beginning saddle point such that trajectories going through this open set shadows the sequence of the heteroclinic orbits (Theorem 3.3.1).
4 Non-symmetric Critical Points in Tensor Approximations

The work in this chapter is joint work with Dr. Martin Mohlenkamp. Section 4.1 is
based on excerpts from Dr. Mohlenkamp’s previous work. He proposed the problem,
introduced the notations and formulated all the equations in Section 4.1. The remainder of
this chapter, Sections 4.2 to 4.5, is my work.

4.1 Introduction

4.1.1 Background

Numerical computations for functions of many variables may require huge
computational cost and suffer from an effect called the curse of dimensionality. For example, let $f : \mathbb{R}^d \to \mathbb{R}$ be a function with dimension $d$. By dimension, we mean the
number of variables in the function. We can write this function of $d$ variables as
$f(x_1, \ldots, x_d)$ where each $x_i \in \mathbb{R}$. If $M$ data points are used in each coordinate $x_i$, then there
are a total of $M^d$ data points and $M^d$ values of the function $f$ that need to be stored in a
computer. We see that the total number of stored entries grows exponentially as the
dimension $d$ increases. This effect is referred to as the curse of dimensionality [7].

One way to avoid the curse of dimensionality is to approximate the multivariate
function as a sum of separable functions. We can write this approximation as

$$f(x_1, x_2, \ldots, x_d) \approx \sum_{l=1}^{r} \prod_{i=1}^{d} g^l_i(x_i) \quad \text{for } x_i \in \mathbb{R},$$

(4.1)

where each $g^l_i : \mathbb{R} \to \mathbb{R}$ is a single-variable function. For this sum of separable function on
the right-hand side, the number of entries that need to be stored is $rM^d$ if there are $M$
values for each variable $x_i$. For this function, the number of stored entries grows linearly
as $d$ increases. When $d$ is relatively large and $r$ is relatively small, we have $rM^d \ll M^d$. In
this way, we can avoid the curse of dimensionality.
A tensor $T$ in dimension $d$ is a discrete representation of a function in dimension $d$ on a rectangular domain. We write it as $T = T(j_1, j_2, \ldots, j_d)$, where $j_i = 1, \ldots, M_i$. In this expression, $M_i$ is the number of allowed values in the index $j_i$. We call it the \textit{resolution} in the index $j_i$. The number of indices is the \textit{dimension} of a tensor. Thus, generally speaking, a tensor is an object with some number of indices and its value at given indices is a real or complex number. A vector has only one index, so it is a tensor in dimension 1; while a matrix has two indices, so it is a tensor in dimension 2.

Because a tensor can be a discrete representation of a function, the approximation of a tensor $T$ by a sum of separable tensors $G$ is similar to (4.1). We write it in the following form:

$$T(j_1, j_2, \ldots, j_d) \approx G(j_1, j_2, \ldots, j_d) = \sum_{l=1}^{r} \prod_{i=1}^{d} G^l_i(j_i) \quad \text{for} \quad j_i = 1, 2, \ldots, M_i,$$

where $r$ is called the \textit{separation rank} of tensor $G$. Note that in order to avoid the curse of dimensionality, the rank $r$ cannot grow exponentially as $d$ increases. Same as the approximation of multivariate function using sum of separable functions, the value of $r$ should be relatively small.

The author in [25] showed that the problem (4.2) is a special case of (4.1) in terms of their essential multilinear structures. In this chapter, we study the tensor approximation problem (4.2). For simplicity, we write the sum of separable tensors $G$ as

$$G = \sum_{l=1}^{r} G^l = \sum_{l=1}^{r} \bigotimes_{i=1}^{d} G^l_i,$$

where $\bigotimes$ is the tensor product. We call $T$ the “target” tensor and $G$ the “guess” tensor.

Throughout this chapter, for any two complex-valued tensors $F$ and $G$ indexed by $j_i = 1, 2, \ldots, M_i$ for $i = 1, \ldots, d$, the inner product is defined as

$$\langle F, G \rangle = \sum_{j_1=1}^{M_1} \cdots \sum_{j_d=1}^{M_d} F(j_1, \ldots, j_d) G(j_1, \ldots, j_d).$$

The norm of a tensor $F$ is defined as $\|F\| = \sqrt{\langle F, F \rangle}$. This is the Frobenius norm.
To approximate the target tensor $T$ with the guess tensor $G$, we would like to minimize the regularized error, which is defined as

$$E_{\lambda}(G) = \|T - G\|^2 + \lambda \sum_{i=1}^{r} \|G_i\|^2 \quad \text{for} \quad \lambda \geq 0. \quad (4.3)$$

For a fixed rank $r$, we want to go over all choices of $\{G_i\}$ to find the minimum of $E_{\lambda}(G)$. Then we can update $r$ and find another set of $\{G_i\}$ that minimizes $E_{\lambda}(G)$. By comparing the minimum values of $E_{\lambda}(G)$ at different $r$ values, we can find the optimal $r$ and $\{G_i\}$ which yield the best approximation $G$. This general tensor approximation is a very broad problem. In this chapter, we will study a special case of the tensor approximation problems.

### 4.1.2 The exemplar target tensors and the rank-1 guess tensors

In this chapter, we study the approximation of an exemplar rank-2 target tensor $T$ using a rank-1 guess tensor $G_1$. This allows us to study the situation when the rank is underestimated.

According to (4.2), a rank-1 guess tensor is written as

$$G_1 = \prod_{i=1}^{d} G_i(j_i), \quad j_i = 1, 2, \ldots, M_i. \quad (4.4)$$

For simplicity, we let $M_i = M$ for all $i = 1, \ldots, d$. We can see that $G_1$ is a separable tensor in dimension $d$.

In order to write a general form of $G_1$, we let $e_1$ and $e_2$ be orthogonal unit vectors in $\mathbb{R}^2$. We also define the unit vector with angle $\theta$ in $\mathbb{R}^2$ to be $u(\theta) = \cos \theta e_1 + \sin \theta e_2$. Note that for a vector of angles $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, the tensor product $\bigotimes_{i=1}^{d} u(\alpha_i)$ results in a unit separable tensor with dimension $d$. This is true because of the concept of tensor product.
For example, let \( \mathbf{e}_1 = [1, 0] \) and \( \mathbf{e}_2 = [0, 1] \). When \( d = 2 \), we have

\[
\mathbf{u}(\alpha_1) \bigotimes \mathbf{u}(\alpha_2) = [\cos \alpha_1, \sin \alpha_1] \bigotimes [\cos \alpha_2, \sin \alpha_2] = \begin{bmatrix}
\cos \alpha_1 \cos \alpha_2 & \cos \alpha_1 \sin \alpha_2 \\
\sin \alpha_1 \cos \alpha_2 & \sin \alpha_1 \sin \alpha_2
\end{bmatrix}.
\]

We can also verify that \( \|\mathbf{u}(\alpha_1) \bigotimes \mathbf{u}(\alpha_2)\| = 1 \) by calculating the Frobenius norm of the above matrix. Thus, \( \mathbf{u}(\alpha_1) \bigotimes \mathbf{u}(\alpha_2) \) is a unit separable tensor in dimension 2. Similarly, \( \mathbf{u}(\alpha_1) \bigotimes \mathbf{u}(\alpha_2) \bigotimes \mathbf{u}(\alpha_3) \) is a unit separable tensor in dimension 3, and \( \bigotimes_{i=1}^{d} \mathbf{u}(\alpha_i) \) is a unit separable tensor in dimension \( d \).

Because the rank-1 guess tensor is a separable tensor in dimension \( d \) according to (4.4), we can write the general form of \( G_1 \) as

\[
G_1 = a \bigotimes_{i=1}^{d} \mathbf{u}(\alpha_i),
\]

where \( a \) is a scalar that corresponds to the norm of \( G_1 \).

The exemplar rank-2 target tensor we want to study is

\[
T = \frac{\bar{T}}{\|\bar{T}\|} = \left(1 + 2z \prod_{i=1}^{d} \cos(\phi_i) + z^2 \right)^{-1/2} \left( \bigotimes_{i=1}^{d} \mathbf{u}(0) + z \bigotimes_{i=1}^{d} \mathbf{u}(\phi_i) \right).
\]

Note that \( \|T\| = 1 \). We consider the parameters \( |z| \leq 1, z \neq 0 \) and \( \phi_i \in [0, \pi/2] \). We do not let \( z = 0 \) because the target tensor \( T \) is only a rank-1 tensor when \( z = 0 \). We also cannot have \( \phi_i = 0 \) for all \( i \) because this reduces the rank of \( T \) to 1 as well. We choose this exemplar target tensor \( T \) because any rank-2 tensor can be transformed into this form by applying a separable unitary change of basis [24]. To study the approximation of the symmetric target tensor, we let \( \phi_i = \phi \) for all \( i = 1, \ldots, d \) and \( \phi \in (0, \pi/2] \).

4.1.3 Regularized error function

When using the rank-1 guess tensor \( G_1 \) to approximate the exemplar rank-2 symmetric target tensor \( T \), we would like to minimize the regularized error function defined in (4.3). For this particular approximation problem, the rank of the guess tensor
\( r = 1 \) in (4.3). So we can write the regularized error function as

\[
E_\lambda(G_1) = \|T - G_1\|^2 + \lambda\|G_1\|^2, \tag{4.7}
\]

where \( \lambda \geq 0 \).

When approximating \( T \) using \( G_1 = a \bigotimes u(\alpha_i) \), we need to minimize \( E_\lambda(G_1) \) over all choices of the scalar \( a \) and the angles \( \{\alpha_i\}_{i=1}^d \). For any numerical algorithm that minimizes \( E_\lambda(G_1) \), it would be complicated to update both the scalar and the angles simultaneously in each iteration. However, when the angles \( \{\alpha_i\}_{i=1}^d \) are fixed, the problem of finding an optimal scalar \( a \) that minimizes \( E_\lambda(G_1) \) is a simple linear least-squares problem. We explain this as follows:

In the language of linear algebra, for a linear system \( Ax = b \) that does not have solutions (usually the system is overdetermined), the least-squares approximation is to “solve” the linear system \( Ax = b \) by minimizing the norm of the residual \( b - Ax \) \[39\]. This is a least-squares approximation of the vector \( b \) with a vector \( Ax \) that minimizes \( \|b - Ax\| \).

For fixed \( \{\alpha_i\}_{i=1}^d \), we want to find the optimal \( a \) that minimizes the regularized error in (4.7). This approximation can be viewed as the least-squares approximation of the vector \((T, 0)\) with \((G_1, \sqrt{\lambda}G_1)\), because the squared norm of the residual is

\[
\|(T, 0) - (G_1, \sqrt{\lambda}G_1)\|^2 = \|(T - G_1, -\sqrt{\lambda}G_1)\|^2 = \|T - G_1\|^2 + \lambda\|G_1\|^2 = E_\lambda(G_1).
\]

So the corresponding linear system is

\[
\begin{bmatrix}
\bigotimes_{i=1}^d u(\alpha_i) \\
\sqrt{\lambda} \bigotimes_{i=1}^d u(\alpha_i)
\end{bmatrix} a = \begin{bmatrix} T \\ 0 \end{bmatrix}. \tag{4.8}
\]

Multiplying both sides of (4.8) from the left by the row vector \( \left[ \bigotimes_{i=1}^d u(\alpha_i), \sqrt{\lambda} \bigotimes_{i=1}^d u(\alpha_i) \right] \), we obtain the normal equation

\[
(1 + \lambda)a = \left[ \bigotimes_{i=1}^d u(\alpha_i), T \right].
\]
Therefore,
\[
a = \left( \bigotimes_{i=1}^{d} \mathbf{u}(\alpha_i), T \right) / (1 + \lambda). \tag{4.9}
\]
So we can treat the scalar $a$ as a “fast” variable that instantly optimizes itself according to (4.9) when the angles $\{\alpha_i\}_{i=1}^{d}$ are updated by an algorithm. In this way the regularized error function $E_\lambda(G_1)$ only depends on $\{\alpha_i\}_{i=1}^{d}$. We find the expression for $E_\lambda(G_1)$ as a function of the angular variables in the following way.

Because this approximation can be viewed as the least-squares approximation of the vector $(T, 0)$ with $(G_1, \sqrt{\lambda}G_1)$, we know that
\[
\left( (T, 0) - (G_1, \sqrt{\lambda}G_1), (G_1, \sqrt{\lambda}G_1) \right) = 0.
\]
This implies that
\[
\langle T, G_1 \rangle = ||G_1||^2 + \lambda ||G_1||^2.
\]
Thus, we can simplify the regularized error function in (4.7) to
\[
E_\lambda(G_1) = ||T||^2 - \langle T, G_1 \rangle - \langle G_1, T \rangle + ||G_1||^2 + \lambda ||G_1||^2
= 1 - (||G_1||^2 + \lambda ||G_1||^2)
= 1 - (1 + \lambda)||G_1||^2.
\]
According to (4.5) and the fact that each $\mathbf{u}(\alpha_i)$ is a unit vector, we know that
\[
||G_1||^2 = a^2 \prod_{i=1}^{d} ||\mathbf{u}(\alpha_i)||^2 = a^2.
\]
Thus,
\[
E_\lambda(G_1) = 1 - (1 + \lambda)a^2. \tag{4.10}
\]
Defining $n(\vec{a}) = \left( \bigotimes_{i=1}^{d} \mathbf{u}(\alpha_i), T \right)$ and plugging (4.9) into (4.10), we can write the regularized error as a function of the angular variables:
\[
E_\lambda(G_1) = 1 - \frac{n^2(\vec{a})}{1 + \lambda}. \tag{4.11}
\]
The range of this regularized error function is $[0, 1]$ because $n^2(\vec{a}) \in [0, 1]$. So this is simultaneously the absolute error and the relative error.
4.1.4 Notations

In this subsection, we introduce the notations for some functions that we will use very often in the calculations. For the rank-1 separable tensors with norm 1, we denote their inner product as

\[ p(\vec{\alpha} - \vec{\beta}) = \left( \bigotimes_{i=1}^{d} u(\alpha_i), \bigotimes_{i=1}^{d} u(\beta_i) \right) = \prod_{i=1}^{d} \cos(\alpha_i - \beta_i). \]  (4.12)

Note that the function \( p \) only depends on the difference of vectors \( \vec{\alpha} - \vec{\beta} \).

