# SYMMETRY BREAKING BIFURCATIONS OF THE INFORMATION DISTORTION

by
Albert Edward Parker III

A dissertation submitted in partial fulfillment
of the requirements for the degree
of
Doctor of Philosophy
in
Mathematics

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 2003

APPROVAL

of a dissertation submitted by

Albert Edward Parker III

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

Tomáš Gedeon _____ _____
(Signature) Date

Approved for the Department of Mathematics

Kenneth L. Bowers _____ _____
(Signature) Date

Approved for the College of Graduate Studies

Bruce McLeod _____ _____
(Signature) Date

STATEMENT OF PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U. S. Copyright Law. Requests for extensive copying or reproduction of this dissertation should be referred to Bell & Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Signature _____
Date _____

This thesis is dedicated
to my mother Eirene Parker,
and to my father Albert Edward Parker Jr.

# ACKNOWLEDGEMENTS

First, it is necessary to express my deep gratitude to my advisor, Tomáš Gedeon. It is his insight on which I have relied when the messy details became overbearing. Without his support, encouragement, and occasional cattle prodding, this thesis would not have been possible. His intense dedication and curiosity have been inspiring. Thank you for guiding me on such a rich and interesting problem!

I have also benefited immensely from working closely with Alex Dimitrov, who provided the germ for the class of problems which we examine in this thesis. From our many fruitful discussions, I have learned much more than just about data manipulation, mathematics, and neuroscience.

I am indebted to John Miller and Gwen Jacobs for their dedication to graduate education at Montana State University-Bozeman. Their support of my education as a mathematician striving to learn neuroscience can not be over emphasized. I would also like to thank the National Science Foundation for their support of the IGERT program, which has been the primary source of the funding for three of the last four years of my studies.

Lastly, and most importantly, I thank my sweetheart, Becky Renee Parker, for her unconditional love and support.

TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABSTRACT

The goal of this thesis is to solve a class of optimization problems which originate from the study of optimal source coding systems. Optimal source coding systems include quantization, data compression, and data clustering methods such as the Information Distortion, Deterministic Annealing, and the Information Bottleneck methods. These methods have been applied to problems such as document classification, gene expression, spectral analysis, and our particular application of interest, neural coding. The class of problems we analyze are constrained, large scale, nonlinear maximization problems. The constraints arise from the fact that we perform a stochastic clustering of the data, and therefore we maximize over a finite conditional probability space. The maximization problem is large scale since the data sets are large. Consequently, efficient numerical techniques and an understanding of the bifurcation structure of the local solutions are required. We maximize this class of constrained, nonlinear objective functions, using techniques from numerical optimization, continuation, and ideas from bifurcation theory in the presence of symmetries. An analysis and numerical study of the application of these techniques is presented.

# CHAPTER 1

## INTRODUCTION

The goal of this thesis is the solution of a class of optimization problems which originate from the study of optimal source coding systems. A problem in this class is of the form

$$\max_{q\in\Delta} \left(G(q) + \beta D(q)\right) \tag{1.1}$$

where $\beta \in [0, \infty)$, $\Delta$ is a subset of $\Re^n$, the usual $n$ dimensional vector space on the reals, and $G$ and $D$ are sufficiently smooth real valued functions.

Source coding systems are those which take a set of $K$ objects, $Y = \{y_i\}_{i=1}^K$, and represent it with a set of $N < K$ objects or *classes*, $Y_N = \{\nu_i\}_{i=1}^N$. Examples include data compression techniques (such as converting a large bitmap graphics file to a smaller jpeg graphics file) and data classification techniques (such as grouping all the books printed in 2002 which address the martial art Kempo). Both data compression and data classification techniques are forms of data clustering methods. Some stipulations that one might require of any such method is that the clustered data, $\{\nu_i\}$, represents the original data reasonably well, and that the implementation of the method runs relatively quickly.

Rate Distortion Theory [17, 35] is a mathematical framework which rigourously defines what we mean by "representing the original data reasonably well" by defining a cost function, $D(Y, Y_N)$, called a *distortion function*, which measures the difference between the original data $Y$ and the clustered data $Y_N$. Once one has a distortion function, and a data set, the method of Deterministic Annealing (DA) [61] is an algorithm that could be implemented to cluster the data quickly. The DA method is an approach to data clustering which has demonstrated marked performance improvements over other clustering algorithms [61]. The DA method actually allows for a stochastic assignment of the data $\{y_i\}_{i=1}^K$ to the clusters $\{\nu_i\}_{i=1}^N$. That is, the data $y_j$ belongs to the $i^{th}$ cluster $\nu_i$ with a certain probability, $q(\nu_i|y_j)$. Observe that we may view $q$ as a vector in some subspace $\Delta$ of $\Re^{NK}$. The subspace $\Delta$ is the space of valid discrete conditional probabilities in $\Re^{NK}$. The DA algorithm finds an *optimal* clustering, $q^*$, of the data by maximizing the level of randomness, called the entropy $H(q, C)$, at a specified level of distortion, $D(q, C) = D(Y, Y_N)$. We have written $H$ and $D$ as functions of $q$ and of the *centroids* of the clusters $C = \{c_i\}_{i=1}^N$, where $c_i$ is the centroid (or mean) of cluster $\nu_i$. This optimization problem can be written as

$$\begin{array}{l} \max_{C,q\in\Delta} H(q, C) \quad \text{constrained by} \\ D(q, C) \leq D_0, \end{array} \tag{1.2}$$

where $D_0 > 0$ is some maximum distortion level.

The Information Distortion method [22, 20, 29] uses the DA scheme to cluster neural data $Y = \{y_i\}_{i=1}^K$ into classes $\{\nu_i\}_{i=1}^N$ to facilitate the search for a *neural coding scheme* in the cricket cercal sensory system [29, 25, 24]. The neural coding problem, which we will describe in detail in the next section, is the problem of determining the stochastic correspondence, $p(X, Y)$, between the stimuli, $X = \{x_i\}$, presented to some sensory system, and the neural responses, $Y = \{y_i\}$, elicited by these stimuli. One of the major obstacles facing neuroscientists as they try to find a coding scheme is that of having only limited data [37]. The limited data problem makes a nonparametric determination of $p(X, Y)$ impossible, and makes parametric estimations (using, say, Poisson or Gaussian models, which we describe in the next section) tenuous at best. For example, it is extremely difficult to estimate the covariance matrix $C_{X,Y}$ when fitting a Gaussian model to neural data. One way to make parametric estimations more feasible is to optimally cluster the neural responses into classes $\{\nu_i\}$, and then to fit a Gaussian model to $p(X|\nu)$ for each class $\nu$. This yields $p(X, Y_N)$, by

$$p(X = x, Y_N = \nu) = p(x|\nu)p(\nu),$$

which is an approximation to $p(X, Y)$. This is the approach used by the Information Distortion method to find a neural coding scheme [29, 25, 24]. The optimal clustering $q^*(Y_N|Y)$ of the neural responses is obtained by the Information Distortion method by solving an optimization problem of the form

$$\begin{aligned} &\max_{q\in\Delta} H(q) \quad \text{constrained by} \\ &D_I(q) \leq D_0 \end{aligned} \tag{1.3}$$

where $D_0 > 0$ is some maximum distortion level, and the distortion function $D_I$ is the *information distortion measure*. Before explicitly defining $D_I$, we first explain the concept of the *mutual information* between $X$ and $Y$, denoted by $I(X; Y)$, which is the amount of information that one can learn about $X$ by observing $Y$ (see (2.4) for an explicit definition). The information distortion measure can now be defined as

$$D_I(q) = I(X; Y) - I(X; Y_N).$$

Thus, if one were interested in minimizing $D_I$, one must assure that the mutual information between $X$ and the clusters $Y_N$ is as close as possible to the mutual information between $X$ and the original space $Y$. Since $I(X, Y)$ is a fixed quantity, then if we let $D_{eff} := I(X, Y_N)$, the problem (1.3) can be rewritten as

$$\begin{aligned} &\max_{q\in\Delta} H(q) \quad \text{constrained by} \\ &D_{eff}(q) \geq I_0 \end{aligned}$$

where $I_0 > 0$ is some minimum information rate. Using the method of Lagrange multipliers, this problem can be rewritten as

$$\max_{q\in\Delta} \left(H(q) + \beta D_{eff}(q)\right), \tag{1.4}$$

for some $\beta \in [0, \infty)$, which is of the form given in (1.1).

As we have seen, Rate Distortion Theory provides a rigorous way to determine how well a particular set of clusters $Y_N = \{\nu_i\}$ represents the original data $Y = \{y_i\}$ by defining a distortion function. The basic question addressed by Rate Distortion Theory is that, when compressing the data $Y$, what is the minimum informative compression, $Y_N$, that can occur given a particular distortion $D(Y, Y_N) \leq D_0$ [17]? This question is answered for independent and identically distributed data by the Rate Distortion Theorem, which states that the minimum compression is found by solving the *minimal information problem*

$$\begin{aligned} &\min_{q\in\Delta} I(Y;Y_N) \quad \text{constrained by} \\ &D(Y;Y_N) \leq D_0 \end{aligned} \tag{1.5}$$

where $D_0 > 0$ is some maximum distortion level.

The Information Bottleneck method is a clustering algorithm which has used this framework for document classification, gene expression, neural coding [64], and spectral analysis [70, 78, 69]. The information distortion measure $D_I$ is used, so that an optimal clustering $q^*$ of the data $Y$ is found by solving

$$\begin{aligned} &\min_{q\in\Delta} I(Y;Y_N) \quad \text{constrained by} \\ &D_I \leq D_0. \end{aligned}$$

As we saw with the Information Distortion optimization problem, we rewrite this problem as

$$\begin{aligned} &\max_{q\in\Delta} -I(Y;Y_N) \quad \text{constrained by} \\ &D_{eff} \geq I_0. \end{aligned}$$

Now the method of Lagrange multipliers gives the problem

$$\max_{q\in\Delta} -I(Y;Y_N) + \beta D_{eff}(q), \tag{1.6}$$

for some $\beta \in [0, \infty)$, which is of the form given in (1.1).