For a rank-1 separable tensor with norm 1 and the target tensor \( T \), we denote their inner product as

\[ n(\vec{\alpha}) = \left( \bigotimes_{i=1}^{d} u(\alpha_i), T \right). \]

By (4.12) and (4.6), we obtain that

\[ n(\vec{\alpha}) = p(\vec{\alpha}) + zp(\vec{\alpha} - \vec{\phi}) \left( 1 + 2zp(\vec{\phi}) + z^2 \right)^{1/2}. \]  (4.13)

The partial derivatives of the above functions \( p(\vec{\alpha}) \) and \( n(\vec{\alpha}) \) are

\[ p_j(\vec{\alpha}) = \frac{\partial}{\partial \alpha_j} p(\vec{\alpha}) = -\sin \alpha_j \prod_{i=1, i \neq j}^{d} \cos \alpha_i, \]  (4.14)

and

\[ n_j(\vec{\alpha}) = \frac{\partial}{\partial \alpha_j} n(\vec{\alpha}) = \frac{p_j(\vec{\alpha}) + zp_j(\vec{\alpha} - \vec{\phi})}{\left( 1 + 2zp(\vec{\phi}) + z^2 \right)^{1/2}}. \]  (4.15)

For \( j \neq k \), we can find the second order derivatives:

\[ p_{jk}(\vec{\alpha}) = \frac{\partial}{\partial \alpha_k} p_j(\vec{\alpha}) = \sin \alpha_j \sin \alpha_k \prod_{i=1, i \neq j, k}^{d} \cos \alpha_i, \]  (4.16)

and

\[ n_{jk}(\vec{\alpha}) = \frac{\partial}{\partial \alpha_k} n_j(\vec{\alpha}) = \frac{p_{jk}(\vec{\alpha}) + zp_{jk}(\vec{\alpha} - \vec{\phi})}{\left( 1 + 2zp(\vec{\phi}) + z^2 \right)^{1/2}}. \]  (4.17)

When \( j = k \), we can calculate that

\[ p_{jj}(\vec{\alpha}) = \frac{\partial}{\partial \alpha_j} p_j(\vec{\alpha}) = -\cos \alpha_j \prod_{i=1, i \neq j}^{d} \cos \alpha_i = -p(\vec{\alpha}), \]  (4.18)
and
\[ n_{jj}(\vec{\alpha}) = \frac{\partial}{\partial \alpha_j} n_j(\vec{\alpha}) = \frac{p_{jj}(\vec{\alpha}) + zp_{jj}(\vec{\alpha} - \vec{\phi})}{1 + 2zp(\vec{\phi}) + z^2} = -n(\vec{\alpha}). \] (4.19)

Note that in all the above notations, \( \vec{\phi} = (\phi_1, \phi_2, \ldots, \phi_d) \). So all the notations can be used for general rank-2 target tensors; they do not need to be symmetric. But in this chapter, we only consider the symmetric rank-2 target tensors, so we use \( \vec{\phi} = (\phi, \phi, \ldots, \phi) \) in Equations (4.13), (4.15), and (4.17) throughout the whole chapter.

### 4.1.5 Gradient flow

In order to approximate \( T \) using \( G_1 \) by minimizing \( E_\lambda(G_1) \), we consider how the angular variables change according to the gradient flow, which has the form
\[ \frac{d}{dt} \alpha_j = -\frac{\partial}{\partial \alpha_j} E_\lambda(G_1). \] (4.20)

By Equation (4.11), we know that
\[ -\frac{\partial}{\partial \alpha_j} E_\lambda(G_1) = \frac{\partial}{\partial \alpha_j} n_j^2(\vec{\alpha}) = \frac{2n(\vec{\alpha})n_j(\alpha)}{1 + \lambda}. \]

Therefore, we obtain the system
\[ \frac{d}{dt} \alpha_j = \frac{2n(\vec{\alpha})n_j(\vec{\alpha})}{1 + \lambda}, \quad \text{for} \quad j = 1, \ldots, d. \] (4.21)

In the symmetric set \( \{ \alpha_1 = \alpha_2 = \cdots = \alpha_d = \alpha \} \), a guess tensor \( G_1 \) is symmetric. We can see that in this case \( n_j(\vec{\alpha}) = n_k(\vec{\alpha}) \) for any \( j \) and \( k \). Thus, \( \frac{\partial}{\partial \alpha_j} E_\lambda(G_1) = \frac{\partial}{\partial \alpha_k} E_\lambda(G_1) \) for all \( j \) and \( k \). Therefore, by (4.20), we know that the symmetric set \( \{ \alpha_1 = \alpha_2 = \cdots = \alpha_d = \alpha \} \) is an invariant set under the gradient flow. It was shown in [24] that a local minimum of \( E_\lambda(G_1) \) is attained in this invariant symmetric set. In this chapter, we study the situations when the \( \alpha_i \)'s are not equal and analyze the existence of local minima outside the symmetric set. Note that throughout the whole chapter, we do not consider any particular algorithms in tensor approximation. We only study the critical points of the gradient flow. This can help us understand the gradient-based algorithms in tensor approximation.
4.2 The existence of non-symmetric equilibrium points of the gradient flow

In this section, we study the existence of the equilibrium points of the gradient flow outside of the symmetric set. We show that the non-symmetric equilibrium points of the gradient flow are in the form of $\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d$. Furthermore, we discuss the existence of such equilibrium points according to the values of parameters $z$, $d$ and $m$. The results are summarized in Table 4.1.

Table 4.1: The existence of non-symmetric equilibrium points of the gradient flow in the form of $\alpha_1 = \cdots = \alpha_m \neq \alpha_{d+1} = \cdots = \alpha_d$ under different parameters $d$, $z$ and $m$. We see from this table that such equilibrium points do not exist when $d$ is even, $m$ is odd and $0 < z \leq 1$. When $d$ and $m$ are both even, such an equilibrium point might not exist for certain values of $z \in [-1, 0)$. For all other parameters values, at least one non-symmetric equilibrium point of the gradient flow exists.

<table>
<thead>
<tr>
<th></th>
<th>$m$ odd</th>
<th></th>
<th>$m$ even</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$d$ odd</td>
<td>exist</td>
<td>exist</td>
<td>exist</td>
<td>exist</td>
</tr>
<tr>
<td>$d$ even</td>
<td>exist</td>
<td>do not exist</td>
<td>exist for certain $z$</td>
<td>exist</td>
</tr>
</tbody>
</table>

Consider the angular change according to the gradient flow as in Equation (4.20)

$$\frac{\partial}{\partial \alpha_j} E_{d}(G_1) = -\frac{2n(i)}{1 + \lambda}, \quad j = 1, \ldots, d.$$  

Setting the right-hand side of this equation to be 0, we can obtain the equilibrium points. We do not take into account the situation when $n(\vec{\alpha}) = 0$. Because when $n(\vec{\alpha}) = 0$, we have $E_{d}(G_1) = 1$ by (4.11). It was pointed out in Subsection 4.1.3 that the range of $E_{d}(G_1)$ is $[0, 1]$. So the maximum error $E_{d}(G_1) = 1$ is attained when $n(\vec{\alpha}) = 0$. We are interested in the existence of local minimum points and saddle points, thus we only need to consider $\vec{\alpha}$ that for all $j$,

$$n_j(\vec{\alpha}) = -\sin(\alpha_j) \prod_{i \neq j} \cos(\alpha_i) - z \sin(\alpha_j - \phi) \prod_{i \neq j} \cos(\alpha_i - \phi) = 0.$$  

(4.22)
Note that for the rank-1 guess tensor $G_1$, the value of $\lambda$ plays no role in our calculations of the critical points of $E_{\lambda}(G_1)$. But for the sake of convenience in the future study on guess tensors with higher rank, we keep $\lambda$ in our calculations.

When $\alpha_j \notin \{\phi, \phi - \pi/2\}$ for all $j$, Equation (4.22) can be rewritten as

$$z = -\frac{\sin(\alpha_j) \prod_{i \neq j} \cos(\alpha_i)}{\sin(\alpha_j - \phi) \prod_{i \neq j} \cos(\alpha_i - \phi)}, \text{ for all } j.$$  \hspace{1cm} (4.23)

Assume that $\alpha_1 \neq \alpha_i$ for some $i \neq 1$. Equation (4.23) can be rewritten as

$$z = -\frac{\sin(\alpha_1) \prod_{k=2}^d \cos(\alpha_k)}{\sin(\alpha_1 - \phi) \prod_{k=2}^d \cos(\alpha_k - \phi)} = -\frac{\sin(\alpha_i) \prod_{k=1, k \neq i}^d \cos(\alpha_k)}{\sin(\alpha_i - \phi) \prod_{k=1, k \neq i}^d \cos(\alpha_k - \phi)}.$$

When $\alpha_j \neq \pi/2$ for all $j$, this can be simplified to

$$\frac{\sin(\alpha_1) \cos(\alpha_i)}{\sin(\alpha_1 - \phi) \cos(\alpha_i - \phi)} = \frac{\sin(\alpha_i) \cos(\alpha_1)}{\sin(\alpha_i - \phi) \cos(\alpha_1 - \phi)}.$$

Thus,

$$\frac{\tan(\alpha_1)}{\tan(\alpha_i)} = \frac{\tan(\alpha_1 - \phi)}{\tan(\alpha_i - \phi)}. \hspace{1cm} (4.24)$$

Using the difference identity for tangent, i.e.,

$$\tan(\alpha_i - \phi) = \frac{\tan \alpha_i - \tan \phi}{1 + \tan \alpha_i \tan \phi},$$

we can rewrite Equation (4.24) to

$$\frac{\tan \alpha_1}{\tan \alpha_i} = \frac{(\tan \alpha_1 - \tan \phi)(1 + \tan \alpha_i \tan \phi)}{(\tan \alpha_i - \tan \phi)(1 + \tan \alpha_1 \tan \phi)}.$$

Simplifying this equation, we have

$$\tan \phi(\tan \alpha_1 - \tan \alpha_i)[1 - \tan \alpha_1 \tan \alpha_i + (\tan \alpha_1 + \tan \alpha_i) \tan \phi] = 0.$$

Because $\phi \in (0, \pi/2]$ and $\alpha_1 \neq \alpha_i$, we have $\tan \phi \neq 0$ and $\tan \alpha_1 - \tan \alpha_i \neq 0$, thus

$$\frac{\tan \alpha_1 + \tan \alpha_i}{1 - \tan \alpha_1 \tan \alpha_i} = -\frac{1}{\tan \phi}.$$
Therefore, by the sum identity for tangent, i.e.,
\[
\tan(\alpha_1 + \alpha_i) = \frac{\tan \alpha_1 + \tan \alpha_i}{1 - \tan \alpha_1 \tan \alpha_i},
\]
we obtain that
\[
\tan(\alpha_1 + \alpha_i) = -\frac{1}{\tan \phi}.
\tag{4.25}
\]
Because \(-1/\tan \phi = -\cot \phi = \tan(\phi - \frac{\pi}{2})\), we conclude that for \(\alpha_1 \neq \alpha_i\),
\[
\alpha_1 + \alpha_i = \phi - \frac{\pi}{2} + k\pi, \quad \text{for some integer } k.
\tag{4.26}
\]

Now consider the situation when \(\alpha_j \in \{\phi, \phi - \pi/2, \pi/2\}\) for some \(j\), we can obtain that Equation (4.26) still holds for those special values of \(\alpha_j\).

1. If \(\alpha_j = \phi\), then by (4.22) we obtain that \(n_j(\vec{a}) = -\sin(\phi) \prod_{i \neq j} \cos(\alpha_i)\). The solution to \(n_j(\vec{a}) = 0\) is \(\alpha_i = \pi/2\) for some \(i\). So we obtain that \(\alpha_i + \alpha_j = \phi + \pi/2\) when \(\alpha_j \neq \alpha_i\), which coincides with Equation (4.26).

2. If \(\alpha_j = \phi - \pi/2\), then for all \(i \neq j\), similarly we can find that
\[
n_i(\vec{a}) = -\sin(\alpha_i) \cos(\phi - \pi/2) \prod_{k \neq i, j} \cos(\alpha_k).
\]
The solution to the system of equations \(n_i(\vec{a}) = 0\) for all \(i \neq j\) is \(\alpha_i = 0\). Thus, we have \(\alpha_i \neq \alpha_j\) and \(\alpha_j + \alpha_i = \phi - \pi/2\).

3. If \(\alpha_j = \pi/2\), then for all \(i \neq j\), we obtain
\[
n_i(\vec{a}) = -z \sin(\alpha_i - \phi) \cos(\pi/2 - \phi) \prod_{k \neq i, j} \cos(\alpha_k - \phi)
\]
by (4.22). The solution to the system of equations \(n_i(\vec{a}) = 0\) for all \(i \neq j\) is \(\alpha_i = \phi\). So we obtain that
\(\alpha_j + \alpha_i = \phi + \pi/2\) and \(\alpha_i \neq \alpha_j\).

Note that Equation (4.26) holds for all distinct angles \(\alpha_1\) and \(\alpha_i\). Because we only consider all \(\alpha_i \in (-\pi/2, \pi/2]\) and \(\phi \in (0, \pi/2]\), in (4.26) \(k\) can only be chosen from \(-1, 0, 1\). If there exist angles \(\alpha_2\) and \(\alpha_3\) with \(\alpha_1 \neq \alpha_2 \neq \alpha_3 \neq \alpha_1\), Equation (4.26) implies that
\[
\alpha_1 + \alpha_3 = \phi - \frac{\pi}{2} + k_1\pi, \quad \text{where } k_1 \in \{-1, 0, 1\};
\]
\[
\alpha_2 + \alpha_3 = \phi - \frac{\pi}{2} + k_2\pi, \quad \text{where } k_2 \in \{-1, 0, 1\};
\]
Subtracting the two equations yields

$$\alpha_2 - \alpha_1 = \tilde{k}\pi \quad \text{where} \quad \tilde{k} = k_2 - k_1 \in \{0, \pm 1, \pm 2\}.$$ 

Since $$\alpha_1, \alpha_2 \in (-\pi/2, \pi/2]$$, $$\alpha_2 - \alpha_1 \in (-\pi, \pi)$$. Thus, $$\tilde{k}$$ in the above equation can only be 0. Then we have $$\alpha_2 = \alpha_1$$. So we conclude that there can only be two different values among the angles $$\alpha_1, \ldots, \alpha_d$$. For simplicity, we assume that there are $$m$$ angles that are the same as $$\alpha_1$$ and the other $$d - m$$ angles are the same as $$\alpha_d$$, where $$m = 1, \ldots, [d/2]$$, i.e.,

$$\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \alpha_{m+2} = \cdots = \alpha_d, \quad \text{where} \quad m = 1, \ldots, [d/2]. \quad (4.27)$$

From Equation (4.26) with $$k = -1, 0, 1$$, we can infer:

$$\cos(\alpha_d - \phi) = -\sin(\alpha_1); \quad (4.28)$$

$$\sin(\alpha_d - \phi) = -\cos(\alpha_1); \quad (4.29)$$

$$\cos(\alpha_d) = -\sin(\alpha_1 - \phi); \quad (4.30)$$

$$\sin(\alpha_d) = -\cos(\alpha_1 - \phi). \quad (4.31)$$

Plugging (4.27) into Equation (4.22), we simplify the equation $$\frac{\partial}{\partial \alpha_1} E_\lambda(G_1) = 0$$ to

$$\sin(\alpha_1)[\cos(\alpha_1)]^{m-1}[\cos(\alpha_d)]^{d-m} + z\sin(\alpha_1 - \phi)[\cos(\alpha_1 - \phi)]^{m-1}[\cos(\alpha_d - \phi)]^{d-m} = 0. \quad (4.32)$$

Substituting (4.28) and (4.30) into (4.32) and simplifying it, we obtain that if $$\alpha_1 \neq 0$$ and $$\alpha_1 \neq \phi$$,

$$[\cos(\alpha_1)]^{m-1}[\sin(\alpha_1 - \phi)]^{d-m-1} + z[\cos(\alpha_1 - \phi)]^{m-1}[\sin(\alpha_1)]^{d-m-1} = 0. \quad (4.33)$$

Note that if $$\alpha_1 = 0$$, then $$\alpha_d = \phi - \pi/2$$ and $$n(\tilde{\alpha}) = 0$$. If $$\alpha_1 = \phi$$, then $$\alpha_d = \pi/2$$ and $$n(\tilde{\alpha}) = 0$$. So either $$\alpha_1 = 0$$ or $$\alpha_1 = \phi$$ implies that the regularized error $$E_\lambda(G_1) = 1$$ by (4.11). Therefore, these two equilibrium points are maximum points.

After simplifying the equation $$\frac{\partial}{\partial \alpha_d} E_\lambda(G_1) = 0$$, we obtain that

$$\sin(\alpha_d)[\cos(\alpha_d)]^{d-m-1}[\cos(\alpha_1)]^m + z\sin(\alpha_d - \phi)[\cos(\alpha_d - \phi)]^{d-m-1}[\cos(\alpha_1 - \phi)]^m = 0.$$
Because of equalities (4.28-4.31), we find that the above equation is equivalent to (4.33). Thus, solving Equations (4.33) and (4.26) is sufficient to find the equilibrium points of the gradient flow. Let us denote the left-hand side of (4.33) as

\[
f(x) = [\cos(x)]^{m-1}[\sin(x - \phi)]^{d-m-1} + z[\cos(x - \phi)]^{m-1}[\sin(x)]^{d-m-1}.
\]

We can check that:

- When \(d\) is even and \(m\) is odd, both \(m - 1\) and \(d - m - 1\) are even. When \(z > 0\), it is obvious that the solution \(f(x) = 0\) does not exist. When \(z < 0\), we find that

\[
\begin{align*}
f(0) &= [\sin(-\phi)]^{d-m-1} > 0, \\
f(\phi) &= z[\sin(\phi)]^{d-m-1} < 0.
\end{align*}
\]

Therefore, by the Intermediate Value Theorem, we conclude that if \(d\) is even, \(m\) is odd and \(z < 0\), at least one solution to (4.33) exists.

- When \(d\) is even and \(m\) is even, we know that \(d - m - 1\) is odd. It is easy to find that

\[
\begin{align*}
f(0) &= [\sin(-\phi)]^{d-m-1} = (-1)^{d-m-1}(\sin \phi)^{d-m-1} = -(\sin \phi)^{d-m-1}, \\
f(\phi) &= z[\sin(\phi)]^{d-m-1}.
\end{align*}
\]

So if \(z > 0\), then \(f(0)f(\phi) < 0\). We can conclude from the Intermediate Value Theorem that there is at least one solution between 0 and \(\phi\).

If \(z < 0\), then \(f(0)\) and \(f(\phi)\) are both negative. For any \(\phi \in (0, \pi/2)\), we can find positive integers \(s\) and \(t\) such that \(s < t\) and \(\frac{s+t}{t} \phi \in (0, \pi/2)\). Note that we also have \(\frac{s}{t} \phi \in (0, \pi/2)\). Letting \(x^* = \frac{s}{2} - \frac{s}{t} \phi\), we have

\[
\begin{align*}
f(x^*) &= \left[\cos\left(\frac{\pi}{2} - \frac{s}{t} \phi\right)^{m-1}\right]\left[\sin\left(\frac{\pi}{2} - \frac{s + t}{t} \phi\right)^{d-m-1}\right] \\
&\quad + z\left[\cos\left(\frac{\pi}{2} - \frac{s + t}{t} \phi\right)^{m-1}\right]\left[\sin\left(\frac{\pi}{2} - \frac{s}{t} \phi\right)^{d-m-1}\right] \\
&= \left[\sin\left(\frac{s}{t} \phi\right)^{m-1}\right]\left[\cos\left(\frac{s + t}{t} \phi\right)^{d-m-1}\right] + z\left[\sin\left(\frac{s + t}{t} \phi\right)^{m-1}\right]\left[\cos\left(\frac{s}{t} \phi\right)^{d-m-1}\right].
\end{align*}
\]
We know that $\left[ \sin \left( \frac{s}{t} \phi \right) \right]^{m-1} \left[ \cos \left( \frac{s}{t} + t \phi \right) \right]^{d-m-1} > 0$ and $\left[ \sin \left( \frac{s}{t} \phi \right) \right]^{m-1} \left[ \cos \left( \frac{s}{t} \phi \right) \right]^{d-m-1} > 0$. So for any $z$ satisfying $\left[ \sin \left( \frac{s}{t} \phi \right) \right]^{m-1} \left[ \cos \left( \frac{s}{t} + t \phi \right) \right]^{d-m-1} < z < 0$ and $z \geq -1$, we have $f(x^*) > 0$. Therefore, by the Intermediate Value Theorem, for any $\phi \in (0, \pi/2)$, there exist certain values of $z < 0$ such that the solution to (4.33) exists.

- When $d$ is odd, we find that

$$f(-\pi/2) = z(-1)^{d-m-1}[\cos(-\pi/2 - \phi)]^{m-1} = (-1)^{d-2}z(\sin \phi)^{m-1},$$
$$f(\pi/2) = z[\cos(\pi/2 - \phi)]^{m-1} = z(\sin \phi)^{m-1}.$$ 

Therefore, $f(-\pi/2)$ and $f(\pi/2)$ have different signs when $d$ is odd. Thus, in this case, according to the Intermediate Value Theorem, for any $m$ and $z$ there is always at least one solution to Equation (4.33) in $(-\pi/2, \pi/2)$.

In conclusion, when $d$ is odd the equilibrium point of the gradient flow in the form of $\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d$ to (4.33) always exists. When $d$ is even, the equilibrium point in such form exists:

(i) When $m$ is odd and $z < 0$.

(ii) When $m$ is even and $z > 0$.

(iii) When $m$ is even and only some values of $z < 0$.

For $d$ is even and $z > 0$, there do not exist odd number of equal angular variables to solve Equation (4.33). The results are summarized in Table 4.1.

### 4.3 Stability of the equilibrium points

In this section, we study the stability of the non-symmetric equilibrium points of the gradient flow and prove the following proposition.
Proposition 4.3.1. The equilibrium points of (4.20) satisfy either (a) or (b) listed as below:

(a) The equation \( n(\vec{a}) = 0 \). The solutions to this equation are the maximum point of the regularized error function \( E_\lambda(G_1) \).

(b) Equations (4.26) and (4.33). Such equilibrium points are in the form of
\[
\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d
\]
and they are saddle points for \( z \in [-1, 1] \backslash \{0\} \) and \( \phi \in (0, \pi/2) \). Specifically, when \( m = 1 \) there are two different kinds of dynamics depending on the parameters:

(i) When \( z > 0 \) (\( d \) is odd), or \((-z)^{1/2} > \frac{1 - \sin \phi}{\cos \phi}\), the saddle equilibrium point has a one-dimensional unstable manifold and a \((d - 1)\)-dimensional stable manifold.

(ii) When \( z < 0 \) and \((-z)^{1/2} < \frac{1 - \sin \phi}{\cos \phi}\), the saddle equilibrium point has a \((d - 2)\)-dimensional unstable manifold and a 2-dimensional stable manifold.

Note that part (a) of Proposition 4.3.1 was already shown in Section 4.2. So we will prove part (b) in this section. For a better understanding of the proof, we divide the proof into two cases: \( \alpha_1 \neq \alpha_2 = \cdots = \alpha_d \) and \( \alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d \) where \( 2 \leq m \leq d - 2 \).

4.3.1 Analysis of the case \( \alpha_1 \neq \alpha_2 = \cdots = \alpha_d \)

We first consider the simplest case where only one angle \( \alpha_1 \) is different with the other angles \( \alpha_2 = \cdots = \alpha_d \). In this case, \( m = 1 \) and Equation (4.33) can be rewritten as
\[
\sin(\alpha_1 - \phi)\]^{d-2} + z[\sin(\alpha_1)]^{d-2} = 0.
\] (4.34)
The solution to (4.34) exists when \( d \) is odd, or when \( d \) is even and \( z < 0 \) according to Table 4.1. Under such parameter values, we can solve (4.34) to obtain that
\[
\alpha_1 = \arctan\left( \frac{\sin \phi}{\cos \phi - (-z)^{1/(d-2)}} \right).
\] (4.35)
By Equations (4.28) and (4.30), we can rewrite (4.34) as

\[ [\cos(\alpha_d)]^{d-2} + z[\cos(\alpha_d - \phi)]^{d-2} = 0. \]

The solution to this equation is

\[ \alpha_d = \arctan \left( -\cos \phi + (-z)^{1/(d-2)} \right) \]  \hspace{2cm} (4.36)

We can find the Hessian matrix at this equilibrium point and then study the eigenvalues of this matrix. We know that at the equilibria, \( n_j(\vec{\alpha}) = 0 \) for all \( j \). So

\[ \frac{\partial^2}{\partial \alpha_j \partial \alpha_k} E_j(G_1) = -\frac{2}{1 + \lambda} n(\vec{\alpha}) n_{jk}(\vec{\alpha}), \]  \hspace{2cm} (4.37)

where

\[ n(\vec{\alpha}) = \frac{\cos(\alpha_1)[\cos(\alpha_d)]^{d-1} + z \cos(\alpha_1 - \phi)[\cos(\alpha_d - \phi)]^{d-1}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} = (-1)^{d-1} \frac{\cos(\alpha_1)[\sin(\alpha_1 - \phi)]^{d-1} + z \cos(\alpha_1 - \phi)[\sin(\alpha_1)]^{d-1}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}}. \]

Applying (4.34), we simplify the above equation to

\[ n(\vec{\alpha}) = \frac{(-1)^{d-1}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} \left( \cos(\alpha_1)[\sin(\alpha_1 - \phi)]^{d-1} - \cos(\alpha_1 - \phi) \sin(\alpha_1) [\sin(\alpha_1 - \phi)]^{d-2} \right) \]

\[ = (-1)^{d-1} \frac{(-1)^{d-1}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} \left[ \sin(\alpha_1 - \phi) \right]^{d-2} \left( \cos(\alpha_1) \sin(\alpha_1 - \phi) - \cos(\alpha_1 - \phi) \sin(\alpha_1) \right) \]

\[ = (-1)^d \frac{(-1)^{d-1}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} \left[ \sin(\alpha_1 - \phi) \right]^{d-2} \sin(\phi). \]

When \( j = k \), we already know that \( n_{jj} = -n(\vec{\alpha}) \) by (4.19). For \( j \neq k \), we have

\[ n_{jk}(\vec{\alpha}) = \frac{\sin(\alpha_j) \sin(\alpha_k) \prod_{i \neq j,k} \cos(\alpha_i) + z \sin(\alpha_j - \phi) \sin(\alpha_k - \phi) \prod_{i \neq j,k} \cos(\alpha_i - \phi)}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}}. \]

Therefore, for \( j = 1 \) and any \( k \neq 1 \), we find that

\[ n_{1k}(\vec{\alpha}) = \frac{\sin(\alpha_1) \sin(\alpha_d)[\cos(\alpha_d)]^{d-2} + z \sin(\alpha_1 - \phi) \sin(\alpha_d - \phi)[\cos(\alpha_d - \phi)]^{d-2}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} \]

\[ = (-1)^{d-1} \frac{\sin(\alpha_1) \cos(\alpha_1 - \phi)[\sin(\alpha_1 - \phi)]^{d-2} + z \sin(\alpha_1 - \phi) \cos(\alpha_1)[\sin(\alpha_1)]^{d-2}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}}. \]
Applying (4.34) to the above equation, we obtain that

\[
\begin{align*}
n_{1k}(\vec{\alpha}) &= (-1)^{d-1} \frac{\sin(\alpha_1) \cos(\alpha_1 - \phi)[\sin(\alpha_1 - \phi)]^{d-2} - \cos(\alpha_1)[\sin(\alpha_1 - \phi)]^{d-1}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} \\
&= (-1)^{d-1} \frac{\sin(\alpha_1 - \phi)[\sin(\alpha_1 - \phi)]^{d-2} \{\sin(\alpha_1) \cos(\alpha_1 - \phi) - \cos(\alpha_1)[\sin(\alpha_1 - \phi)]\}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} \\
&= (-1)^{d-1} \frac{\sin(\alpha_1 - \phi)[\sin(\alpha_1 - \phi)]^{d-2} \sin(\phi)}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}}. \tag{4.39}
\end{align*}
\]

For our convenience, we write the notation \(n_{jk}(\vec{\alpha})\) as \(n_{j\bar{k}}\) for the remainder of this chapter.

Combining Equations (4.38) and (4.39), we find that for any \(k \neq 1\):

\[
n_{1k} + n(\vec{\alpha}) = 0.
\]

Thus, for any \(k\) and \(j\),

\[
n_{1k} = n_{k1} = n_{jj} = -n(\vec{\alpha}) = (-1)^{d-1} \frac{\sin(\alpha_1 - \phi)[\sin(\alpha_1 - \phi)]^{d-2} \sin(\phi)}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}}. \tag{4.40}
\]

For \(j \neq k\) and \(j, k \neq 1\), according to (4.28)-(4.31), we have

\[
n_{jk} = \frac{(\sin \alpha_d)^2 \cos(\alpha_1)(\cos \alpha_d)^{d-3} + z[\sin(\alpha_d - \phi)]^2 \cos(\alpha_1 - \phi)(\cos \alpha_1 - \phi)^{d-3}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} \\
= (-1)^{d-1} \frac{\cos \alpha_1[\cos(\alpha_1 - \phi)]^2[\sin(\alpha_1 - \phi)]^{d-3} + z \cos(\alpha_1 - \phi)(\cos \alpha_1)^2(\sin \alpha_1)^{d-3}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} \\
= (-1)^{d-1} \frac{\cos(\alpha_1)[\cos(\alpha_1 - \phi)]}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} \cos(\alpha_1)(\cos(\alpha_1 - \phi)) \{\cos(\alpha_1 - \phi)[\sin(\alpha_1 - \phi)]^{d-3} + z \cos(\alpha_1)(\sin \alpha_1)^{d-3}\}.  
\]
Applying Equation (4.34), we obtain that

\[ n_{jk} = \frac{(-1)^{d-1}}{[1 + 2z(\cos \phi)^d + z^2]^{1/2}} \cos(\alpha_1)[\cos(\alpha_1 - \phi)] \]

\[ \left\{ \cos(\alpha_1 - \phi)[\sin(\alpha_1 - \phi)]^{d-3} - \frac{\cos(\alpha_1)}{\sin(\alpha_1)} [\sin(\alpha_1 - \phi)]^{d-2} \right\} \]

\[ = \frac{(-1)^{d-1} \cos(\alpha_1)}{[1 + 2z(\cos \phi)^d + z^2]^{1/2} \sin(\alpha_1)} [\cos(\alpha_1 - \phi)][\sin(\alpha_1 - \phi)]^{d-3} \]

\[ [\sin(\alpha_1) \cos(\alpha_1 - \phi) - \cos(\alpha_1) \sin(\alpha_1 - \phi)] \]

\[ = \frac{(-1)^{d-1} \cos(\alpha_1)}{[1 + 2z(\cos \phi)^d + z^2]^{1/2} \sin(\alpha_1)} [\cos(\alpha_1 - \phi)][\sin(\alpha_1 - \phi)]^{d-3} \sin(\phi). \] (4.41)

For our convenience in deriving the Hessian matrix, we use the following notations:

\[ n^2 = [n(\vec{a})]^2; \] (4.42)

\[ A = -\frac{n_{jk}(\vec{a})}{n(\vec{a})} \text{ for } j \neq k \text{ and } j, k \neq 1. \] (4.43)

We do not consider the situation where \( n(\vec{a}) = 0 \) because when \( n(\vec{a}) = 0 \), we have a maximum critical point. We consider only the case when \( n(\vec{a}) \neq 0 \), so we know that \( n^2 > 0 \) and \( A \) is well-defined.