A basic *annealing* algorithm, various forms of which have appeared in [61, 22, 29, 78, 70], can be used to solve (1.1) (which includes the cases (1.4) and (1.6)) for $\beta = \mathcal{B}$, where $\mathcal{B} \in [0, \infty)$.

Algorithm 1 (Annealing). *Let*

$$q_0 \text{ be the maximizer of } \max_{q\in\Delta} G(q) \tag{1.7}$$

*and let $\beta_0 = 0$. For $k \geq 0$, let $(q_k, \beta_k)$ be a solution to (1.1). Iterate the following steps until $\beta_{\mathcal{K}} = \mathcal{B}$ for some $\mathcal{K}$.*

1. *Perform $\beta$-step: Let $\beta_{k+1} = \beta_k + d_k$ where $d_k > 0$.*

2. *Take* $q_{k+1}^{(0)} = q_k + \eta$, *where* $\eta$ *is a small perturbation, as an initial guess for the solution* $q_{k+1}$ *at* $\beta_{k+1}$.

3. *Optimization: solve*

$$\max_{q \in \Delta} G(q) + \beta_{k+1} D(q)$$

*to get the maximizer* $q_{k+1}$, *using initial guess* $q_{k+1}^{(0)}$.

The purpose of the perturbation in step 2 of the algorithm is due to the fact that a solution $q_{k+1}$ may get "stuck" at a suboptimal solution $q_k$. The goal is to perturb $q_{k+1}^{(0)}$ outside of the basin of attraction of $q_k$.

To illustrate how Algorithm 1 works, we now examine its results when employed by the Information Distortion method to solve (1.4). We consider the synthetic data set $p(X, Y)$, shown in figure 1(a), which was drawn from a mixture of four Gaussians as the authors did in [22, 29]. In this model, we may assume that $X = \{x_i\}_{i=1}^{52}$ represents a range of possible stimulus properties and that $Y = \{y_i\}_{i=1}^{52}$ represents a range of possible neural responses. There are four *modes* in $p(X, Y)$, where a mode of a probability distribution can be thought of as the areas in the space $(X, Y)$ which have high probability. Each mode corresponds to a range of responses elicited by a range of stimuli. For example, the stimuli $\{x_i\}_{i=1}^{15}$ elicit the responses $\{y_i\}_{i=39}^{52}$ with high probability, and the stimuli $\{x_i\}_{i=25}^{36}$ elicit the responses $\{y_i\}_{i=22}^{38}$ with high probability. One would expect that the maximizer $q^*$ of (1.4) will cluster the neural responses $\{y_i\}_{i=1}^{52}$ into four classes, each of which corresponds to a mode of $p(X, Y)$. This intuition is justified by the Asymptotic Equipartition Property for jointly typical sequences, which we present as Theorem 13 in Chapter 2.

The mutual information $I(X, Y)$ is about 1.8 bits, which is comparable to the mutual information conveyed by single neurons about stimulus parameters in several unrelated biological sensory systems [21, 41, 58, 72]. For this analysis we used the joint probability $p(X, Y)$ explicitly to evaluate $H(q) + \beta D_{eff}(q)$, as opposed to modelling $p(X, Y)$ by $p(X, Y_N)$ as explained in the text. The annealing algorithm (Algorithm 1) was run for $0 \leq \beta \leq 2$.

The optimal clustering $q^*(Y_N|Y)$ for $N = 2$, 3, and 4 is shown in panels (b)–(d) of figure 1. We denote $Y_N$ by the natural numbers, $Y_N = \{1, ..., N\}$. When $N = 2$ as in panel (b), the optimal clustering $q^*$ yields an incomplete description of the relationship between stimulus and response, in the sense that responses $\{y_i\}_{i=1}^{12} \cup \{y_i\}_{i=39}^{52}$ are in class $\nu_1 = 1$ and responses $\{y_i\}_{i=13}^{38}$ are in class $\nu_2 = 2$. The representation is improved for the $N = 3$ case shown in panel (c) since now $\{y_i\}_{i=1}^{12}$ are in class $\nu_1 = 1$, and $\{y_i\}_{i=39}^{52}$ are in a separate class, $\nu_2 = 2$. The responses $\{y_i\}_{i=13}^{38}$ are still lumped together in the same class $\nu_3 = 3$. When $N = 4$ as in panel (d), the elements of $Y$ are separated into the classes correctly and most of the mutual information is recovered (see panel(f)). The mutual information in (f) increases with the number of classes approximately as $\log_2 N$ until it recovers about 90% of the original mutual information (at $N = 4$), at which point it levels off.

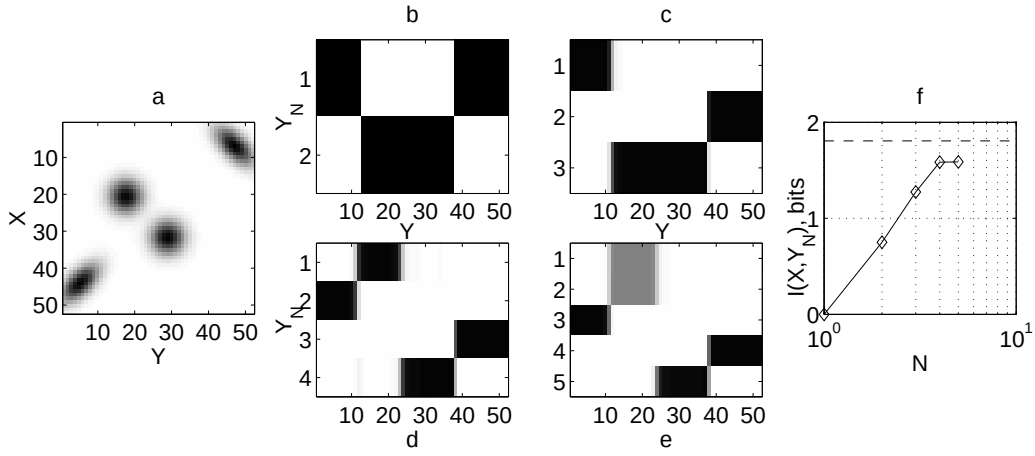



Figure 1. *The Four Blob Problem* from [22, 29]. (a) A joint probability for the relation $p(X, Y)$ between a stimulus set $X$ and a response set $Y$, each with 52 elements. (b–d) The optimal clusterings $q^*(Y_N|Y)$ for $N = 2$, 3, and 4 classes respectively. These panels represent the conditional probability $q(\nu|y)$ of a class $\nu$ being associated with a response $y$. White represents $q(\nu|y) = 0$, black represents $q(\nu|y) = 1$, and intermediate values are represented by levels of gray. In (e), a clustering is shown for $N = 5$. Observe that the data naturally splits into 4 clusters because of the 4 modes of $p(X, Y)$ depicted in panel (a). The behavior of the effective distortion $D_{eff} = I(X; Y_N)$ with increasing $N$ can be seen in the log-linear plot (f). The dashed line is $I(X; Y)$, which is the least upper bound of $I(X; Y_N)$.

It has been observed that the solutions $(q, \beta)$ of (1.1), which contain the sequence $\{(q_k, \beta_k)\}$ found in step 3 of Algorithm 1, undergo *bifurcations* or *phase transitions* as $\beta \to \mathcal{B}$ [61, 22, 29, 78, 70]. (see Figure 2). The explicit form of some of these solutions about bifurcation points for the Information Distortion problem (1.4) are given in Figure 3.

The behavior of $D_{eff}$ as a function of $\beta$ can be seen in the top panel. Some of the solutions $\{(q_k, \beta_k)\}$ for different values of $\beta_k$ are presented on the bottom row (panels 1 – 6). One can observe the bifurcations of the solutions (1 through 5) and the corresponding transitions of $D_{eff}$. The abrupt transitions ($1 \to 2$, $2 \to 3$) are similar to the ones described in [61] for a different distortion function. One also observes transitions ($4 \to 5$) which appear to be smooth in $D_{eff}$ even though the solution from $q_k$ to $q_{k+1}$ seems to undergo a bifurcation.

The bifurcation structure outlined in Figure 3 raises some interesting questions. Why are there only 3 bifurcations observed? In general, are there only $N - 1$ bifurcations observed when one is clustering into $N$ classes? In Figure 3, observe that $q \in \Re^{4K} = \Re^{208}$. Why should we observe only 3 bifurcations to local solutions of $H + \beta D_{eff}$ in such a large dimensional space? What types of bifurcations should

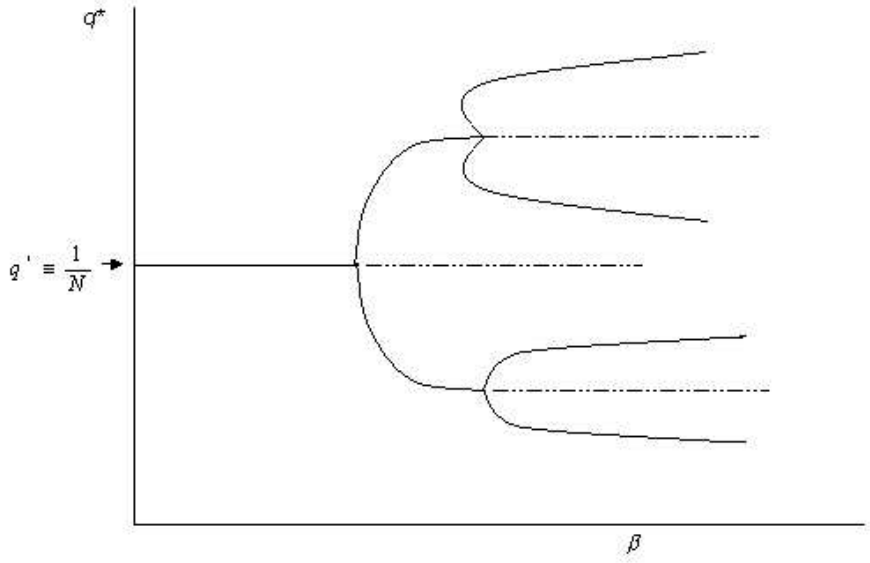



Figure 2. Conceptual bifurcation structure of solutions $(q^*, \beta)$ to the problem (1.1) as a function of the parameter $\beta$. In this instance, the first solution is denoted as $q_{\frac{1}{N}}$, the clustering of the data such that $q(Y_N|Y) = \frac{1}{N}$ for every $\nu \in Y_N$ and every $y \in Y$.

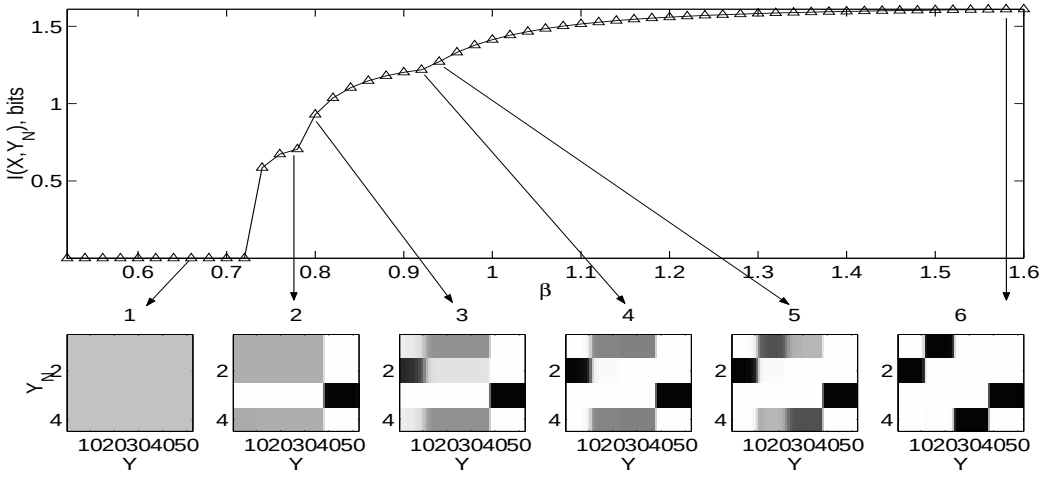



Figure 3. [22, 29] Observed bifurcations of the solutions $(q^*, \beta)$ to the Information Distortion problem (1.4). For the data set in Figure 1a, the behavior of $D_{eff} = I(X; Y_N)$ (top) and the solutions $q(Y_N|Y)$ (bottom) as a function of $\beta$.

we expect: pitchfork-like, transcritical, saddle-node, or some other type? At bifurcation, how many bifurcating branches are there? What do the bifurcating branches look like: are they *subcritical* or *supercritical* (sometimes called *first order* and *second order* phase transitions respectively)? What is the stability of the bifurcating branches? In particular, from bifurcation of a solution, is there always a bifurcating branch which contains solutions of the original optimization problem?

For problems of the form

$$\max_{q \in \Delta} F(q, \beta), \tag{1.8}$$

where

$$F(q, \beta) = G(q) + \beta D_{eff}(q),$$

which include the problems posed by the Information Distortion (1.4) and Information Bottleneck (1.6) methods, we have addressed these questions. We considered the bifurcation structure of all *stationary points* of (1.8), which are points $q \in \Re^{NK}$ that satisfy the necessary conditions of constrained optimality, known as the Karush-Kuhn-Tucker Conditions (see Theorem 16). In this way, we have been able to answer many of the questions about the bifurcation structure just posed.

The foundation upon which we have relied to effect these answers is the theory of bifurcations in the presence of symmetries [33, 34, 71]. The symmetries in the case of (1.8) are based upon the observation that any solution $(q^*(Y_N|Y), \beta)$ to (1.8) gives another equivalent solution simply by permuting the labels of the classes of $Y_N$ (see chapter 6). This symmetry can be seen in Figure 1 in any of the panels (a)–(e). Permuting the numbers on the vertical axis just changes the labels of the classes $Y_N = \{1, ..., N\}$, and does not affect the value of the cost function $G(q) + \beta D_{eff}(q)$ (this is proved rigorously for the problem (1.4) in Theorem 73). For example, if $P_1$ and $P_2$ are two $K \times 1$ vectors such that for a solution $q^*(Y_N|Y)$, $q^*(1|Y) = P_1$ and $q^*(2|Y) = P_2$, then the clustering $\hat{q}$ where $\hat{q}(1|Y) = P_2$, $\hat{q}(2|Y) = P_1$, and $\hat{q}(Y_N|Y) = q^*(Y_N|Y)$ for all other classes $\nu$, is also a maximizer of (1.8), since $F(\hat{q}, \beta) = F(q^*, \beta)$.

We will use $S_N$ to denote the well known algebraic group of all permutations on $N$ symbols [8, 27]. We say that $F(q, \beta)$ is $S_N$-*invariant* if $F(q, \beta) = F(\sigma(q), \beta)$ where $\sigma(q)$ denotes the action on $q$ by permutation of the classes of $Y_N$ as defined by the element $\sigma \in S_N$. Now suppose that a solution $q^*$ is fixed by all the elements of $S_M$ for $1 < M \leq N$. A bifurcation at $\beta = \beta^*$ in this scenario is called *symmetry breaking* if the bifurcating solutions are fixed (and only fixed) by subgroups of $S_M$. Under some generic conditions (Assumptions 81), we are able to use the Equivariant Branching Lemma [34] (Theorem 47) and the Smoller-Wasserman Theorem [71] (Theorem 49) to show that if there is a bifurcation point on a solution branch that is fixed by $S_M$ for $1 < M \leq N$, then symmetry breaking bifurcation occurs. The Equivariant Branching Lemma in this instance gives explicit bifurcating directions of the $M$ bifurcating solutions, each of which has symmetry $S_{M-1}$.

The theory of bifurcation in the presence of symmetries gives us the following answers to the questions posed above. There are only $N - 1$ bifurcations observed

when one is clustering into $N$ classes because there are only $N-1$ symmetry breaking bifurcations along certain paths of bifurcating branches. In particular, there are $N-1$ subgroups of $S_N$ in the *partial lattice* or "chain of subgroups"

$$1 < S_2 < ... < S_{N-1} < S_N.$$

The first solution branch, $(q_0, \beta)$, where $q_0$ is the uniform distribution $q_{\frac{1}{N}}$, has symmetry of the full group $S_N$. When bifurcation occurs on this branch, the symmetry dictates that there are at least $N$ bifurcating branches, each with symmetry $S_{N-1}$ (Corollary 111 and the Equivariant Branching Lemma). Each of these branches undergoes symmetry breaking bifurcation at some point later on, with at least $N-1$ bifurcating branches, each with symmetry $S_{N-2}$ (Theorem 110 and the Equivariant Branching Lemma), and so on. Once we are on a solution branch where there is no symmetry (in other words, symmetry $S_1$), then we have shown that, generically, further bifurcations are not possible (Theorem 114).

We have shown that all symmetry breaking bifurcations from $S_M$ to $S_{M-1}$ are pitchfork-like (Theorem 120 and see Figures 16–24). Furthermore, we have ascertained the existence of other types of bifurcating branches from symmetry breaking bifurcation which we did not expect (see Figure 25).

In fact, we have shown that the observed bifurcation structure given in Figure 3, although qualitatively correct, is "shifted" in $\beta$ (see Figure 20 and Remark 152).

We have derived a condition, called the *bifurcation discriminator*, which predicts whether all of the branches from a symmetry breaking bifurcation from $S_M$ to $S_{M-1}$ are either subcritical or supercritical (Theorems 127 and 128). We have confirmed this result numerically for the subcritical bifurcations that occur, for example, from the $q_{\frac{1}{N}}$ solution branch for $N \geq 3$ for the Four Blob Problem (see Table 3 and Figures 16, 17 and 24). We have also numerically confirmed that subcritical bifurcations occur on other branches as well (Figure 22).

It is a well known fact that subcritical bifurcating branches are unstable (Theorem 127). We have also provided a condition which ascertains the stability of supercritical branches (Theorem 128). We have shown that, in some instances, unstable branches can not contain solutions to (1.9) (Theorem 129). For example, the subcritical bifurcating branches in Figure 16 contain stationary points which are not solutions of the problem (1.8). Thus, we have shown that a local solution to the optimization problem (1.8) does not always persist from a symmetry breaking bifurcation. This would explain why, in practice, solving (1.1) after bifurcation incurs significant computational cost [29, 61].

Symmetry breaking bifurcations are not the only bifurcations. The existence of subcritical bifurcating branches implies that *saddle-node* bifurcations or *folds* may occur. We have confirmed numerically that these "non-symmetry breaking" bifurcations do indeed exist (Figures 16, 17, 22, and 24). Furthermore, we show that, generically, saddle-node bifurcations are the only type of non-symmetry breaking bifurcations. We also give necessary and sufficient conditions for the existence of saddle-node bifurcations (chapter 8).