We can express \( n^2 \) and \( A \) using (4.38) and (4.41):

\[ n^2 = \frac{1}{[1 + 2z(\cos \phi)^d + z^2]^{2d-4} \sin(\alpha_1 - \phi)^2}. \] (4.44)

Similarly,

\[ A = \frac{\cos(\alpha_1) \cos(\alpha_1 - \phi)}{\sin(\alpha_1) \sin(\alpha_1 - \phi)} = \cot(\alpha_1) \cot(\alpha_1 - \phi). \] (4.45)
According to Equations (4.37), (4.40), (4.42) and (4.43) we can obtain the following Hessian matrix:

$$H = -\frac{2n(\vec{a})}{1 + \lambda} \begin{bmatrix}
  n_{11}(\vec{a}) & n_{12}(\vec{a}) & \cdots & \cdots & n_{1d}(\vec{a}) \\
  n_{12}(\vec{a}) & n_{22}(\vec{a}) & n_{23}(\vec{a}) & \cdots & n_{2d}(\vec{a}) \\
  \vdots & n_{23}(\vec{a}) & \ddots & \ddots & \vdots \\
  \vdots & \vdots & \ddots & \ddots & \vdots \\
  n_{1d}(\vec{a}) & n_{2d}(\vec{a}) & \cdots & \cdots & n_{dd}(\vec{a})
\end{bmatrix}$$

$$= \frac{2n^2}{1 + \lambda} \begin{bmatrix}
  1 & 1 & \cdots & \cdots & 1 \\
  1 & 1 & A & \cdots & A \\
  \vdots & A & \ddots & \ddots & \vdots \\
  \vdots & \vdots & \ddots & \ddots & \vdots \\
  1 & A & \cdots & A & 1
\end{bmatrix} \tag{4.46}$$

Now we use the Householder reflector [21] to tridiagonalize the Hessian matrix given by (4.46). The Householder matrix for this Hessian matrix $H$ is

$$P = \frac{1}{2d - 2 \sqrt{d - 1} - 2} \begin{bmatrix}
  0 & 0 & \cdots & \cdots & 0 \\
  0 & 2\sqrt{d - 1} - 2 & 2\sqrt{d - 1} - 2 & \cdots & 2\sqrt{d - 1} - 2 \\
  0 & 2\sqrt{d - 1} - 2 & 2d - 2\sqrt{d - 1} - 4 & -2 & \cdots & -2 \\
  \vdots & \vdots & -2 & \ddots & \ddots & \vdots \\
  \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\
  0 & 2\sqrt{d - 1} - 2 & -2 & \cdots & -2 & 2d - 2\sqrt{d - 1} - 4
\end{bmatrix}.$$
Multiplying matrix $P$ by the Hessian matrix $H$, we obtain that

$$PH = \frac{2n^2}{1 + \lambda}$$

$$
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
\frac{(d-1)\sqrt{d-1}}{d-\sqrt{d-1}} & \frac{(\sqrt{d-1})(1+(d-2)A)}{d-\sqrt{d-1}} & \cdots & \frac{(\sqrt{d-1})(1+(d-2)A)}{d-\sqrt{d-1}} \\
0 & \frac{(\sqrt{d-1})(1-A)}{d-\sqrt{d-1}} & 1-A & \cdots & \frac{1-A}{d-\sqrt{d-1}} \\
\vdots & \vdots & 0 & \ddots & \vdots \\
0 & \frac{(\sqrt{d-1})(1-A)}{d-\sqrt{d-1}} & \cdots & \frac{1-A}{d-\sqrt{d-1}} & 1-A
\end{bmatrix}
$$

Because the Householder matrix $P$ here is a symmetric unitary matrix with real entries, we calculate the matrix product $PHP$ to find that

$$PHP = \frac{2n^2}{1 + \lambda}$$

$$
\begin{bmatrix}
1 & \sqrt{d-1} & 0 & \cdots & 0 \\
\sqrt{d-1} & 1 + (d-2)A & 0 & \cdots & 0 \\
0 & 0 & 1-A & 0 & \cdots \\
\vdots & \vdots & 0 & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 1-A
\end{bmatrix}
$$

(4.47)

We know that eigenvalues of the Hessian matrix $H$ are the same as the eigenvalues of $PHP$ under the Householder reflection. So we can easily see the $d - 2$ eigenvalues on the diagonal of the matrix $PHP$:

$$x_3 = \cdots = x_d = \frac{2n^2}{1 + \lambda} (1 - A).$$

(4.48)

To find the remaining two eigenvalues of the matrix $PHP$ given by (4.47), we need to study the upper left block matrix

$$B = \frac{2n^2}{1 + \lambda} \begin{bmatrix}
1 & \sqrt{d-1} \\
\sqrt{d-1} & 1 + (d-2)A
\end{bmatrix}.$$
To obtain the eigenvalues of matrix $B$, we can find its characteristic equation (ignoring the coefficient $\frac{2n^2}{1+\lambda}$):

$$x^2 - [(d - 2)A + 2]x - (d - 2)(1 - A) = 0. \tag{4.49}$$

Thus, we calculate the two eigenvalues to be

$$x_{1,2} = \frac{n^2}{1 + \lambda} \left[(d - 2)A + 2 \pm \sqrt{(d - 2)^2A^2 + 4(d - 1)}\right].$$

Note that the two eigenvalues $x_1$ and $x_2$ must satisfy:

$$x_1 + x_2 = \frac{2n^2}{1 + \lambda} [(d - 2)A + 2]; \tag{4.50}$$

$$x_1x_2 = -\frac{4(d - 2)n^4(1 - A)}{(1 + \lambda)^2}. \tag{4.51}$$

From Equations (4.50), (4.51) and (4.48) we can conclude that

- If $A < 1$, then $x_3 = \cdots = x_d > 0$ and $x_1x_2 < 0$. Therefore, there are $d - 1$ positive eigenvalues (with $d - 2$ repeated eigenvalues) and one negative eigenvalue.

- If $A > 1$, then $x_3 = \cdots = x_d < 0$ and $x_1x_2 > 0$. In this case, $A > 1$, so $x_1 + x_2 > 0$ according to Equation (4.50). Thus, there are two positive eigenvalues and $d - 2$ repeated negative eigenvalues.

- If $A = 1$, then we find that $x_2 = x_3 = \cdots = x_d = 0$ and $x_1 > 0$. There is only one non-zero eigenvalue and it is positive.

The last step is to determine the sign of $1 - A$. By Equations (4.44) and (4.45) we have

$$1 - A = 1 - \cot(\alpha_1) \cot(\alpha_1 - \phi)$$

$$= \frac{\sin(\alpha_1) \sin(\alpha_1 - \phi) - \cos(\alpha_1) \cos(\alpha_1 - \phi)}{\sin(\alpha_1) \sin(\alpha_1 - \phi)}$$

$$= \frac{-1}{\left[\sin(\alpha_1 - \phi)\right]^2} \frac{\sin(\alpha_1)}{\sin(\alpha_1)} \cos(2\alpha_1 - \phi). \tag{4.52}$$
According to Equation (4.34), we have

\[ (-z)^{1/(d-2)} = \frac{\sin(\alpha_1 - \phi)}{\sin(\alpha_1)}. \]

Substituting this into (4.52), we obtain that

\[ 1 - A = \frac{-1}{[\sin(\alpha_1 - \phi)]^2} (-z)^{1/(d-2)} \cos(2\alpha_1 - \phi). \]  

Therefore, the sign of \(1 - A\) depends on the signs of \((-z)^{1/(d-2)}\) and \(\cos(2\alpha_1 - \phi)\). We know from (4.35) that

\[ \tan(\alpha_1) = \frac{\sin \phi}{\cos \phi - (-z)^{1/(d-2)}}. \]

Thus, we can use this to further simplify \(\cos(2\alpha_1 - \phi)\) as

\[
\cos(2\alpha_1 - \phi) = \cos(2\alpha_1) \cos(\phi) + \sin(2\alpha_1) \sin(\phi) \\
= (\cos^2 \alpha_1 - \sin^2 \alpha_1) \cos \phi + 2 \sin \alpha_1 \cos \alpha_1 \sin \phi \\
= \cos^2 \alpha_1 [ (1 - \tan^2 \alpha_1) \cos \phi + 2 \tan \alpha_1 \sin \phi] \\
= \cos^2 \alpha_1 \left[ 1 - \frac{\sin^2 \phi}{[\cos \phi - (-z)^{1/(d-2)}]^2} \right] \cos \phi + 2 \frac{\sin \phi}{[\cos \phi - (-z)^{1/(d-2)}]} \sin \phi \\
= \frac{\cos^2 \alpha_1}{[\cos \phi - (-z)^{1/(d-2)}]^2} \left[ \cos \phi (-z)^{2/(d-2)} - 2 (-z)^{1/(d-2)} + \cos \phi \right] \\
= \frac{\cos^2 \alpha_1 \cos \phi}{[\cos \phi - (-z)^{1/(d-2)}]^2} \left[ (-z)^{1/(d-2)} - \frac{1 + \sin \phi}{\cos \phi} \right] \left[ (-z)^{1/(d-2)} - \frac{1 - \sin \phi}{\cos \phi} \right].
\]

(4.54)

Therefore, plugging (4.54) into (4.53), we obtain that

\[
1 - A = - \frac{\cos^2(\alpha_1) \cos \phi}{[\sin(\alpha_1 - \phi)]^2[\cos \phi - (-z)^{1/(d-2)}]^2} (-z)^{1/(d-2)} \\
\left[ (-z)^{1/(d-2)} - \frac{1 + \sin \phi}{\cos \phi} \right] \left[ (-z)^{1/(d-2)} - \frac{1 - \sin \phi}{\cos \phi} \right].
\]

(4.55)

When \(\phi \in (0, \frac{\pi}{2})\), we have \(\cos \phi > 0\), \(\sin \phi > 0\) and \(0 < \frac{1 - \sin \phi}{\cos \phi} < \frac{1 + \sin \phi}{\cos \phi}\). We can derive the following results from Equation (4.55):
• When \((-z)^{1/(d-2)} < 0\) or \((-z)^{1/(d-2)} \in \left(\frac{1-\sin \phi}{\cos \phi}, \frac{1+\sin \phi}{\cos \phi}\right)\), we have \(1 - A > 0\). These restrictions on \((-z)^{1/(d-2)}\) are equivalent to \(z > 0\) or \(\frac{1-\sin \phi}{\cos \phi} < (-z)^{1/(d-2)} \leq 1\) because \((-z)^{1/(d-2)} < \frac{1+\sin \phi}{\cos \phi}\) is always true when \(-1 < z < 1\).

• When \((-z)^{1/(d-2)} \in (0, \frac{1-\sin \phi}{\cos \phi})\), we have \(1 - A < 0\). Note that \((-z)^{1/(d-2)} > \frac{1+\sin \phi}{\cos \phi}\) cannot hold because \(-1 < z < 1\).

Note that \(1 - A = 0\) only happens when \((-z)^{1/(d-2)} = \frac{1-\sin \phi}{\cos \phi}\) because of the restriction that \(z \neq 0\). In this case, we can use (4.35) and (4.36) to see that \(\alpha_1 = \alpha_d\). Thus, \(1 - A = 0\) occurs in the symmetric set. Therefore, we can finally conclude that:

1. When the parameters satisfy \(z > 0\) (\(d\) is odd), or \((-z)^{1/(d-2)} > \frac{1+\sin \phi}{\cos \phi}\), the Hessian matrix at the equilibrium point has \(d - 1\) positive eigenvalues and one negative eigenvalue. Thus, the equilibrium point is a saddle and there is only one unstable direction.

2. When the parameters satisfy \(z < 0\) and \((-z)^{1/(d-2)} < \frac{1-\sin \phi}{\cos \phi}\), the Hessian matrix at the equilibrium point has two positive eigenvalues and \(d - 2\) repeated negative eigenvalues. Then we know that the equilibrium point is also a saddle, but there are \(d - 2\) unstable directions.

4.3.2 Analysis of the case \(\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d\) where \(2 \leq m \leq d - 2\)

Now we consider the case where there are \(m\) angles that are the same as \(\alpha_1\) and the other \(d - m\) angles are the same as \(\alpha_d\). We let \(2 \leq m \leq d - 2\) so that we do not repeat the case when \(m = 1\). When \(m = d - 1\), we can swap \(\alpha_1\) and \(\alpha_d\) and then apply the results for \(m = 1\). We show in this subsection that all non-symmetric equilibrium points of the gradient flow in the form of \(\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d\) are saddle points when \(n(\vec{a}) \neq 0\).
We already know that the equilibria of the gradient flow are the solutions to Equation (4.33). By (4.13), we find that in this case,

\[ n(\vec{a}) = \frac{1}{\|\vec{T}\|} \{(\cos \alpha_1)^m \cos \alpha_d\}^{d-m} + z[\cos(\alpha_1 - \phi)]^m \cos^m(\alpha_d - \phi) \] 

\[ = \frac{(-1)^{d-m}}{\|\vec{T}\|} \{(\cos \alpha_1)^m \sin(\alpha_1 - \phi)\}^{d-m} + z[\cos(\alpha_1 - \phi)]^m \sin(\alpha_1)^{d-m} \]. \quad (4.56) \]

By Equation (4.33),

\[ z = \frac{\cos(\alpha_1)^{m-1} \sin(\alpha_1 - \phi)^{d-m-1}}{\cos(\alpha_1 - \phi)^{m-1} \sin(\alpha_1)^{d-m-1}}. \quad (4.57) \]

Plugging this into (4.56), we have

\[ n(\vec{a}) = \frac{(-1)^{d-m}}{\|\vec{T}\|} \{(\cos \alpha_1)^m \sin(\alpha_1 - \phi)\}^{d-m} - \cos(\alpha_1 - \phi) \sin \alpha_1 \cos^{m-1}(\alpha_1) \sin(\alpha_1 - \phi)^{d-m-1} \]

\[ = \frac{(-1)^{d-m}}{\|\vec{T}\|} \{(\cos \alpha_1)^m \sin(\alpha_1 - \phi)\}^{d-m-1} \cos \alpha_1 \sin(\alpha_1 - \phi) - \cos(\alpha_1 - \phi) \sin(\alpha_1) \]

\[ = \frac{(-1)^{d-m+1}}{\|\vec{T}\|} \{(\cos \alpha_1)^m \sin(\alpha_1 - \phi)\}^{d-m-1} \sin(\phi). \quad (4.58) \]

For \( i = 1, \ldots, m \), the partial derivatives of \( n(\vec{a}) \) are

\[ n_i(\vec{a}) \]

\[ = \frac{1}{\|\vec{T}\|} \left\{ - \sin \alpha_1 (\cos \alpha_1)^{m-1} (\cos \alpha_d) \right\}^{d-m} - z \sin(\alpha_1 - \phi) [\cos(\alpha_1 - \phi)]^{m-1} \cos(\alpha_1 - \phi) \}

\[ = \frac{(-1)^{d-m}}{\|\vec{T}\|} \left\{ \sin \alpha_1 (\cos \alpha_1)^{m-1} (\sin(\alpha_1 - \phi)) \right\}^{d-m} + z \sin(\alpha_1 - \phi) [\cos(\alpha_1 - \phi)]^{m-1} \sin(\alpha_1)^{d-m} \} \]

For \( j = m + 1, \ldots, d \), we have

\[ n_j(\vec{a}) \]

\[ = \frac{1}{\|\vec{T}\|} \left\{ - \sin \alpha_d (\cos \alpha_1)^{m} (\cos \alpha_d) \right\}^{d-m-1} - z \sin(\alpha_1 - \phi) [\cos(\alpha_1 - \phi)]^{m} \cos(\alpha_1 - \phi) \}

\[ = \frac{(-1)^{d-m}}{\|\vec{T}\|} \left\{ \cos(\alpha_1 - \phi) (\cos \alpha_1)^{m} (\sin(\alpha_1 - \phi)) \right\}^{d-m-1} + z \cos \alpha_1 [\cos(\alpha_1 - \phi)]^{m} (\sin(\alpha_1)^{d-m-1} \} \).
We already know that \( n_{ij}(\vec{a}) = -n(\vec{a}) \) for any \( i \) by (4.19), we just need to find the other second order derivatives. For \( i, k = 1, \ldots, m \), but \( k \neq i \), we can find that

\[
\begin{align*}
n_{ik}(\vec{a}) &= \frac{1}{\|\vec{T}\|} \left\{ (\sin \alpha_1)^2 (\cos \alpha_1)^{m-2} (\cos \alpha_d)^{d-m} + z [\sin(\alpha_1 - \phi)]^2 [\cos(\alpha_1 - \phi)]^{m-2} [\cos(\alpha_d - \phi)]^{d-m} \right\} \\
&= (-1)^{d-m} \frac{1}{\|\vec{T}\|} \left\{ (\sin \alpha_1)^2 (\cos \alpha_1)^{m-2} [\sin(\alpha_1 - \phi)]^{d-m} + z [\sin(\alpha_1 - \phi)]^2 [\cos(\alpha_1 - \phi)]^{m-2} (\sin \alpha_1)^{d-m} \right\} .
\end{align*}
\]

Applying (4.57), we can simplify the this equation to

\[
\begin{align*}
n_{ik}(\vec{a}) &= (-1)^{d-m} \frac{1}{\|\vec{T}\|} \left\{ (\sin \alpha_1)^2 (\cos \alpha_1)^{m-2} [\sin(\alpha_1 - \phi)]^{d-m} - \frac{\cos \alpha_1^{m-1} [\sin(\alpha_1 - \phi)]^{d-m+1} \sin \alpha_1}{\cos(\alpha_1 - \phi)} \right\} \\
&= (-1)^{d-m} \frac{1}{\|\vec{T}\| \cos(\alpha_1 - \phi)} \sin \alpha_1 (\cos \alpha_1)^{m-2} [\sin(\alpha_1 - \phi)]^{d-m} [\sin \alpha_1 \cos(\alpha_1 - \phi) - \cos \alpha_1 \sin(\alpha_1 - \phi)] \\
&= (-1)^{d-m} \frac{1}{\|\vec{T}\| \cos(\alpha_1 - \phi)} \sin \alpha_1 (\cos \alpha_1)^{m-2} [\sin(\alpha_1 - \phi)]^{d-m} \sin \phi. 
\end{align*}
\]

For \( j = m + 1, \ldots, d \) and \( i = 1, \ldots, m \), the second order derivatives are

\[
\begin{align*}
n_{ji}(\vec{a}) &= n_{ij}(\vec{a}) = \frac{1}{\|\vec{T}\|} \left\{ \sin \alpha_1 \sin \alpha_d (\cos \alpha_1)^{m-1} (\cos \alpha_d)^{d-m-1} + z \sin(\alpha_1 - \phi) \sin(\alpha_d - \phi) [\cos(\alpha_1 - \phi)]^{m-1} [\cos(\alpha_d - \phi)]^{d-m-1} \right\} \\
&= (-1)^{d-m} \frac{1}{\|\vec{T}\|} \left\{ \sin \alpha_1 \cos(\alpha_1 - \phi) (\cos \alpha_1)^{m-1} [\sin(\alpha_1 - \phi)]^{d-m-1} + z \sin(\alpha_1 - \phi) \cos \alpha_1 [\cos(\alpha_1 - \phi)]^{m-1} (\sin \alpha_1)^{d-m-1} \right\} .
\end{align*}
\]

Similarly, we can simplify this by (4.57):

\[
\begin{align*}
n_{ji}(\vec{a}) &= n_{ij}(\vec{a}) = (-1)^{d-m} \frac{1}{\|\vec{T}\|} \left\{ \sin \alpha_1 \cos(\alpha_1 - \phi) (\cos \alpha_1)^{m-1} [\sin(\alpha_1 - \phi)]^{d-m-1} - [\sin(\alpha_1 - \phi)]^{d-m} (\cos \alpha_1)^{m} \right\} \\
&= (-1)^{d-m} \frac{1}{\|\vec{T}\|} (\cos \alpha_1)^{m-1} [\sin(\alpha_1 - \phi)]^{d-m-1} [\sin \alpha_1 \cos(\alpha_1 - \phi) - \cos \alpha_1 \sin(\alpha_1 - \phi)] \\
&= (-1)^{d-m} \frac{1}{\|\vec{T}\|} (\cos \alpha_1)^{m-1} [\sin(\alpha_1 - \phi)]^{d-m-1} \sin \phi.
\end{align*}
\]
For \( j, l = m + 1, \ldots, d \) but \( l \neq j \), we obtain in a similar way that

\[
n_{jl}(\vec{\alpha}) = \frac{1}{||\vec{T}||} \left\{ (\sin \alpha_d)^2 (\cos \alpha_1)^m (\cos \alpha_d)^{d-m-2} + z [\sin(\alpha_d - \phi)]^2 [\cos(\alpha_1 - \phi)]^m [\cos(\alpha_d - \phi)]^{d-m-2} \right\}
\]

\[
= -\frac{(-1)^{d-m+1}}{||\vec{T}||} \left\{ \cos^2(\alpha_1 - \phi) \cos^m(\alpha_1) [\sin(\alpha_1 - \phi)]^{d-m-2} + z \cos^2(\alpha_1) \cos^m(\alpha_1 - \phi) [\sin(\alpha_1 - \phi)]^{d-m-2} \right\}.
\]

\[
= -\frac{(-1)^{d-m+1}}{||\vec{T}||} \left\{ \cos^2(\alpha_1 - \phi) \cos^m(\alpha_1) [\sin(\alpha_1 - \phi)]^{d-m-2} - \frac{[\cos(\alpha_1)]^{m+1} [\sin(\alpha_1 - \phi)]^{d-m-1} \cos(\alpha_1 - \phi)}{\sin \alpha_1} \right\}
\]

\[
= -\frac{(-1)^{d-m+1}}{||\vec{T}|| \sin \alpha_1} \cos(\alpha_1 - \phi) \cos^m(\alpha_1) [\sin(\alpha_1 - \phi)]^{d-m-2} \{ \sin \alpha_1 \cos(\alpha_1 - \phi) - \cos \alpha_1 \sin(\alpha_1 - \phi) \}
\]

\[
= -\frac{(-1)^{d-m+1}}{||\vec{T}|| \sin \alpha_1} \cos(\alpha_1 - \phi) \cos^m(\alpha_1) [\sin(\alpha_1 - \phi)]^{d-m-2} \sin \phi.
\] (4.61)

Note that according to Equations (4.58) and (4.60),

\[
n_{ij}(\vec{\alpha}) = -n(\vec{\alpha}).
\]

Therefore, combining with (4.19), we have \( n_{ij}(\vec{\alpha}) = n_{ji}(\vec{\alpha}) = n_{kk}(\vec{\alpha}) = -n(\vec{\alpha}) \) for

\[i = 1, \ldots, m, \ j = m + 1, \ldots, d \] and \( k = 1, \ldots, d \). Let us define

\[
B = -\frac{n_{ik}(\vec{\alpha})}{n(\vec{\alpha})} \quad \text{where} \quad i, k = 1, \ldots, m, \ \text{and} \ i \neq k;
\]

\[
A = -\frac{n_{jl}(\vec{\alpha})}{n(\vec{\alpha})} \quad \text{where} \quad j, l = m + 1, \ldots, d, \ \text{and} \ j \neq l.
\]

Note that \( n = 0 \) results in the maximum points of the error function. So we study only

\[n \neq 0\]. Thus, in this case \( A \) and \( B \) are well-defined. Plugging in Equations (4.58), (4.59) and (4.61), we can obtain that

\[
B = \tan(\alpha_1) \tan(\alpha_1 - \phi),
\]

\[
A = \cot(\alpha_1) \cot(\alpha_1 - \phi).
\]

From the above equations, we find that \( AB = 1 \). Thus, we can replace \( B \) by \( A^{-1} \) in our calculations.
Now we calculate the Hessian matrix:

\[
H = \frac{2n(\vec{a})}{1 + \lambda} \cdot \begin{bmatrix}
n(\vec{a}) & -n_1(\vec{a}) & \cdots & -n_m(\vec{a}) & n(\vec{a}) & \cdots & n(\vec{a}) \\
-n_1(\vec{a}) & n(\vec{a}) & \cdots & -n_m(\vec{a}) & n(\vec{a}) & \cdots & n(\vec{a}) \\
\vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\
-n_m(\vec{a}) & -n_{2m}(\vec{a}) & \cdots & n(\vec{a}) & n(\vec{a}) & \cdots & n(\vec{a}) \\
n(\vec{a}) & n(\vec{a}) & \cdots & n(\vec{a}) & n(\vec{a}) & \cdots & -n_{m+1}(\vec{a}) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots \\
n(\vec{a}) & n(\vec{a}) & \cdots & -n_{m+1}(\vec{a}) & -n_{m+2}(\vec{a}) & \cdots & n(\vec{a}) \\
\end{bmatrix}
\]

\[
= \frac{2[n(\vec{a})]^2}{1 + \lambda}
\]

\[
= \frac{2n^2}{1 + \lambda}
\]

\[
= 2 \left( \frac{n(\vec{a})}{1 + \lambda} \right)^2
\]

\[
= 2 \left( \frac{n^2}{1 + \lambda} \right)
\]

\[
= \frac{2n^2}{1 + \lambda}
\]

(4.62)
As in Subsection 4.3.1, we write \( n(\vec{a}) \) as \( n \) for our convenience. We know that \( n^2 \) is positive because we do not consider \( n = 0 \) here.

The Hessian matrix in (4.62) can be rewritten as

\[
H = \frac{2n^2}{1 + \lambda} \begin{bmatrix}
X & Y \\
Y' & Z
\end{bmatrix},
\]

where the block matrix \( X \) is an \( m \times m \) matrix with the diagonal elements 1 and the remaining entries \( A^{-1} \), \( Y \) is an \( (m - d) \times m \) matrix with all entries 1 and \( Z \) is an \( (m - d) \times (m - d) \) matrix with 1 on the diagonal and \( A \) in the other entries.

The Hessian matrix in (4.62) needs two steps of Householder reflection. For the first step, we use the Householder matrix

\[
P_1 = \frac{1}{d - m - \sqrt{d - m}} \begin{bmatrix}
W_1 & 0 \\
0' & U_1
\end{bmatrix},
\]

where \( 0 \) is the \( (d - m) \times m \) zero matrix, \( W_1 \) is an \( m \times m \) diagonal matrix with \( (d - m - \sqrt{d - m}) \) on the diagonal, and \( U_1 \) is a \( (d - m) \times (d - m) \) matrix of the following form:

\[
U_1 = \begin{bmatrix}
\sqrt{d - m - 1} & \sqrt{d - m - 1} & \cdots & \cdots & \sqrt{d - m - 1} \\
\sqrt{d - m - 1} & d - m - \sqrt{d - m - 1} & -1 & \cdots & -1 \\
\vdots & -1 & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & -1 \\
\sqrt{d - m - 1} & -1 & \cdots & -1 & d - m - \sqrt{d - m - 1}
\end{bmatrix}.
\]

Applying \( P_1 \) on both sides of the Hessian matrix \( H \), we have

\[
P_1 H P_1 = \frac{2n^2}{(1 + \lambda)(d - m - \sqrt{d - m})^2} \begin{bmatrix}
V & \vec{u} & 0 \\
\vec{u}' & w & \vec{0}' \\
\vec{0}' & \vec{0} & \Lambda_1
\end{bmatrix},
\]
where \( V = (d - m - \sqrt{d - m})X \) is an \( m \times m \) matrix,
\[
u = (\sqrt{d - m} - 1)(d - m - \sqrt{d - m})(d - m)[1, 1, \ldots, 1]'\]
is a length-\( m \) column vector, \( \mathbf{0} \) is the \( m \times (d - m - 1) \) zero matrix, \( w = (d - m)(\sqrt{d - m} - 1)^2[1 + (d - m - 1)A] \) is a scalar, \( \mathbf{0} \) is a zero vector with length \( d - m - 1 \) and \( \Lambda_1 \) is a \( (d - m - 1) \times (d - m - 1) \) diagonal matrix with diagonal elements \( (d - m - \sqrt{d - m})^2(1 - A) \).