Although we had (1.8) in mind as we developed the mathematical framework in this thesis, we have been able to generalize the theory so that it applies to a class of optimization problems. We conclude this section by giving the form of a problem in this class, which is

$$\max_{q\in\Delta} F(q,\beta), \tag{1.9}$$

where

$$F(q,\beta) = G(q) + \beta D(q), \tag{1.10}$$

and $q$ is a discrete conditional probability $q(Y_N|Y)$, a stochastic map of the realizations of some random variable $Y$ to the realizations of a random variable $Y_N$. The space $\Delta$ is the linear constraint space of valid conditional probabilities,

$$\Delta := \left\{ q(Y_N|Y) \mid \sum_{\nu} q(\nu|y) = 1 \text{ and } q(\nu|y) \geq 0 \ \forall y \in Y \right\}. \tag{1.11}$$

The goal is to solve (1.9) for $\beta = \mathcal{B} \in [0,\infty)$. Further assumptions on the functions $G$ and $D$ are the following.

Assumption 2.

1. *$G$ and $D$ are real valued functions of $q(Y_N|Y)$, which depend on $Y_N$ only through $q$, are invariant to relabelling of the elements or* classes *$\nu$ of $Y_N$. That is, $G$ and $D$ are $S_N$-invariant.*

2. *$G$ and $D$ are sufficiently smooth in $q$ on the interior of $\Delta$.*

3. *The Hessians of $G$ and $D$ are block diagonal.*

As we have seen, similar problems arise in Rate Distortion Theory (1.5), Deterministic Annealing (1.2), the Information Distortion method (1.4), and the Information Bottleneck method (1.6).

## Neural Coding

The motivating factor for the work presented in this thesis is the efficient implementation of the Information Distortion method [22, 20, 29]. The objective of the Information Distortion is to allow a quantitative determination of the type of information encoded in neural activity patterns and, at the same time, identify the code with which this information is represented. In spite of the fact that the explicit objective of the method is deciphering the neural code, the method could be applied

to cluster any system of pairs of the inputs and outputs. This versatility has already been exhibited by the Information Bottleneck method [70, 78, 69].

This section is organized as follows. First, we describe in detail the neural coding problem, first with words, then by building the mathematical framework. We continue by surveying some of the methods used to determine coding schemes in many different sensory systems. This prepares the reader for the following section, which provides an overview of how the Information Distortion method searches for an answer to the neural coding problem.

We begin with Dimitrov and Miller's formulation of the neural coding problem [22].

> The early stages of neural sensory processing encode information about sensory stimuli into a representation that is common to the whole nervous system. We will consider this encoding process within a probabilistic framework [4, 41, 59].
>
> One of the steps toward understanding the neural basis of an animal's behavior is characterizing the code with which its nervous system represents information. All computations underlying an animal's behavioral decisions are carried out within the context of this code.
>
> Deciphering the neural code of a sensory system means determining the correspondence between neural activity patterns and sensory stimuli. This task can be reduced further to three related problems: determining the specific stimulus parameters encoded in the neural ensemble activity, determining the nature of the neural symbols with which that information is encoded, and finally, quantifying the correspondence between these stimulus parameters and neural symbols. If we model the coding problem as a correspondence between the elements of an input set $\mathcal{X}$ and an output set $\mathcal{Y}$, these three tasks are: finding the spaces $\mathcal{X}$ and $\mathcal{Y}$, and the correspondence between them.
>
> Any neural code must satisfy at least two conflicting demands. On the one hand, the organism must recognize the same natural object as identical in repeated exposures. On this level the response of the organism needs to be *deterministic*. On the other hand, the neural code must deal with uncertainty introduced by both external and internal noise sources. Therefore the neural responses are by necessity *stochastic* on a fine scale [19, 86](see Figure 4).
>
> In this respect the functional issues that confront the early stages of any biological sensory system are similar to the issues encountered by communication engineers in their work of transmitting messages across noisy media. Thus, tools from information theory can be used to characterize the neural coding scheme of a simple sensory system.

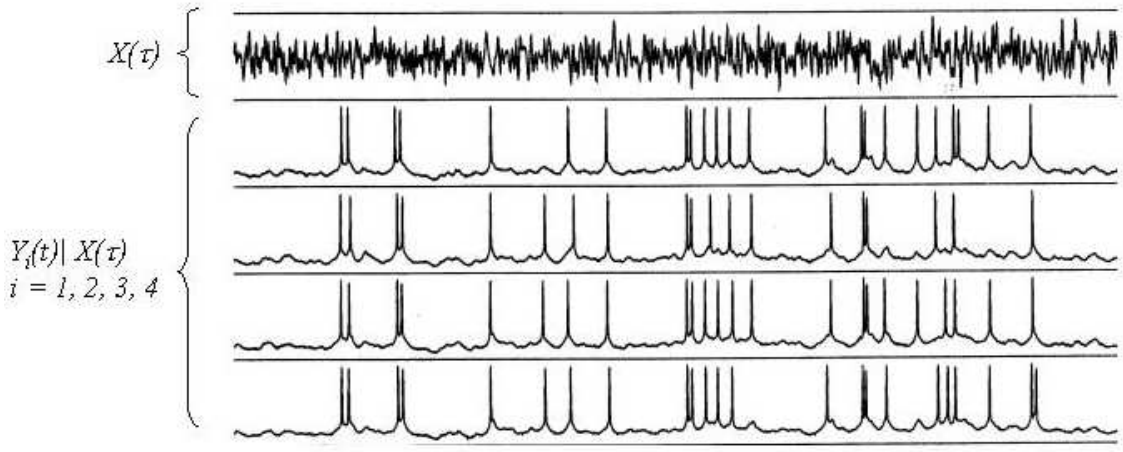



Figure 4. The neural response to a static stimulus is stochastic. Presenting an identical stimulus, $X(\tau) = x$, four separate times to a biological sensory system produces four distinct neural responses, $Y = y_1, y_2, y_3, y_4$.

One can model the input/output relationship present in a biological sensory system as an *optimal information channel* $(X, Y)$ [68], where $X$, is a random variable of inputs

$$X : \Omega_X \to \mathcal{X}, \tag{1.12}$$

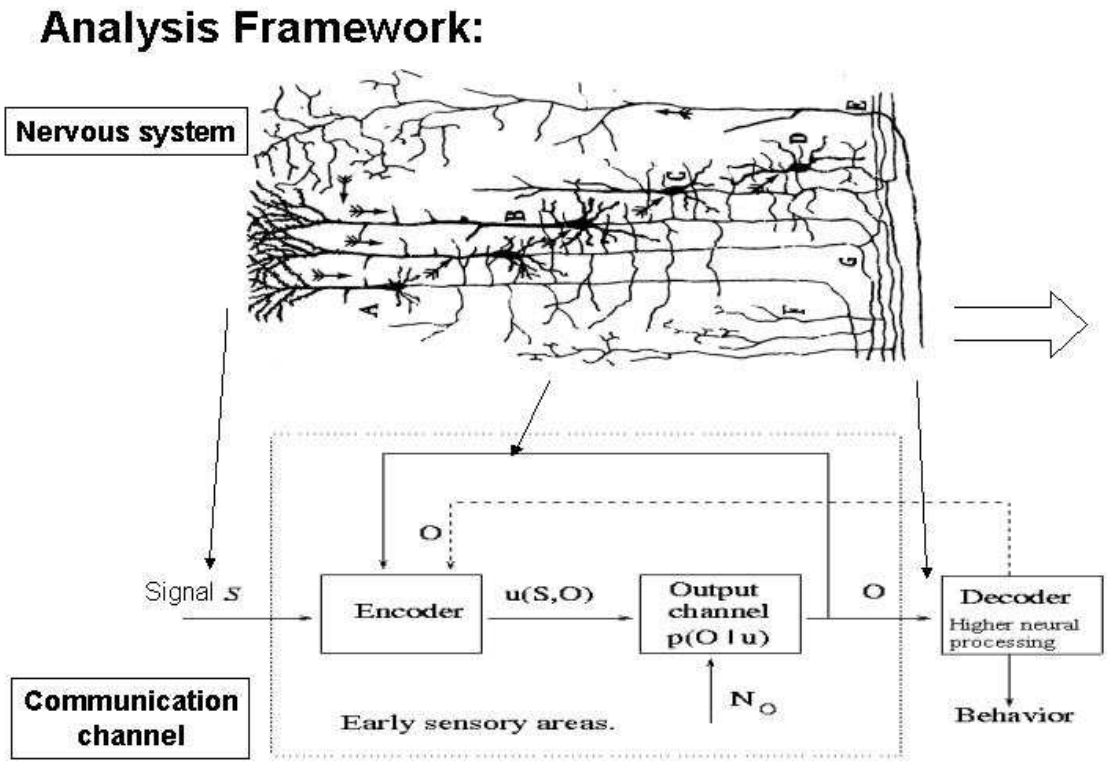
and $Y$ is a random variable of outputs

$$Y : \Omega_Y \to \mathcal{Y} \tag{1.13}$$

(see Figure 5).

When translating the structure of an information channel to neural systems, the output space $\Omega_Y$ from (1.13) is usually the set of activities of a group of neurons, which is potentially an infinite dimensional space, since we assume that the neural response is some function of the voltage at each point in physical space of the cell's membrane, for each cell in the group, at each instance of time. Instead of considering the membrane potential at every instance of time, it is common practice to assume that the *spikes* (the sharp modes of the neural responses in Figure 4) are the only relevant features of the neural response. If the neural response is divided up into $k$ time bins, and if we let a 1 indicate the presence and 0 indicate the absence of a spike in a particular time bin of the neural response, then we let $Y$ represent $\Omega_Y$ as the finite dimensional measurable space $\mathcal{Y} = \{0, 1\}^k$. Thus, each neural response is modelled as a sequence of $k$ zeroes and ones, $Y = Z^k$, where $Z \in \{0, 1\}$, so that only the temporal patterns of spikes is taken into account. For the physiological data presented in this thesis, the length of a time bin is on the order of $100\mu s$ and $k = 100$.