For the second step of the Householder reflection, we use the matrix
\[
P_2 = \frac{1}{m - \sqrt{m}} \begin{bmatrix} W_2 & \mathbf{0} \\ \mathbf{0}' & U_2 \end{bmatrix},
\]
where \( \mathbf{0} \) is the \( (d - m) \times m \) zero matrix, \( U_2 \) is the \( (d - m) \times (d - m) \) diagonal matrix with \( m - \sqrt{m} \) on the diagonal and \( W_2 \) is an \( m \times m \) matrix with the form:
\[
W_2 = \begin{bmatrix}
m - \sqrt{m} - 1 & -1 & \ldots & -1 & \sqrt{m} - 1 \\
-1 & m - \sqrt{m} - 1 & \ddots & \vdots & \sqrt{m} - 1 \\
\vdots & \ddots & \ddots & -1 & \vdots \\
-1 & \cdots & -1 & m - \sqrt{m} - 1 & \sqrt{m} - 1 \\
\sqrt{m} - 1 & \sqrt{m} - 1 & \cdots & \sqrt{m} - 1 & \sqrt{m} - 1
\end{bmatrix}.
\]
Thus, after this second step of Householder reflection, we obtain that

\[
P_2 P_1 H P_1 P_2 = \frac{2n^2}{1 + \lambda} \begin{bmatrix}
1 - A^{-1} & \cdots & 0 \\
\vdots & \ddots & \vdots \\
1 - A^{-1} & & 0 \\
1 + (m - 1)A^{-1} & \sqrt{m(d - m)} & \\
\sqrt{m(d - m)} & 1 + (d - m - 1)A & \sqrt{m(d - m)} \\
\end{bmatrix}.
\]

From (4.63), we can see that \(d - 2\) of the eigenvalues of the Hessian matrix \(H\) are

\[
x_1 = \cdots = x_{m-1} = \frac{2n^2(1 - A^{-1})}{1 + \lambda},
\]

\[
x_{m+2} = \cdots = x_d = \frac{2n^2(1 - A)}{1 + \lambda}.
\]

The remaining two of the eigenvalues \(x_m\) and \(x_{m+1}\) are the same as the eigenvalues of matrix

\[
D = \frac{2n^2}{1 + \lambda} \begin{bmatrix}
1 + (m - 1)A^{-1} & \sqrt{m(d - m)} \\
\sqrt{m(d - m)} & 1 + (d - m - 1)A \\
\end{bmatrix}.
\]

The characteristic equation of matrix \(D\) is (ignoring the constant \(2n^2/(1 + \lambda)\))

\[
x^2 - [2 + (d - m - 1)A + (m - 1)A^{-1}]x + (1 + m^2 - md) + [(d - m - 1)A + (m - 1)A^{-1}]
+ (m - 1)(d - m - 1) = 0.
\]

By solving Equation (4.66), we obtain the eigenvalues of matrix \(D\). They are
\[
x_{m,m+1} = \frac{n^2}{1 + \lambda} \left\{ 2 + (d - m - 1)A + (m - 1)A^{-1} \pm \sqrt{[(d - m - 1)A + (m - 1)A^{-1}]^2 + 4(d - 1)} \right\}.
\]

(4.67)

From Equation (4.66), we also know that the sum and product of the solutions satisfy:

\[
x_m + x_{m+1} = \frac{n^2}{1 + \lambda} \left[ 2 + (d - m - 1)A + (m - 1)A^{-1} \right]
\]

\[
= \frac{n^2}{1 + \lambda} \left[ -(d - m - 1) - (m - 1) + d + (d - m - 1)A + (m - 1)A^{-1} \right] n^2
\]

\[
= \frac{n^2}{1 + \lambda} [d - (d - m - 1)(1 - A) - (m - 1)(1 - A^{-1})] n^2,
\]

(4.68)

and

\[
x_m x_{m+1} = \frac{n^4}{(1 + \lambda)^2} \left\{ (1 + m^2 - md) + [(d - m - 1)A + (m - 1)A^{-1}] + (m - 1)(d - m - 1) \right\}
\]

\[
= \frac{n^4}{(1 + \lambda)^2} \left\{ (1 + m^2 - md) + [(d - m - 1)A + (m - 1)A^{-1}] + (m - 1)(d - m - 1) \right\}
\]

\[
= -\frac{n^4}{(1 + \lambda)^2} [(d - m - 1)(1 - A) + (m - 1)(1 - A^{-1})].
\]

(4.69)

Because \( A \) and \( A^{-1} \) have the same sign, we need to only consider two cases:

- When \( A > 0 \), we can find that \( x_m > 0 \) by (4.67). However, the sign of \( x_{m+1} \) cannot be determined.

- When \( A < 0 \), we have \( x_m x_{m+1} < 0 \) according to (4.69). Because \( x_m > x_{m+1} \), we obtain that \( x_m > 0 \) and \( x_{m+1} < 0 \).

Consider the signs of the other eigenvalues in (4.64) and (4.65). There are three cases for us to study:

1. If \( 0 < A < 1 \), then \( 1 - A > 0 \) and \( 1 - A^{-1} < 0 \). There are \( m - 1 \) repeated negative eigenvalues, \( d - m - 1 \) repeated positive eigenvalues, one positive distinct eigenvalue and another distinct eigenvalue of unknown sign.
2. If \( A > 1 \), then \( 1 - A < 0 \) and \( 1 - A^{-1} > 0 \). Similarly, there are \( m - 1 \) repeated positive eigenvalues, \( d - m - 1 \) repeated negative eigenvalues, one positive distinct eigenvalue and and another distinct eigenvalue of unknown sign.

3. If \( A < 0 \), then \( 1 - A > 0 \) and \( 1 - A^{-1} > 0 \). Thus, there are \( d - 1 \) positive eigenvalues \((m - 1)\) repeated eigenvalues that are the same as \( x_1 \), \( d - m - 1 \) repeated eigenvalues that are the same as \( x_d \), and one distinct eigenvalue) and one distinct negative eigenvalue.

Therefore, when there are some \( \alpha_i \)'s that are not the same as the others, the equilibrium points are saddle points when \( n(\vec{a}) \neq 0 \). We can conclude that there does not exist a local minimum outside of the symmetric set.

Let us remark that when \( A > 0 \), the eigenvalue \( x_{m+1} \) can indeed be positive, zero, or negative. We show this in the following way:

We first consider the possibility of \( x_{m+1} = 0 \). By (4.67) this implies that

\[
\sqrt{[(d - m - 1)A + (m - 1)A^{-1}]^2 + 4(d - 1)} = 2 + (d - m - 1)A + (m - 1)A^{-1}. 
\]

This equation can be simplified to

\[
(d - m - 1)A + (m - 1)A^{-1} = d - 2. \tag{4.70}
\]

When \( A > 0 \), we know that \( A^{-1} > 0 \). We can apply the inequality \( a + b \geq 2 \sqrt{ab} \) to obtain that

\[
(d - m - 1)A + (m - 1)A^{-1} \geq 2 \sqrt{(d - m - 1)(m - 1)}. \tag{4.71}
\]

Let us define a function \( g(m) = (d - m - 1)(m - 1) \). Because the restriction on \( m \) is \( 2 \leq m \leq d - 2 \) in this case, we can find that the minimum of \( g \) is \( g(2) = d - 3 \). Therefore, using (4.71) we can obtain that

\[
(d - m - 1)A + (m - 1)A^{-1} \geq 2 \sqrt{d - 3}. \tag{4.72}
\]
Plugging (4.70) into (4.72), we have \( d - 2 \geq 2 \sqrt{d - 3} \), which can be easily checked to be true for all \( d \geq 3 \) by squaring both sides of the inequality. Therefore, it is possible for (4.70) to hold and then \( x_{m+1} = 0 \). We also know that \( x_{m+1} > 0 \) if 
\[(d - m - 1)A + (m - 1)A^{-1} > d - 2, \text{ and } x_{m+1} < 0 \text{ if } (d - m - 1)A + (m - 1)A^{-1} < d - 2. \]
We cannot tell whether \( x_{m+1} \) is positive, negative or 0, but we will show in Section 4.4.3 that the sign of \( x_{m+1} \) does not affect our results.

### 4.4 Possibility of a swamp

In numerical approximations, the swamp phenomenon occurs when the error function decreases extremely slowly for a great number of iterations before converging at a reasonable rate at the end. In our study of using the gradient flow to minimize the error function, we are not considering any particular numerical algorithms. We only want to study how the angular variables change according to the gradient flow as defined in (4.20). This results of this study can shed light on the behavior of gradient-based algorithms.

Consider the dynamical system (4.20), the swamp phenomenon could occur if a trajectory is close to a saddle point and the saddle point has a stable direction that is stronger than all the unstable directions.

For an equilibrium point \((\alpha_1, \alpha_2, \ldots, \alpha_d)\) of the gradient flow, \(x_1, x_2, \ldots, x_d\) denote the eigenvalues to the Hessian matrix at \((\alpha_1, \alpha_2, \ldots, \alpha_d)\). Note that in (4.20), the angular variables change according to the negative gradient flow. Thus, the eigenvector corresponding to a negative eigenvalue gives us an unstable direction and the eigenvector corresponding to a positive eigenvalue gives us a stable direction. When all eigenvalues are nonzero, let us define the **saddle value** as

\[
\nu = \frac{\min\{x_i : x_i < 0\}}{\min\{x_j : x_j > 0\}}. \tag{4.73}
\]
When one or more eigenvalues are 0, say $x_k = 0$ for some $k$, if there exists at least one negative eigenvalue, we can define the saddle value as

$$
\nu = \lim_{x_k \to 0^+} \frac{\min\{x_i : x_i < 0\}}{x_k} = \infty.
$$

When $\nu < 1$, the strongest unstable direction is weaker than the weakest stable direction, then a trajectory near such an equilibrium point and following a stable direction might stay in a neighborhood of it for a relatively long time. It is generally believed that when $\nu < 1$, it is possible for the swamp behavior to occur under some gradient-based algorithm.

We can study the values of $\nu$ for the saddles in the form of $\alpha_1 \neq \alpha_2 = \cdots = \alpha_d$ and $\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d$ ($m > 1$) respectively. For $\alpha_1 \neq \alpha_2 = \cdots = \alpha_d$, we can find the parameter regions where $\nu < 1$ analytically. The results are summarized as Table 4.2 and the detailed calculations are in Subsection 4.4.1. For $\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d$ and $m > 1$, we prove in Subsection 4.4.2 that there is no parameter region where the saddle value $\nu < 1$. Numerical results are illustrated in Subsection 4.4.3.

<table>
<thead>
<tr>
<th>Dimension $d$</th>
<th>Parameter regions where $\nu &lt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = 3$</td>
<td>$\cos \phi \in (0, h_4(z))$ when $z &gt; 0$ and $(-z)^1/(d-2) \in \left(\frac{1-\cos \phi}{\sin \phi}, 1\right)$</td>
</tr>
<tr>
<td>$d = 4$</td>
<td>$(-z)^1/(d-2) \in \left(\frac{1-\cos \phi}{\sin \phi}, 1\right)$ and $\phi \neq \pi/2$</td>
</tr>
<tr>
<td>$d = 5$</td>
<td>$(-z)^1/(d-2) \in \left(\frac{1-\cos \phi}{\sin \phi}, \cos \phi\right)$</td>
</tr>
<tr>
<td>$d \geq 6$</td>
<td>$(-z)^1/(d-2) \in \left(\frac{1-\cos \phi}{\sin \phi}, h_1^{-1}(\phi)\right)$</td>
</tr>
</tbody>
</table>

### 4.4.1 Saddle value at $\alpha_1 \neq \alpha_2 = \cdots = \alpha_d$

According to the results in 4.3.1, the Hessian matrix at the equilibrium point in the form of $\alpha_1 \neq \alpha_2 = \cdots = \alpha_d$ has the following eigenvalues:
\[ x_1 = [(d - 2)A + 2 + \sqrt{(d - 2)^2A^2 + 4(d - 1)}]n^2/(1 + \lambda), \quad (4.74) \]
\[ x_2 = [(d - 2)A + 2 - \sqrt{(d - 2)^2A^2 + 4(d - 1)}]n^2/(1 + \lambda), \quad \text{and (4.75)} \]
\[ x_3 = \cdots = x_d = 2(1 - A)n^2/(1 + \lambda), \quad (4.76) \]

where \( A, n^2 \) and \( 1 - A \) are calculated in Equations (4.45), (4.44) and (4.55). We distinguish three cases, based on the signs of the eigenvalues, and discuss the values of \( \nu \) in each case.

### 4.4.1.1 Case 1: \( A < 0 \)

For this case, we know that \( 1 - A > 0 \). Thus, \( x_3 = \cdots = x_d > 0 \) and \( x_1, x_2 < 0 \). Because \( x_1 > x_2 \), we know \( x_1 > 0 \) and \( x_2 < 0 \). Therefore, \( \nu = \max \left\{ \frac{-x_2}{x_3}, \frac{-x_2}{x_1} \right\} \). By Equation (4.55), \( 1 - A > 0 \) implies

\[ (-z)^{\frac{1}{1-d}} < 0 \quad \text{or} \quad (-z)^{\frac{1}{1-d}} \in \left[ \frac{1 - \sin \phi}{\cos \phi}, 1 \right]. \quad (4.77) \]

According to (4.45), we have

\[ A^{-1} = \tan(\alpha_1) \tan(\alpha_1 - \phi). \quad (4.78) \]

In this case \( A < 0 \) implies that \( \tan(\alpha_1) \tan(\alpha_1 - \phi) < 0 \). Plugging in the value of this equilibrium point given by Equation (4.35) into (4.78), we have

\[ A^{-1} = \frac{\sin \phi}{\cos \phi - (-z)^{1/(d-2)}} \frac{\sin \phi}{\left[\cos \phi - (-z)^{-1/(d-2)}\right] \left[(-z)^{-1/(d-2)} - \cos \phi\right]}. \quad (4.79) \]

We consider two scenarios of the parameter values:

1. (i) When \( z < 0 \), i.e., \( (-z)^{1/(d-2)} > 0 \), using Equation (4.79), \( A < 0 \) is equivalent to

\[ \cos^2 \phi - \left[ (-z)^{-1/(d-2)} + (-z)^{1/(d-2)} \right] \cos \phi + 1 \]
\[ = \left( \cos \phi - (-z)^{-1/(d-2)} \right) \left( \cos \phi - (-z)^{1/(d-2)} \right) > 0. \quad (4.80) \]
Because \(0 < (-z)^{1/(d-2)} \leq 1\), we know \((-z)^{-1/(d-2)} \geq 1\). Thus, the solution to the inequality (4.80) is \(\cos \phi \in (0, (-z)^{1/(d-2)})\). This is equivalent to \((-z)^{1/(d-2)} > \cos \phi\).

Note that when \((-z)^{1/(d-2)} > 0\), we know that
\[
(-z)^{1/(d-2)} + (-z)^{-1/(d-2)} \geq 2 \sqrt{(-z)^{1/(d-2)}(-z)^{-1/(d-2)}} = 2.
\]

Then by (4.79),
\[
-A^{-1} = \frac{\sin^2 \phi}{\cos^2 \phi + 1 - [(-z)^{1/(d-2)} + (-z)^{-1/(d-2)}] \cos \phi} \\
\geq \frac{1 - \cos^2 \phi}{\cos^2 \phi + 1 - 2 \cos \phi} = \frac{1 + \cos \phi}{1 - \cos \phi} \\
\geq 1.
\]

Thus, \(1 + A \geq 0\) when \((-z)^{1/(d-2)} > 0\) (i.e., \(z < 0\)).

(ii) When \(z > 0\) and \(d\) is odd, \(-1 \leq (-z)^{1/(d-2)} < 0\). We also know that \(A < 0\) is equivalent to inequality (4.80). It is obvious that (4.80) is always true for \((-z)^{1/(d-2)} < 0\). Therefore, this case 1 requires the parameters to satisfy

\[
(-z)^{1/(d-2)} \in [-1, 0) \cup (\cos \phi, 1].
\]

Similar to the argument in (i), when \((-z)^{1/(d-2)} < 0\) (i.e., \(z > 0\) and \(d\) is odd), we have
\[
(-z)^{1/(d-2)} + (-z)^{-1/(d-2)} \leq -2. \quad \text{We can obtain that } 1 + A \leq 0.
\]

Now we can consider the value of \(v\). We want to study whether \(v < 1\) can occur in this case. Let us assume that \(-\frac{x_2}{x_3} < 1\). Using (4.75) and (4.76), we obtain
\[
\sqrt{(d-2)^2A^2 + 4(d-1) - 2 - (d-2)A} < 1.
\]

Because \(1 - A < 0\), we can multiply \(2(1 - A)\) on both sides of the above inequality:
\[
\sqrt{(d-2)^2A^2 + 4(d-1) - 2 - (d-2)A} < 2(1 - A).
\]
Simplifying this inequality yields
\[ \sqrt{(d - 2)^2 A^2 + 4(d - 1)} < dA + 4(1 - A). \]

Squaring both sides of the above inequality and simplifying it, we obtain that
\[ -dA^2 + A^2 + d - 1 < 2(1 - A)dA + 4(1 - A)^2. \]

This is equivalent to
\[ (d - 1)(1 - A)(1 + A) - 2(1 - A)(dA + 2 - 2A) < 0. \]

Thus, we can obtain that
\[ (1 - A)[d - 5 - (d - 3)A] < 0. \]

Because $1 - A > 0$ and $A < 0$, this is never true when $d \geq 5$. It is even impossible to have
\[ (1 - A)[d - 5 - (d - 3)A] \leq 0, \]
so we showed that $-\frac{z_k}{x_3} > 1$ when $d \geq 5$. Then $\nu = \max \left\{ -\frac{z_k}{x_3}, -\frac{z_1}{x_1} \right\} > 1$ for $d \geq 5$. We consider the situations when $d = 3$ and $d = 4$ separately.