A




B

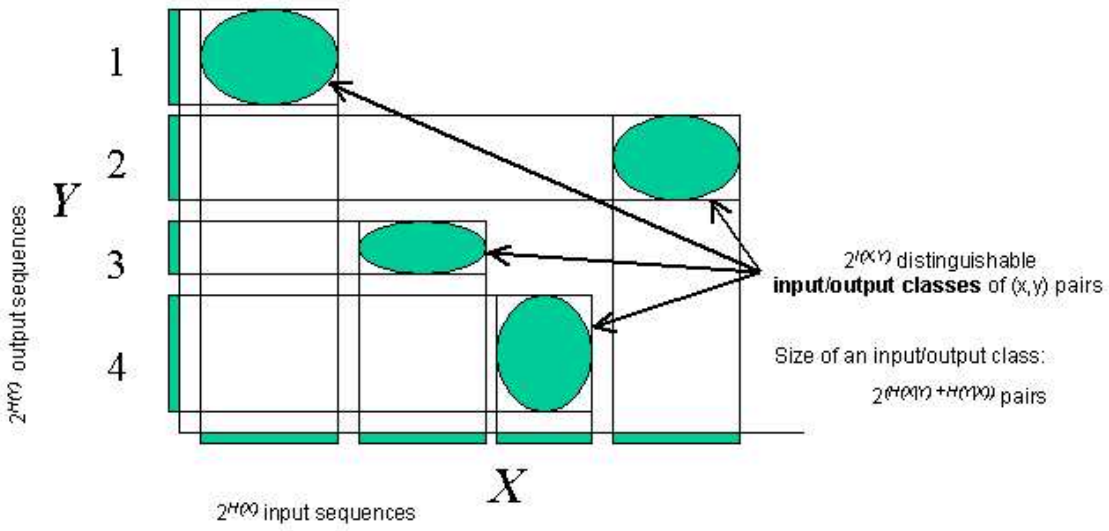



Figure 5. A: Modelling a sensory system as a communication channel. B: The structure, $p(X, Y)$, of an optimal communication system.

Thus, a neural response of length 10 ms is represented by $Y$ as a sequence of 100 zeros and ones.

Another common representation of the neural response, called the *firing rate*, is given by

$$\tilde{Y} : \mathcal{Y} \to \tilde{\mathcal{Y}}, \tag{1.14}$$

where $\tilde{\mathcal{Y}}$ is the space of real numbers $\Re$. $\tilde{Y}$ represents either the number of spikes which occur in some window of time which is large with respect to the time bins which contain the individual spikes, or it is the *mean firing rate*, an average of the spike count over several time bins. These time windows ranges anywhere from $10 - 500$ms in the neurophysiological literature [59, 67].

The input space $\Omega_X$ can be sensory stimuli from the environment or the set of activities of another group of neurons. It is also potentially an infinite dimensional space. Elements of the space of visual stimuli, for example, would represent the visual scene at different locations in physical space at each instance in time. Many times when the input is sensory stimuli from the environment, one assumes that $\mathcal{X} = \Re^K$, where $\Re^K$ is the $K$ dimensional vector space on the real numbers. If we let $K = km$ for some positive integers $k$ and $m$, then we have that $\mathcal{X} = \Re^{km} = (\Re^m)^k$. In this context, $X$ can be written as $X = W^k$ where $W$ is a random variable

$$W : \Omega_X \to \Re^m,$$

and interpreted as an $m$ dimensional representation of the stimulus $X \in \mathcal{X}$ at time $k$.

The correspondence between stimuli and responses, the joint probability $p(X, Y)$, is called a *coding scheme* [22, 73]. The input $X = W^k$ is produced by a source with a probability $p(X)$. The output $Y = Z^k$ is produced with probability $p(Y)$. The *encoder* $p(Y|X)$ is a stochastic mapping from $\mathcal{X}$ to $\mathcal{Y}$. From the point of view of information theory, the designation of spaces $\mathcal{X}$ and $\mathcal{Y}$ as an input and output space is arbitrary. Thus we can choose to characterize the same information channel as a source $Y$ with probability $p(Y)$ and a *decoder* stochastic mapping $p(X|Y)$ from $\mathcal{Y}$ to $\mathcal{X}$ (see Figure 6).

Neural Coding through the Ages

We continue by surveying some of the methods used to determine coding schemes in many different sensory systems. These methods can be partitioned into two categories. *Neural encoding* methods find approximations of the encoder $p(Y|X)$. *Neural decoding* methods find approximations to the decoder $p(X|Y)$.

Neural Encoding. Perhaps the simplest description of neural encoding is *spike count coding*, commonly called *rate coding*, first observed in the classic early work of Adrian and Zotterman [2, 3] in 1926. Adrian and Zotterman hung weights of different masses from a muscle, and measured the activity of a stretch receptor neuron

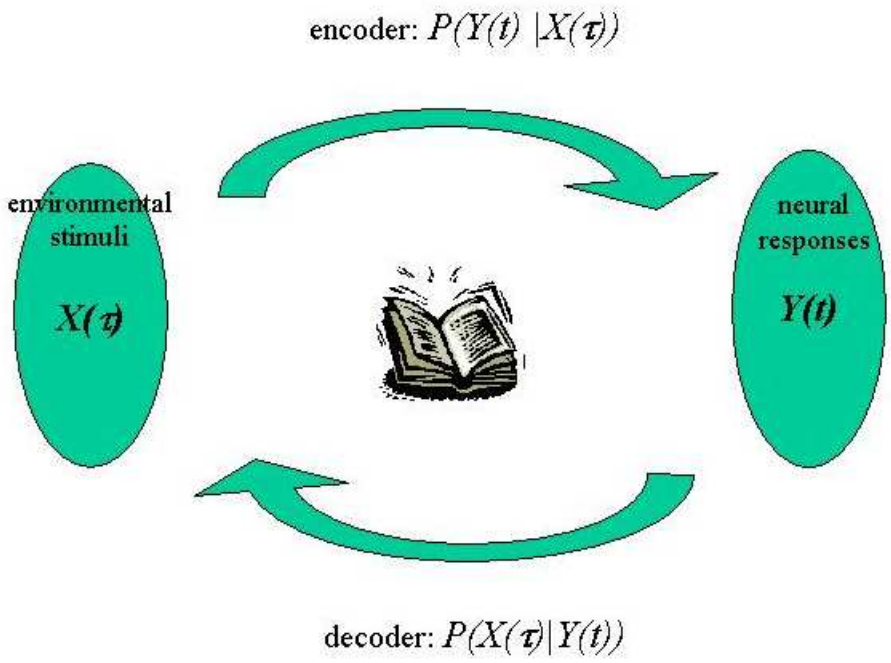



Figure 6. Probability framework, showing the spaces produced by $X(\tau)$ and $Y(t)$, and the stochastic mappings $p(Y|X)$ and $p(X|Y)$ between them. Discovering either of these mappings defines a dictionary between classes of stimuli and classes of responses, where the classes are defined by $p(X, Y)$ as in Figure 5B. We use two different time variables, $\tau$ and $t$, to make the distinction that the stimuli $X$ may occur during different intervals of time than do the neural responses $Y$.

embedded in the muscle [59]. They found that the firing rate, $\tilde{Y}$ as defined in (1.14), of the stretch receptor cell increased with increasing stimulus strength (weights with more mass). This common relationship, called the *response tuning curve*, (Figure 7A) is evidenced in many sensory systems [59]. For example, moving a static pattern across the visual field of a blowfly [59] and recording from the fly's motion sensitive neuron $H1$, also yields a response tuning curve as in Figure 7A. In this case, the stimulus amplitude is the average velocity of the pattern, over a 200ms window. Similarly, blowing wind with uniform intensity from many different directions across a cricket yields the *directional tuning curve* when recording from the four interneurons of the cricket cercal sensory system [48] as in Figure 7B.

Figures 7A and 7B suggest that, even in this simple encoding regime, neural encoding is *not* a linear process.

To estimate the encoder $p(\tilde{Y}|X)$, an experimenter could, in principle, repeat each stimulus $x \in \mathcal{X}$ many times, giving the density depicted in Figure 8. Since the experimenter controls $p(X = x)$ (the probability of observing a realization of the

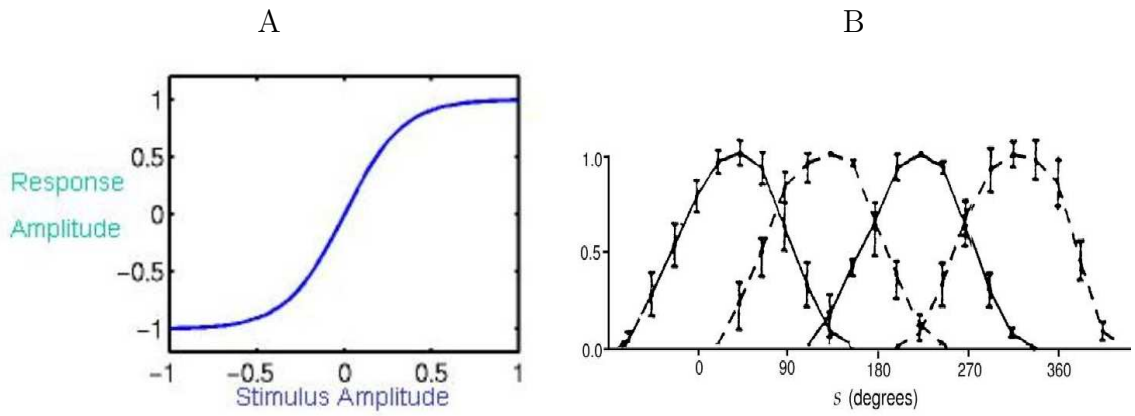



Figure 7. A: The response tuning curve. In *spike count* or *rate* coding, the response amplitude is $\tilde{Y}$, which we define as the number of spikes present in some time window. The stimulus amplitude is represented by some scalar. B: The Directional Tuning Curve. Another example of spike count coding. The response or directional tuning curves for the 4 interneurons in the cricket cercal sensory system, where the stimulus amplitude is given by direction of the wind with respect to the cricket in degrees, and the response amplitude is $\tilde{Y}$. The *preferred directions*, (the *center of mass* or *modes* of the tuning curves) are orthogonal to each other [48].