(a) When $d = 3$, we have $x_1 = (A + 2 + \sqrt{A^2 + 8})n^2/(1 + \lambda)$,
\[ x_2 = (A + 2 - \sqrt{A^2 + 8})n^2/(1 + \lambda) \]
and $x_3 = \cdots = x_d = 2(1 - A)n^2/(1 + \lambda)$. We know that $x_1 + x_2 = \frac{2(A + 2)n^2}{1 + \lambda}$.

(1) $z < 0$: As shown from above in case (i) that when $z < 0$, $1 + A > 0$. Then $x_1 + x_2 > 0$,
i.e., $-\frac{z_2}{x_1} < 1$ in this case. Because of the restriction $1 - A > 0$ and $1 + A > 0$,
we know that $A^2 < 1$. We can also obtain that
\[ -\frac{x_2}{x_3} = \frac{\sqrt{A^2 + 8} - A - 2}{2(1 - A)} \]
\[ = \frac{3 - A - 2}{2(1 - A)} \]
\[ < 1/2. \]

Therefore, when $d = 3$ and $z < 0$, $\nu = \max \left\{ -\frac{z_2}{x_1}, -\frac{z_2}{x_3} \right\} < 1$. 
(2) \( z > 0 \): When \( d = 3 \) and \( z > 0 \), if \( \nu < 1 \) then both \( \frac{-z}{x_3} < 1 \) and \( \frac{-z}{x_1} < 1 \). We can find that \( \frac{-z}{x_3} < 1 \) implies \( \sqrt{A^2 + 8} < 4 - A \), which can be checked to be always true when \( A < 0 \).

As for \( \frac{-z}{x_1} < 1 \), it is equivalent to \( x_1 + x_2 > 0 \). By (4.50) and \( d = 3 \), we know that \( A > -2 \), i.e., \( -A^{-1} > 1/2 \). From Equation (4.79) we can derive that

\[
\cos^2 \phi - [(-z) + (-z)^{-1}] \cos \phi + 1 < 2 \sin^2 \phi.
\]

This inequality can be rewritten as

\[
3 \cos^2 \phi - [(-z) + (-z)^{-1}] \cos \phi - 1 < 0.
\]

The solution to this is

\[
\cos \phi \in \left( \frac{(-z) + (-z)^{-1} - \sqrt{\Delta_1}}{6}, \frac{(-z) + (-z)^{-1} + \sqrt{\Delta_1}}{6} \right),
\]

where \( \Delta_1 = \left[ (-z) + (-z)^{-1} \right]^2 + 12 \). Because \( \cos \phi \in [0, 1] \), we obtain that the parameter region for \( \nu < 1 \) is

\[
\cos \phi \in (0, h_4(z)),
\]

where \( h_4(z) = \left\{ (-z) + (-z)^{-1} + \sqrt{\left[ (-z) + (-z)^{-1} \right]^2 + 12} \right\} /6. \)

(b) When \( d = 4 \), we have \( x_1 = (2A + 2 + \sqrt{4A^2 + 12})n^2/(1 + \lambda) \),

\[
x_2 = (2A + 2 - \sqrt{4A^2 + 12})n^2/(1 + \lambda) \text{ and } x_3 = \cdots = x_d = 2(1 - A)n^2/(1 + \lambda).
\]

When \( z > 0 \) and \( d = 4 \), the equilibrium point does not exist. Thus, we only need to consider \( z < 0 \). Then \( x_1 + x_2 = \frac{4(A+1)n^2}{1+A} \geq 0 \) because \( 1 + A \geq 0 \). Thus, \( \frac{-z}{x_1} \leq 1 \) for \( d = 4 \). Here \( \frac{-z}{x_1} = 1 \) when \( A = -1 \).
Because \( A^2 \leq 1 \) given that \( 1 + A \geq 0 \) and \( 1 - A > 0 \), we can also calculate that

\[
-\frac{x_2}{x_3} = \frac{\sqrt{4A^2 + 12 - 2A - 2}}{2(1 - A)} \\
\leq \frac{4 - 2A - 2}{2(1 - A)} \\
= 1.
\]

Therefore, when \( d = 4 \), we have \( \nu = \max \left\{ -\frac{x_2}{x_1}, -\frac{x_2}{x_3} \right\} \leq 1 \). Note that \( \nu = 1 \) when \( A = -1 \), i.e., \( A^{-1} = -1 \). By (4.79), we have

\[
\cos^2 \phi - [(-z)^{1/(d-2)} + (-z)^{-1/(d-2)}] \cos \phi + 1 = \sin^2 \phi.
\]

This is equivalent to

\[
\cos \phi \left( 2 \cos \phi - [(-z)^{1/(d-2)} + (-z)^{-1/(d-2)}] \right) = 0.
\]

The solution is \( \cos \phi = 0 \) or \( \cos \phi = [(-z)^{1/(d-2)} + (-z)^{-1/(d-2)}]/2 \). Note that when \( (-z)^{1/(d-2)} > 0 \),

\[
[(-z)^{1/(d-2)} + (-z)^{-1/(d-2)}]/2 \geq 2/2 = 1.
\]

Thus, \( \cos \phi \in [0, 1) \) cannot be equal to \( [(-z)^{1/(d-2)} + (-z)^{-1/(d-2)}]/2 \). So the only condition for \( \nu = 1 \) is \( \phi = \pi/2 \).

### 4.4.1.2 Case 2: \( 0 < A < 1 \)

In this case, \( 1 - A > 0 \) is still true. Thus, \( x_1 > 0, x_2 < 0 \) and \( x_3 = \cdots = x_d > 0 \). We also know from (4.77) in Case 1 that \( 1 - A > 0 \) implies that \( (-z)^{1/2} < 0 \) or

\( (-z)^{1/2} \in \left[ \frac{1 - \sin \phi}{\cos \phi}, 1 \right] \).

On the other hand, because \( A > 0 \), by Equation (4.79), we need

\[
\frac{\sin \phi}{\cos \phi} - \frac{\sin \phi}{(-z)^{1/(d-2)} + (-z)^{-1/(d-2)} - \cos \phi} > 0.
\]

In Case 1 we showed that \( A < 0 \) is equivalent to \( (-z)^{1/2} \in [-1, 0) \cup (\cos \phi, 1] \). Therefore, on the contrary, when \( A > 0 \) we have \( (-z)^{1/2} \in (0, \cos \phi) \).
Combining with the inequality (4.77) we got from $1 - A > 0$, we have

\[ (-z)^{\frac{1}{d-2}} \in \left\{ [-1, 0) \cup \left( \left( \frac{1 - \sin \phi}{\cos \phi}, 1 \right) \right) \right\} \cap (0, \cos \phi]. \]

We can check that \( \frac{1 - \sin \phi}{\cos \phi} < \cos \phi \) for any \( \phi \in (0, \pi/2) \). Therefore, this Case 2 requires

\[ (-z)^{1/(d-2)} \in \left( \frac{1 - \sin \phi}{\cos \phi}, \cos \phi \right). \] \hfill (4.81)

Because \( x_1 + x_2 = 2n^2[(d - 2)A + 1]/(1 + A) > 0 \), then \( -\frac{x_2}{x_1} < 1 \). In order for \( \nu < 1 \) to hold, we need to have \( -\frac{x_2}{x_1} < 1 \). Similar as Case 1, we can obtain

\[ (1 - A)[d - 5 - (d - 3)A] < 0. \] \hfill (4.82)

Since \( 1 - A > 0 \) and \( A > 0 \), we only need to discuss the case when \( d - 5 - (d - 3)A < 0 \). When \( d = 3, 4 \) or 5 and \( 0 < A < 1 \), we can verify that \( d - 5 - (d - 3)A < 0 \) is always true. So \( \nu < 1 \) holds for \( d = 3, 4, \) or 5.

For \( d > 5 \), by (4.79), \( d - 5 - (d - 3)A < 0 \) is equivalent to

\[ A^{-1} = \frac{\sin \phi}{\cos \phi} - \frac{\sin \phi}{(-z)^{1/(d-2)}} (\frac{1 - \sin \phi}{\cos \phi} - \cos \phi) < \frac{d - 3}{d - 5}. \]

Because \( z < 0 \) in Case 2 according to (4.81), we can simplify the above inequality to

\[ 2 \cos^2 \phi - (d - 3) \left( (-z)^{1/(d-2)} + (-z)^{-1/(d-2)} \right) \cos \phi + (2d - 8) < 0. \] \hfill (4.83)

The discriminant is

\[ \Delta = (d - 3)^2 \left[ (-z)^{1/(d-2)} + (-z)^{-1/(d-2)} \right]^2 - 16(d - 4). \]

Due to the fact that \( (-z)^{1/(d-2)} + (-z)^{-1/(d-2)} \geq 2 \) when \( z < 0 \) in Case 2, we know that for \( d > 5 \),

\[ \Delta \geq 4(d - 3)^2 - 16(d - 4) = 4(d - 5)^2 > 0. \]

Therefore, we can solve inequality (4.83) to obtain that

\[ \cos \phi \in \left( \frac{(d - 3)}{4} \left[ (-z)^{1/(d-2)} + (-z)^{-1/(d-2)} \right] - \sqrt{\Delta}, \frac{(d - 3)}{4} \left[ (-z)^{1/(d-2)} + (-z)^{-1/(d-2)} \right] + \sqrt{\Delta} \right). \]
We can check that \(\frac{(d-3)(-z)^{1/(d-2)} + (-z)^{-1/(d-2)}}{4} \leq \sqrt{\Delta} \geq 1\). So we only need
\[
\cos \phi \in \left( \frac{(d-3)(-z)^{1/(d-2)} + (-z)^{-1/(d-2)} - \sqrt{\Delta}}{4}, 1 \right).
\] (4.84)

This Case 2 requires inequality (4.81), which is equivalent to
\[
\cos \phi \in \left( (-z)^{1/(d-2)}, \frac{2}{(-z)^{1/(d-2)} + (-z)^{-1/(d-2)}} \right).
\]

Define
\[
h_1(z) = \frac{(d-3)(-z)^{1/(d-2)} + (-z)^{-1/(d-2)} - \sqrt{\Delta}}{4},
\]
\[
h_2(z) = (-z)^{1/(d-2)}, \text{ and } h_3(z) = \frac{2}{(-z)^{1/(d-2)} + (-z)^{-1/(d-2)}}.
\]

Because \(\frac{2}{(-z)^{1/(d-2)} + (-z)^{-1/(d-2)}} < 1\), in order for \(\nu < 1\) to hold, we need
\[
\cos \phi \in (\max\{h_1(z), h_2(z)\}, h_3(z)),
\]
if \(\max\{h_1(z), h_2(z)\} < h_3(z)\) and \(d > 5\).

4.4.1.3 Case 3: \(A > 1\)

In this case, \(1 - A < 0\). According to the results in 4.3.2, we have
\[x_3 = \cdots = x_d < 0 < x_2 < x_1\] and we also know that \(1 - A < 0\) implies that
\[(-z)^{1/(d-2)} \in \left( \frac{0}{1}, \frac{1 - \sin \phi}{\cos \phi} \right).
\]

We want to study the value of \(\nu = -\frac{x_2}{x_3}\). Assuming that \(\nu = -\frac{x_2}{x_3} < 1\), we have
\[
\frac{2(A - 1)}{(d-2)A + 2 - \sqrt{(d-2)^2A^2 + 4(d-1)}} < 1.
\]

Simplifying this inequality in a similar way as what we did in Case 1, we obtain that
\[\nu > 1\] for all \(d \geq 3\) in this case.
4.4.1.4 Summary for the saddle values when $\alpha_1 \neq \alpha_2 = \cdots = \alpha_d$

Note that each case we studied above yields an open set of parameters. We can also study the value of $\nu$ on the boundary of those sets.

- On the boundary of Case 1 and Case 2, $\cos \phi = (-z)^{1/(d-2)}$. This is equivalent to $A = 0$. Then we can easily find that $-\frac{\alpha_2}{x_3} \geq 1$ when $d \geq 5$. Thus, $\nu \geq 1$ when $d \geq 5$. If $d = 3$ or 4, the results are the same as in Case 1, i.e., $\nu < 1$ for $d = 3$ and $\nu \leq 1$ for $d = 4$.

- On the boundary of Case 2 and Case 3, $(-z)^{1/(d-2)} = (1 - \sin \phi)/\cos \phi$ and $A = 1$. We can see that the eigenvalues $x_2 = x_3 = \cdots = x_d = 0$, so $\nu$ is undefined. Note that this saddle point is non-hyperbolic. When $A = 1$, we obtain the solutions that $\alpha_1 = \alpha_d$. So this saddle point is in the symmetric set.

- On the boundary of Case 1 and Case 3, $z = 0$. This would reduce the rank of the target tensor to 1, so we do not consider this situation.

Figure 4.1 illustrates the three cases of the parameter space $z \in [-1, 1] \setminus \{0\}$ and $\phi \in (0, \pi/2]$. We can summarize our results as follows (see also Table 4.2):

Case 1: When $(-z)^{1/(d-2)} \in [-1, 0] \cup [\cos \phi, 1]$, $\nu \geq 1$ except for $d = 3$ and 4.

- When $d = 3$ and $z < 0$, we have $\nu < 1$. When $d = 3$ and $z > 0$, if $\cos \phi \in (0, h_d(z))$, then $\nu < 1$.

- When $d = 4$ and $z < 0$, we have $\nu \leq 1$. The equality $\nu = 1$ happens when $\phi = \pi/2$.

Case 2: When $(-z)^{1/(d-2)} \in \left(\frac{1-\sin \phi}{\cos \phi}, \cos \phi\right)$,

- if $d = 3, 4$ or 5, then $\nu < 1$;
• if $d \geq 6$, $\max\{h_1(z), h_2(z)\} < h_3(z)$ and $\cos \phi \in (\max\{h_1(z), h_2(z)\}, h_3(z))$, then $\nu < 1$.

**Case 3:** When $(-z)^{1/(d-2)} \in \left(0, \frac{1-\sin \phi}{\cos \phi}\right)$, $\nu > 1$ for any $d$.

Figure 4.1: The three cases in the parameter space $z \in [-1, 1] \setminus \{0\}$ and $\phi \in (0, \pi/2]$ when $d = 7$. The only region where $\nu < 1$ is between the green and red curves. On the green curve $\nu$ is undefined.

### 4.4.2 Saddle value at $\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d$

For the equilibrium points in the form of $\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d$, we know from Subsection 4.3.2 that the eigenvalues of the Hessian matrix are given by Equations (4.64), (4.65) and (4.67). We can discuss the value of $\nu$ in three cases again. We show in this subsection that $\nu > 1$ for all the non-symmetric saddle points in the form of $\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d$ where $2 \leq m \leq d - 2$.