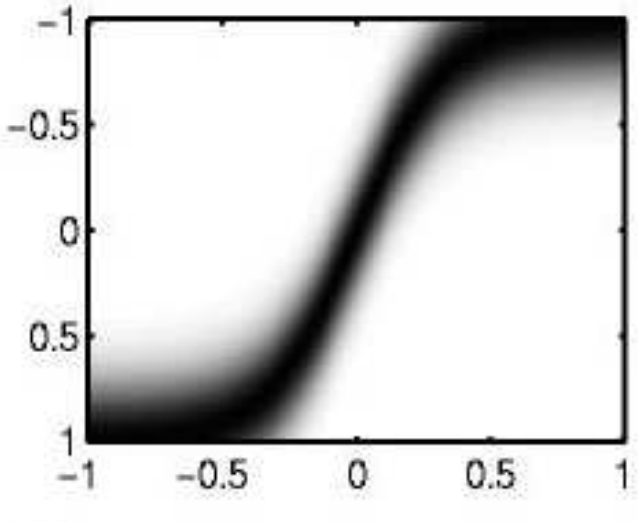



Figure 8. An estimate of the encoder $p(\tilde{Y}|X)$, using spike count coding, by repeating each stimulus $x \in \mathcal{X}$ many times, creating a histogram for each $\tilde{y}|X$, and then normalizing.

stimulus $X = x$), one can then calculate

$$p(\tilde{Y} = \tilde{y}) = \sum_{\mathcal{X}} p(\tilde{y}|x)p(x).$$

Bayes Rule [28] then yields the decoder

$$p(x|\tilde{y}) = p(\tilde{y}|x)p(x)\frac{1}{p(\tilde{y})}.$$

Spike count coding does seem to describe some sensory systems well [59], and is an attractive method due to its simplicity, especially when the stimulus space is small (i.e. a few dimensions), as in the case of coding direction in the cricket cercal sensory system [48, 63]. There are at least three points arguing why spike count coding is not a feasible way to describe an arbitrary sensory system. First, counting spikes per unit of time neglects the temporal precision of the spikes of the neural response, which potentially decreases the information conveyed by the response [52, 53, 66, 62, 57, 56]. In the visual system, it has been conjectured that firing rates are useful for gross discrimination of stimuli, while a temporal code is necessary for more subtle differences [57]. Secondly, the known short behavioral decision times (for, say, defensive maneuvering of a blowfly or of a cricket) imply that these decisions are made based on the observation of just a few spikes (1 or 2 in a 10-30ms window in some instances [59, 77]) from the sensory system which instigates the decision, and not on some large window of time. The third reason is that many sensory systems, such as the visual, auditory and olfactory systems, respond to stimulus attributes that are very complex. In other words, $\Omega_X$, the space of possible stimuli for some systems, is a very large space, which is not clearly representable by a small space $\mathcal{X}$ to be presented in an experiment. Hence, it is not feasible to present all possible stimuli in experiment to estimate $p(\tilde{Y}|X)$.

Another way to describe neural encoding , first used by Fatt and Katz in 1952 [79], is by fitting a Poisson model [28] to the data

$$p(\tilde{Y} = \tilde{y}|X = x) = \text{Poisson}(\lambda) := \frac{e^{-\lambda}\lambda^{\tilde{y}}}{\tilde{y}!}$$

for some rate $\lambda$. This model presupposes that the spikes are independent from each other given a stimulus $X = x$. Determining $\lambda$ for a given realization $X = x$ of the stimulus is straightforward. One starts by computing the *peristimulus time histogram* (PSTH), $r(t|X = x)$, the normalized histogram of the neural responses $Y|x$ over many repetitions of the stimulus $X = x$ (see Figure 9A). The PSTH $r(t|X = x)$ gives the probability per unit time of observing a spike given that $X = x$ occurred [79, 59]. The Poisson rate is

$$\lambda = \int r(t|X = x)dt,$$

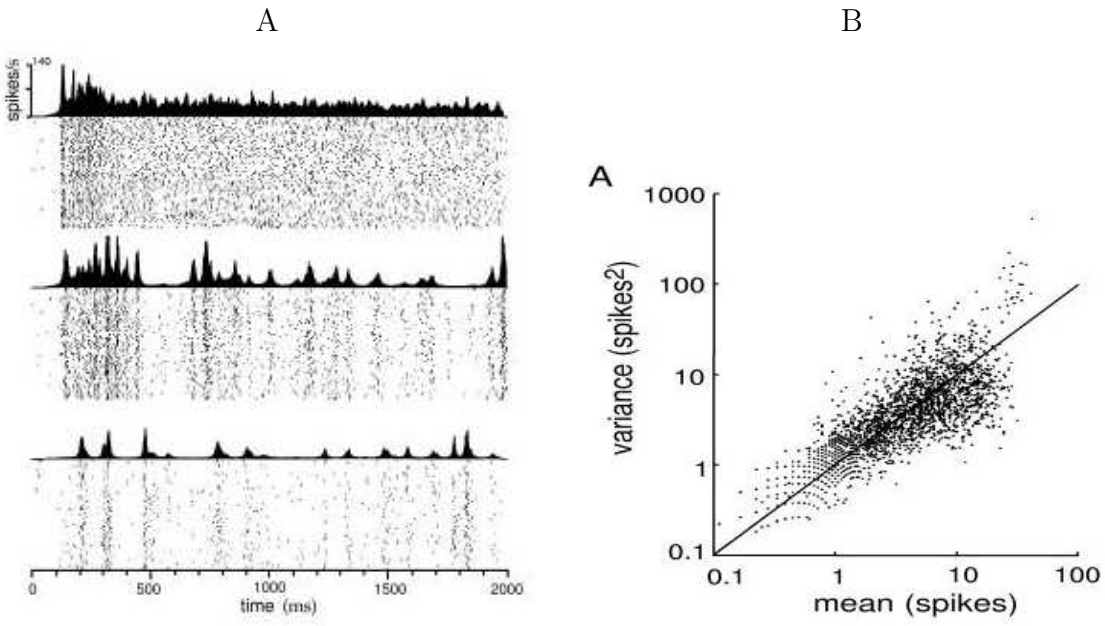



Figure 9. Both panels are from [1]. A: Examples of a peristimulus time histogram for three different stimuli $x_1, x_2, x_3$, not shown. Below each PSTH is the raster plot of associated neural responses $Y|x_i$ over many repetitions of the stimulus $X = x_i$. The PSTH is the normalized histogram of the raster plot. B: Testing to see if the firing rate given a particular realization of a stimulus, $\tilde{Y}|X = x$ is *not* a Poisson process. A true Poisson process has population mean equal to population variance, and so by the large Law of Large Numbers, for a large enough data size, the sample mean and sample variance must be very nearly equal.

which is the average number of spikes given that $X = x$. Thus

$$p(\tilde{Y}|X = x) = \text{Poisson}\left(\int r(t|X = x)dt\right). \tag{1.15}$$

The relation (1.15) yields an explicit form of $p(\tilde{Y}|X = x)$, which is alluring since a Poisson process is a basic, well studied process. But when is the assumption that the spikes are independent met? One way to test whether a process is *not* a Poisson process is to test whether the sample mean is equal to the sample variance. Such a test for neurological data is shown in figure 9B.

Rieke et al. contend that if the *refractory period* of a neuron is small compared to the mean *interspike interval* (ISI), then a Poisson model may be appropriate [59]. Berry and Meister have proposed a variant of the Poisson model which deals with the refractory period and its implications regarding the independence assumption [5].

Another shortcoming of the Poisson model as posed in (1.15) is that it only considers the neural response as the firing rate $\tilde{Y}$. In order to model a spike train

$Y = Z^N$, Rieke et al. suggest a "Poisson-like" model [59]. If $t_i$ is the beginning of one of the $N$ time bins which define $Y = y$, and $T$ is the total length of time of the neural response $Y = y$, then

$$p(Y = y|X = x) = \frac{1}{N!}\Pi_{i=1}^{N} r(t_i|X = x) \exp\left(\int_0^T r(t|X = x)dt\right).$$

In this case, the implicit assumption is that the neural responses $Y$ are independent.

Other Poisson-like processes which dispense with the independence assumption are the so called Inhomogeneous Poisson Gaussian and Inhomogeneous Poisson Zernike models used by Brown et al. to model the encoder $p(Y|X)$ [11]. These models use a generalization of the Poisson rate parameter $\lambda$ which is history dependent and so independence of the neural responses is not necessary.

The strongest argument posed against the spike count coding model applies here as well: since the space of possible stimuli for some systems is a very large space, it is not possible to present all possible stimuli in experiments to estimate $r(t|X)$ (and hence to estimate $p(Y|X)$).

The last neural encoding model which we investigate here employs the celebrated *Wiener/Volterra series*. The Volterra series, discovered by Volterra in 1930, is a series expansion for a continuous function, such as $\tilde{Y}(t)$, provided that $\tilde{Y}(t) = G(X(\tau))$ for some functional $G$ that satisfies some regularity conditions [85, 59, 80]. The series is given by

$$\begin{array}{rcl} Y(t) = f_0 & + & \int f_1(\tau_1)X(t-\tau_1)d\tau_1 \\ & + & \int\int f_2(\tau_1,\tau_2)X(t-\tau_1)X(t-\tau_2)d\tau_1 d\tau_2 + ... \end{array}. \tag{1.16}$$

Wiener in 1958 reformulated the Volterra series in a way such that the coefficient functions or *kernels* $f_i$ could be measured from experiment [59, 87, 80]. The *first Wiener kernel* is

$$f_1 = \frac{X * Y}{S_X},$$

where $X * Y$ is the convolution of $X$ and $Y$, and $S_X = X * X$ is the power spectrum of $X$ [59]. $f_1$ is proportional to the *spike triggered average*. Rieke et al. (as well as many others) have satisfactorily used just the first Wiener kernel, and hence only the first term of (1.16), to approximate $\tilde{Y}|X$. The benefits of encoding in this fashion is two-fold: computing the first Wiener kernel is inexpensive, and not much data is required to compute it. On the other hand, there are many instances (the cricket cercal sensory system for example [24, 25]) where this practical low order approximation, does not work well [60, 32]. Although it is theoretically possible to compute many terms in the Wiener series to improve the encoding approximation [42, 59], such computations can be quite costly, and they are rarely done in practice. The necessity of higher order terms in the approximation of $\tilde{Y}|X$ is another indication that neural encoding is not a linear process. To deal with this deficiency, van Hateren and Snippe use the Wiener filter in conjunction with various nonlinear models to estimate the response of the photoreceptor cells in the blowfly [81].

Another issue is that the Wiener/Volterra series is an expansion for a continuous function, which is appropriate for neural responses modelled as the firing rate $\tilde{Y}$. But how does one construct a Wiener/Volterra series to model the discrete spiking of neurons $Y$?

Furthermore, the result of calculating $\tilde{Y}$ using a Wiener series approximation gives a specific $\tilde{Y}(t)|X(\tau)$. Since we view encoding within a probabilistic framework, we wish to determine an approximation to $p(\tilde{Y}|X)$, the encoder. In principle, one could repeat realizations of the stimulus to estimate $p(\tilde{Y}|X)$. But now one is once again faced with fact that the space of possible stimuli for some systems is a very large space. Thus, it is not feasible to present all possible stimuli in experiment to estimate $p(\tilde{Y}|X)$.

Neural Decoding. We now turn our attention to the problem of estimating the neural decoder $p(X|Y)$. This problem may be more tractable than the task of determining the encoder $p(Y|X)$ since it is easier to estimate $p(X|Y)$ over an ensemble of responses, since $\mathcal{Y} := \{0,1\}^k$ is in many cases a much smaller space than the space of stimuli $\mathcal{X}$.

The Linear Reconstruction Method, espoused by Rieke et al in 1997 [59], considers a linear Wiener/Volterra approximation of $X|Y$

$$\begin{aligned} X(t) &= \int K_1(\tau)Y(t-\tau)d\tau \\ &= \sum_i K_1(t-t_i). \end{aligned} \tag{1.17}$$

The last equation follows if one models a spike train as a sum of delta functions

$$Y(t) = \sum_i \delta(t-t_i),$$

where the $i^{th}$ spike occurs at time $t_i$. To determine $K_1$, one minimizes the mean squared error [59]

$$\min_{K(t)} \left( \sum_{x\in\mathcal{X}} \int_{\Re} \left( x(t) - \sum_i K(t-t_i) \right)^2 dt \right),$$

which has the explicit solution [59]

$$K_1 = \mathcal{F}^{-1}\left( \frac{\langle \mathcal{F}(X(\omega)) \sum_j e^{-i\omega t_j} \rangle_Y}{\langle | \sum_j e^{-i\omega t_j} | \rangle_Y} \right). \tag{1.18}$$

Here, $\langle\cdot\rangle_Y$ indicates averaging over the values of $y \in \mathcal{Y}$, $\mathcal{F}$ indicates a Fourier Transform, and $\omega$ is frequency. The numerator of (1.18) is the Fourier transform of average

stimulus surrounding a spike, and the denominator is the power spectrum of the spike train.

This method deals with one of the problems from the Wiener/Volterra series method of encoding by modelling $Y(t)$ as a delta function, and so the temporal structure of spikes is considered. This does not violate the continuity assumption of the Wiener series as in the encoding regime because in decoding, we need only assume that $X(t)$ is a continuous function, not $Y(t)$.

Computing only one kernel (from (1.18)), which is computationally inexpensive, presupposes that decoding is linear. Furthermore, this method yields only a point estimate of $X|Y$. To estimate $p(X|Y)$, one would need to continue an experiment for a long period of time in the hope of producing many instances of the same neural response for each observed $y \in \mathcal{Y}$. Unfortunately, as pointed out in [37], the amount of data needed to support non-parametric estimates of coding schemes which contain long sequences of length $T$ across $N$ neurons grows exponentially with $T$ and $N$. For some systems, the required data recording time may well exceed the expected lifespan of the system.

The linear reconstruction method models a single neuron, and it is not clear how the regime can be extended to account for populations of neurons. Although there is evidence that neural coding is performed independently by single neurons [49], coding by a population of neurons has been shown to be important in some sensory systems [55, 77], as well as from a theoretical point of view [45, 77]. Other linear methods have been developed which do model populations of neurons, but, unfortunately, for each of the ones that we introduce here, the neural response is assumed to be spike counts in a time window, $\tilde{Y}$. Georgopoulos et al. in 1983 proposed the Population Vector Method [30] which decodes a stimulus using a convolution similar to (1.17) to estimate $X|\tilde{Y}$

$$X(t) = \sum_i \tilde{Y}_i C_i.$$

Here, $C_i$ is the *preferred stimulus* for neuron $i$. Abbot and Salinas in 1994 [63] proposed their Optimal Linear Estimator (OLE), which decodes by

$$X(t) = \sum_i \tilde{Y}_i D_i$$

where $D_i$ is chosen so that

$$\langle\langle \int_{\Re} \left( x(t) - \sum_i \tilde{Y}_i D_i \right)^2 dt \rangle_{\tilde{Y}} \rangle_X,$$

the mean squared error averaged over all stimuli and all neural responses observed in experiment, is minimized. As in (1.18), $\langle\cdot\rangle_X$ and $\langle\cdot\rangle_{\tilde{Y}}$ indicate averaging over the spaces $\mathcal{X}$ and $\tilde{\mathcal{Y}}$ respectively. The analytic solution for such a $D_i$ is given by [63]

$$D_i = \sum_j Q_{ij}^{-1} L_j$$

where $L_j$ is center of mass of the tuning curve for cell $i$ (see Figure 7B), and $Q_{ij}$ is the correlation matrix of $\tilde{Y}_i$ and $\tilde{Y}_j$.

There are other linear methods for decoding as well, which use either a Maximum Likelihood Estimator or a Bayesian estimator instead of the OLE [63].

To get a good sampling of points $\tilde{y} \in \tilde{\mathcal{Y}}$, Abbot and Salinas advocate presenting a randomly chosen, continuously varying stimulus $X$, such as a Gaussian White Noise (GWN) stimulus, to the sensory system. This enables an experimenter to take a "random walk" through the stimulus space, thereby eliciting a wide range of neural responses from $\mathcal{Y}$ [63, 47, 74].

The Population Vector Method is inexpensive to implement, and is ideal when the tuning curve is a (half) cosine as in the case of the cricket cercal sensory system (Figure 7B). Furthermore, small error (difference of the estimated stimulus from the true stimulus) is incurred when decoding $\{\tilde{Y}_i\}$ if the preferred stimuli $\{C_i\}$ are orthogonal. The OLE in fact has smallest average mean squared error of all linear methods over a population of neurons [63]. For the Population Vector Method, however, it is not always obvious what the preferred stimulus $C_i$ is for generic, complex stimuli. Furthermore, the method does not work well if the preferred stimuli $\{C_i\}$ are not uniformly distributed, and it requires a lot of neurons in practice [63]. Neither of these linear methods give an explicit estimate of $p(X|Y)$.

A parametric approach, in which a particular probability distribution is assumed, could yield an explicit form of $p(X|Y)$ as is the case when one considers Poisson encoding models. Such a model for decoding was proposed by de Ruyter van Steveninck and Bialek in 1988 [59]. In experiment, they let $X(t)$ be a randomly chosen and continuously varying stimulus. $p(X|Y)$ is then approximated with a Gaussian with mean $E(X|Y)$ and covariance $\mathrm{Cov}(X|Y)$ computed from data as in Figure 10.

In this regime, the temporal pattern of the spikes is considered and one has an explicit form for $p(X|Y)$. But why should $p(X|Y)$ be Gaussian? This choice is justified by the following remark.

REMARK 3. *Jayne's maximum entropy principle [36] states that of all models that satisfy a given set of constraints, one ought to choose the one that maximizes the entropy, since a maximum entropy model does not implicitly introduce additional constraints in the problem. Rieke et al. show that over all models with a fixed mean and covariance, the Gaussian is the maximum entropy model [59].*

However, an inordinate amount of data is required to obtain good estimates of $\mathrm{Cov}(X|Y = y)$ over all observed $y \in \mathcal{Y}$, which requires one to continue an experiment for a long period of time. Another way to deal with the problem of not having enough data is to cluster the responses together and then to estimate a gaussian model for each response cluster.