- **Case 1:** $A < 0$.

When $A < 0$, we know from Subsection 4.3.2 that $x_1 = \cdots = x_{m-1} > 0$, $x_{m+2} = \cdots = x_d > 0$, $x_m > 0$ and $x_{m+1} < 0$. So $\nu = -\frac{x_{m+1}}{\min\{x_1, x_d, x_m\}}$. By (4.64)) and
(4.67), we can find that

\[ x_m - x_1 = \frac{n^2}{1 + A} \left\{ (d - m - 1)A + (m + 1)A^{-1} - \sqrt{[(d - m - 1)A + (m + 1)A^{-1}]^2 + 4(d - 1)} \right\}. \]

Because \( A < 0 \) and \( A^{-1} < 0 \), we obtain that \( x_m - x_1 < 0 \). Therefore, \( x_m < x_1 \).

Similarly, according to (4.65) and (4.67), we have

\[ x_m - x_d = \frac{n^2}{1 + A} \left\{ (d - m + 1)A + (m - 1)A^{-1} - \sqrt{[(d - m - 1)A + (m + 1)A^{-1}]^2 + 4(d - 1)} \right\}. \]

Thus, \( x_m - x_d < 0 \) when \( A < 0 \). So \( x_m < x_d \). Therefore, \( \nu = -\frac{x_{m+1}}{x_m} \).

We can rewrite Equation (4.68) as

\[ x_m + x_{m+1} = \frac{2n^2}{1 + A} [2 + A + A^{-1} + (d - m - 2)A + (m - 2)A^{-1}]. \]

Because \( A < 0 \), we find that \( A + A^{-1} \leq -2 \) and \( A^{-1} = -1 \) only when \( A = A^{-1} = -1 \). We also know that when \( A < 0 \), then \((d - m - 2)A \leq 0\) and \((m - 2)A^{-1} \leq 0\) because \( 2 \leq m \leq d - 2 \). However, when \( m \) is fixed, \((d - m - 2)A\) and \((m - 2)A^{-1}\) cannot both be 0. Therefore, \( x_m + x_{m+1} < 0 \), i.e., \( -x_{m+1} > x_m \). So we conclude that in this case \( \nu > 1 \).

**Case 2:** \( 0 < A < 1 \) and \((d - m - 1)A + (m - 1)A^{-1} \neq d - 2 \).

From the results in Subsection 4.3.2, we know that \( x_1 = \cdots = x_{m-1} < 0 \),
\( x_{m+2} = \cdots = x_d > 0 \) and \( x_m > 0 \). Note that we let \((d - m - 1)A + (m - 1)A^{-1} \neq d - 2\), thus \( x_{m+1} \neq 0 \) according to the discussion in Subsection 4.2.3 about Equation 4.70.

So we only consider \( x_{m+1} \) is either positive or negative in this case, as well as in the following Case 3. If \( x_{m+1} > 0 \), then \( \nu = -\frac{x_1}{\min\{x_m, x_{m+1}, x_d\}} \). It is easy to see from (4.64) and (4.65) that

\[ -\frac{x_1}{x_d} = \frac{1}{A} > 1. \]
Then we obtain that \( \nu \geq -\frac{x_1}{x_d} > 1 \). If \( x_{m+1} < 0 \), then
\[
\nu = -\frac{\min\{x_1, x_{m+1}\}}{\min\{x_m, x_d\}} \geq -\frac{x_1}{\min\{x_m, x_d\}} \geq -\frac{x_1}{x_d} > 1.
\]
Therefore, \( \nu > 1 \) regardless of the sign of \( x_{m+1} \).

- **Case 3:** \( A > 1 \) and \( (d - m - 1)A + (m - 1)A^{-1} \neq d - 2 \).

Similar as Case 2, we know that \( x_1 = \cdots = x_{m-1} > 0, x_{m+2} = \cdots = x_d < 0, x_m > 0 \) and \( x_{m+1} \) can be either positive or negative. If \( x_{m+1} > 0 \), we can derive that
\[
\nu = -\frac{x_d}{\min\{x_1, x_m, x_{m+1}\}} \geq -\frac{x_d}{x_1} = A > 1.
\]
If \( x_{m+1} < 0 \), we have
\[
\nu = -\frac{\min\{x_d, x_{m+1}\}}{\min\{x_1, x_m\}} \geq -\frac{x_d}{\min\{x_1, x_m\}} \geq -\frac{x_d}{x_1} = A > 1.
\]
Therefore, \( \nu > 1 \) in this case as well.

We showed that \( \nu > 1 \) when \( 2 \leq m \leq d - 2 \) for all the parameter values except when \( (d - m - 1)A + (m - 1)A^{-1} = d - 2 \). If \( (d - m - 1)A + (m - 1)A^{-1} = d - 2 \), then we have a non-hyperbolic equilibrium point of the gradient flow. At this equilibrium point, the eigenvalues of the Hessian matrix are
\[
x_1 = \cdots = x_{m-1} = \frac{2n^2(d - m - 1)(A - 1)}{(1 + \lambda)(m - 1)},
x_m = \frac{2n^2d}{1 + \lambda},
x_{m+1} = 0,
x_{m+2} = \cdots = x_d = \frac{2n^2(1 - A)}{1 + \lambda}.
\]
We can see that this non-hyperbolic equilibrium point is not in the symmetric set. However, \( x_1 = \cdots = x_{m-1} \) and \( x_{m+2} = \cdots = x_d \) always have opposite signs as long as \( A \neq 1 \). When \( A \neq 1 \), this non-hyperbolic equilibrium point is not interesting for us to
study the swamps because we show below that the limits of the saddle value \( \nu \) at
\[ x_{m+1} \to 0^+ \text{ and } x_{m+1} \to 0^- \] are greater than 1.

For Case 2, we can find that,
\[
\lim_{x_{m+1} \to 0^+} \nu = -\frac{x_d}{\min\{x_1, x_m, x_{m+1}\}} = \infty,
\]
\[
\lim_{x_{m+1} \to 0^-} \nu = -\frac{\min\{x_d, x_{m+1}\}}{\min\{x_1, x_m\}} = -\frac{x_d}{\min\{x_1, x_m\}} > 1.
\]

For Case 3, the same results can be obtained in a similar way.

In conclusion, we show in this subsection that \( \nu > 1 \) for all the non-symmetric saddle
points in the form of \( \alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d \) where \( 2 \leq m \leq d - 2 \).

### 4.4.3 Numerical results

Given parameters \( z, \phi \) and \( d \), the equilibrium point of the gradient flow in the form of
\( \alpha_1 \neq \alpha_2 = \cdots = \alpha_d \) can be found by solving Equations (4.35) and (4.36). The eigenvalues
of the Hessian matrix at this equilibrium points are given by (4.74), (4.75) and (4.76).

Therefore, for any \( |z| < 1 \), \( z \neq 0 \) and \( \phi \in (0, \pi/2) \), we can numerically find the saddle value
\( \nu \) and plot \( \nu \) as a function of \( z \) and \( \phi \). For the special cases \( d = 3 \) and \( d = 4 \), the plots for \( \nu \)
are in Figure 4.2. When \( d = 3 \) and \( d = 4 \), \( \nu < 1 \) happens both in Case 1 and Case 2. According
to the results in 4.4.1, for \( d = 3 \), the inequality \( \nu < 1 \) holds when \( (-z)^{1/(d-2)} \in (\cos \phi, 1] \),
\( z > 0 \), and \( (-z)^{1/(d-2)} \in \left(\frac{1-\sin \phi}{\cos \phi}, \cos \phi\right) \). For \( d = 4 \), when \( (-z)^{1/(d-2)} \in (\cos \phi, 1] \) and
\( (-z)^{1/(d-2)} \in \left(\frac{1-\sin \phi}{\cos \phi}, \cos \phi\right) \), we showed that \( \nu < 1 \). Therefore, we can see a relatively large
region in the parameter space for \( \nu < 1 \) when \( d = 3 \) or 4.

When \( d \geq 5 \), the parameter region for \( \nu < 1 \) is reduced to only \( \cos \phi \in (h_2(z), h_1(z)) \) as
shown in Figure 4.3. We can see that as \( d \) gets larger, the region where \( \nu < 1 \) becomes
narrower.

For the equilibrium points in the form of \( \alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \alpha_d \ (m > 1) \), we can
find the saddle value \( \nu \) numerically in a similar way. Given \( z, \phi, d \) and \( m \), we know the
Figure 4.2: Saddle value $\nu$ at each parameter pair $((-z)^{1/(d-2)}, \cos \phi)$ for (a) $d = 3$ and (b) $d = 4$. The dark red region corresponds to where $\nu$ is very large, the maximum of $\nu$ in both figures is on the order of $10^7$. The region where $\nu < 1$ is under the black curves. We can see that in (a) the inequality $\nu < 1$ holds in a small region where $z > 0$ and $\cos \phi$ is small as well as where $(-z)^{1/(d-2)} \in \left(\frac{1-\sin \phi}{\cos \phi}, 1\right]$. For (b), the inequality $\nu < 1$ holds only when $(-z)^{1/(d-2)} \in \left(\frac{1-\sin \phi}{\cos \phi}, 1\right]$.

equilibrium point satisfies Equation (4.33). We cannot solve this equation analytically, but we can use the bisection method to find its numerical solution between $-\pi/2$ and $\pi/2$. The eigenvalues of the Hessian matrix at this equilibrium point are given by Equations (4.64), (4.65) and (4.67). Then we can obtain the value of $\nu = -\frac{\min|x_i; x_i < 0\}|}{\min|x_i; x_i > 0\}}$ for the given parameters. Note that the equilibrium point with $\alpha_1 = \cdots = \alpha_m \neq \alpha_{m+1} = \cdots = \alpha_d$ has the same eigenvalues as the equilibrium point in the form of

$$\alpha_1 = \cdots = \alpha_{d-m} = \alpha_{d-m+1} = \cdots = \alpha_d.$$ So we only need to consider $2 \leq m \leq \lfloor d/2 \rfloor$.

Figures 4.4, 4.5 and 4.6 show the results from numerical computations. When $d$ is odd, the equilibrium point of the gradient flow exists for all $z, \phi$ and $m$. So we need to look at the parameter space where $z \in [-1, 1] \setminus \{0\}$ and $\phi \in (0, \pi/2]$ (Figure 4.4). When $d$ is even and $m$ is odd, such equilibrium point only exists when $z < 0$. Then in this situation we can restrict the parameter space to $z \in [-1, 0)$ and $\phi \in (0, \pi/2]$ (Figure 4.5). When $d$ and $m$ are
Figure 4.3: Saddle value \( \nu \) at each parameter pair \(((-z)^{1/(d-2)}, \cos \phi)\) for (a) \( d = 5 \), (b) \( d = 6 \), (c) \( d = 10 \) and (d) \( d = 15 \). The regions where \( \nu < 1 \) are enclosed by black curves. (a) For \( d = 5 \), \( \nu < 1 \) occurs under all parameter values in Case 2. When \( z > 0 \), \( \nu < 1 \) cannot happen. (b) For \( d = 6 \), only part of the parameter region \((\cos \phi \in (h_2(z), h_1(z)))\) in Case 2 yields \( \nu < 1 \). (c) When \( d = 10 \), the region \( \cos \phi \in (h_2(z), h_1(z)) \) becomes smaller. (d) When \( d = 15 \), \( \nu > 1 \) in the region where \((-z)^{1/(d-2)} < 0\), so we did not plot this left half of the parameter space. We see only a narrow region \( \cos \phi \in (h_2(z), h_1(z)) \) where \( \nu < 1 \).

both even, the equilibrium point exists when \( z > 0 \) and certain values of \( z < 0 \). So we can see an irregular region in the parameter space where \( \nu \) is calculated (Figure 4.6).
Figure 4.4: Values of $\nu$ in the parameter space $z \in [-1, 1] \setminus \{0\}$ and $\phi \in (0, \pi/2]$ when (a) $d = 5$, $m = 2$ and (b) $d = 9$, $m = 4$. On the dark blue line at $z = 0$, $\nu$ is undefined. We see $\nu > 1$ on the entire parameter space.

Figure 4.5: Values of $\nu$ when $d = 6$ and $m = 3$ on the parameter space $z \in [-1, 0)$ and $\phi \in (0, \pi/2]$. We can see that there is no region where $\nu < 1$.

From Figures 4.4, 4.5 and 4.6, we do not see any parameter region where $\nu < 1$.

Actually, numerical results show that as $d$ gets larger, the region where $\nu < 2$ becomes smaller and it vanishes for certain $d$ (Figure 4.4(b)). When $d = 4$ and $m = 2$
Figure 4.6: Values of $\nu$ for $m = 2$ on the parameter space $z \in [-1, 1] \setminus \{0\}$ and $\phi \in (0, \pi/2]$ when (a) $d = 4$ and (b) $d = 6$. The dark blue region in $z < 0$ is where the equilibrium point of the gradient flow does not exist. We see that there is no region in the parameter space where $\nu < 1$. When $d = 4$ there is a yellow region where $\nu > 1$ and $\nu$ is close to 1.

(Figure 4.6(a)), in the lower right yellow region $\nu$ is close to 1 but still greater than 1. Therefore, we numerically verified that when $2 \leq m \leq \lfloor d/2 \rfloor$, the inequality $\nu > 1$ is always true for such non-symmetric equilibrium points of the gradient flow.

4.5 Summary

In this chapter, we studied the application of gradient flow to minimize the error function when approximating the rank-2 symmetric target tensor using a rank-1 guess tensor. Note that we studied how the angular variables $\{\alpha_i\}_{i=1}^d$ change according to the gradient flow of the error function instead of the actual approximation algorithms. We had already known that the local minimum of the error function exists in the symmetric set $\{\alpha_1 = \cdots = \alpha_d\}$. Outside of the symmetric set, we have found that all the equilibrium points of the gradient flow are either the solutions to (4.26) and (4.33), or the solutions to $n(\tilde{\alpha}) = 0$. These solutions to $n(\tilde{\alpha})$ are the maximum point of the error function. Those
non-symmetric equilibria which are solutions to (4.26) and (4.33) exist under a relatively wide range of parameters as shown in Table 4.1.

It was shown in Section 7.3 of [9] and [15] that the best rank-1 approximation to a real symmetric tensor can be chosen symmetric. In other words, if there is a local minimum outside of the symmetric set, then there is an equally good or better local minimum in the symmetric set. We have shown that in fact all non-symmetric equilibria of the gradient flow are saddle points or maximum points in Proposition 4.3.1. So there is no local minimum outside of the symmetric set. For such non-symmetric saddle points, we also defined and studied the saddle value $\nu$. It is generally believed that when $\nu < 1$, it is possible for the swamp behavior to occur. We showed that only when $d$ is small and the saddle point is in the form of $\alpha_1 \neq \alpha_2 = \cdots = \alpha_d$ there is a wide parameter region (as in Table 4.2) where the strongest unstable direction is weaker than the weakest stable direction. In this parameter region, when the trajectory approaches the saddle point following a direction that is close to a stable direction, it will stay in a neighborhood of the saddle point for a relatively long time before it leaves this neighborhood, which might result in the swamp behavior.
REFERENCES