The last approach we study here is the Metric Space Approach of Victor and Purpura (1996) [84, 83], which actually constructs an estimate of the joint probability $p(X, Y)$. From the previous decoders we have examined, we see that we are in search

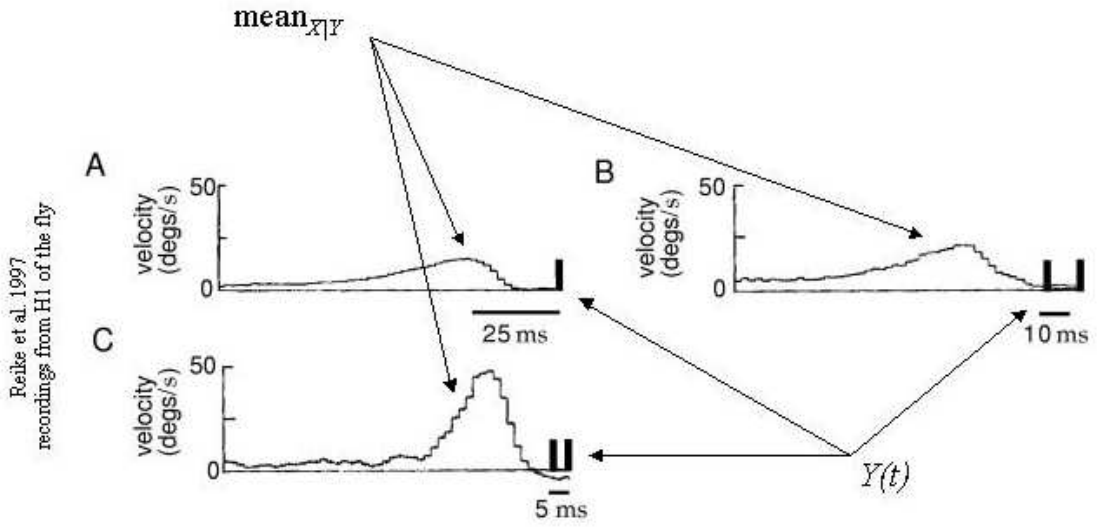



Figure 10. Estimating $p(X|Y)$ with a Gaussian. Examples of three spike trains recorded from the H1 neuron of the blowfly and the corresponding conditional means of the stimuli (velocity of a pattern) which elicited each of these responses. These conditional means, as well as conditional variances, are used to construct a Gaussian decoder $p(X|Y)$ of the stimuli [59].

of a decoding method that estimates $p(X|Y)$, takes the temporal structure of the spikes of the neural responses $Y(t)$ into account, and deals with the insufficient data problem. The Metric Space Approach satisfies all these goals, and without assuming a distribution on $X|Y$ a priori, as was necessary for the Poisson and Gaussian models we have examined. Instead, as the name implies, a metric is assumed on $\mathcal{Y}$. Choosing some scalar $r \geq 0$ and given two spike trains, $Y_i$ and $Y_j$ , the distance between them is defined by the metric

$$D[r](Y_i, Y_j), \tag{1.19}$$

which is the minimum cost required to transform $Y_i$ into $Y_j$ via a path of elementary steps (see Figure 11):

1. Adding or deleting a spike has a cost of 1.

2. Shifting a spike in time by $\Delta t$ has a cost of $r|\Delta t|$.

The quantity $\frac{1}{r}$ can be interpreted as a measure of the temporal precision of the metric. The metric

$$D[r = 0](Y_i, Y_j)$$

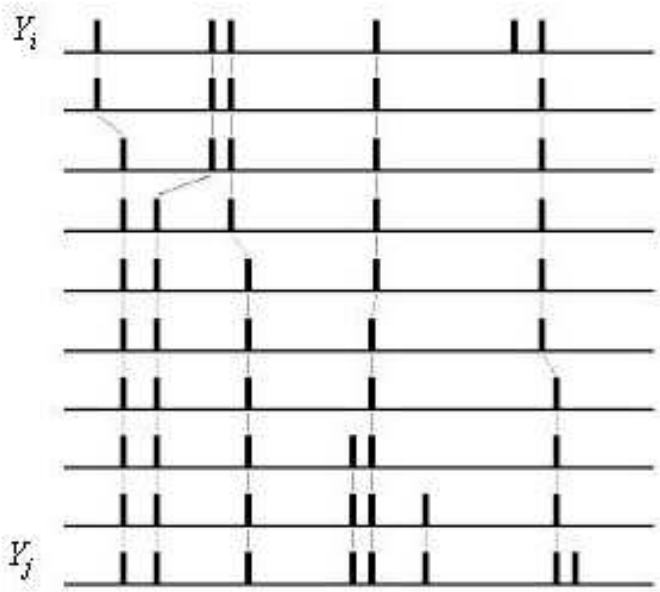



Figure 11. Computing the Spike Train Metric [84]. One path of elementary steps used to transform a spike train $Y_i$ into a spike train $Y_j$..

is just the difference in the number of spikes between the spike trains $Y_i$ and $Y_j$. Coding based on this measure is just counting spikes since no cost is incurred when shifting spikes in time. The metric

$$D[r = \infty](Y_i, Y_j)$$

gives infinitesimally precise timing of the spikes.

Unfortunately, the Metric Space Approach suffers from the same problem that all of the encoders that we have investigated do: the stimuli, $x_1, x_2, ..., x_C$ must be repeated multiple times, a problem when $\mathcal{X}$ is large. The Metric Space Approach is described by the following Algorithm.

Algorithm 4 (Metric Space Method). *[84] Choose $r \geq 0$ and an integer $z$. Suppose that there are $C$ stimuli, $x_1, x_2, ..., x_C$, presented multiple times each, all of which elicit a total of $K$ neural responses $y_1, y_2, ..., y_K$. Initialize $\mathcal{C}$, the $C \times C$ classification matrix, to zeros, and let $\nu_1$, $\nu_2$, ..., $\nu_C$ be $C$ abstract response classes. Start the algorithm with $i = 1$.*

1. *Suppose that $y_i$ was elicited by $x_\alpha$. Assign $y_i$ to response class $\nu_\beta$ if*

$$\langle D[r](y_i, \hat{y})^z \rangle_{\hat{y} \text{ elicitedby } x_\beta}^{\frac{1}{z}}$$

*is the minimum over all $x_k$ for $k = 1, ..., C$.*

A

| $\nu_1$ | $\nu_2$ | $\nu_3$ | $\nu_4$ | $\nu_5$ | |
|---|---|---|---|---|---|
| 3 | 11 | 3 | 2 | 1 | $x_1$ |
| 5 | 10 | 3 | 2 | 0 | $x_2$ |
| 1 | 1 | 15 | 1 | 2 | $x_3$ |
| 1 | 0 | 4 | 2 | 13 | $x_4$ |
| 2 | 3 | 2 | 5 | 8 | $x_5$ |

B

| $\nu_1$ | $\nu_2$ | $\nu_3$ | $\nu_4$ | $\nu_5$ | |
|---|---|---|---|---|---|
| .25 | .44 | .11 | .17 | .043 | $x_1$ |
| .42 | .40 | .11 | .17 | 0 | $x_2$ |
| .08 | .04 | .56 | .08 | .08 | $x_3$ |
| .08 | 0 | .15 | .17 | .54 | $x_4$ |
| .17 | .12 | .07 | .42 | .33 | $x_5$ |

Table 1. A: An example of the Metric Space method for clustering data where $K = 100$ neural responses were clustered into $C = 5$ classes. Observe that there were 20 neural responses elicited by each $C = 5$ stimulus. B: The $i^{th}$ column of the normalized matrix $\mathcal{C}$ gives the decoder $p(X|\nu_i)$. In this example, any of the neural responses which belong to $\nu_1$ are decoded as the stimulus $x_2$ with certainty .42. Any of the neural responses in class $\nu_3$ are decoded as the stimulus $x_3$ with certainty .56.

*2. Increment the component* $[\mathcal{C}]_{\alpha\beta}$ *of the matrix* $\mathcal{C}$ *by 1.*

*3. Repeat step 1 and 2 for* $i = 2, ..., K$

One normalizes the columns of the matrix $\mathcal{C}$ to get the decoder $p(X|\nu)$ (see Table 1). Decode a neural response $y$ and the certainty of the assignment $p(X|y)$ by looking up its response class $\nu$ in the normalized matrix $\mathcal{C}$ (see Table 1B). The responses are clustered together to obtain $p(X|\nu)$, an estimate of $p(X|Y)$ given the available amount of data.

Minimizing the cost function $D[r]$ in step 1 of Algorithm 4 is intuitively a nice way to quantify jitter in the spike trains. As we have seen, in Rate Distortion Theory, this type of cost function is called a distortion function. The values for $q$ and $z$ that Victor and Purpura recommend to use in Algorithm 4 are those that maximize the transmitted information from stimulus to response [84].

## The Information Distortion

The brief survey in the last section gives insight into what types of characteristics that an encoding/decoding algorithm ought to have. First, the algorithm ought to produce an estimate of $X|Y$ (or of $Y|X$) as well as a measure of the certainty of the estimate, $p(X|Y)$ (or $p(Y|X)$). The temporal structure of the spike trains of the neural responses need to be considered. Assumptions about the linearity of encoding or decoding ought not to be required. Presentation of all stimuli must not be required. Rather, $X(t)$ ought to be randomly chosen and continuously varying. A population of neurons ought to be able to be considered. And lastly, the algorithm needs to deal

with the problem of having limited data, perhaps by clustering the neural responses. The Information Distortion method [22, 20, 29] satisfies these prerequisites.

It searches for approximations of the decoder $p(X|Y)$ by *quantizing* the neural responses $\mathcal{Y}$ to a small *reproduction* set of $N$ classes, $\mathcal{Y}_N$, by defining the random variable

$$Y_N : \Omega_Y \to \mathcal{Y}_N.$$

The random variables

$$X \to Y \to Y_N$$

form a Markov chain [22]. The *quantization* or stochastic assignment [17, 35] of the elements of $\mathcal{Y}$ to $\mathcal{Y}_N$ is defined by the *quantizer* $q(Y_N|Y)$

$$q(Y_N|Y) : \mathcal{Y} \to \mathcal{Y}_N. \tag{1.20}$$

The Information Distortion method computes an optimal quantizer $q^*(Y_N|Y)$ that minimizes an information-based distortion function, called the *information distortion measure*,

$$D_I(Y, Y_N),$$

which is defined in (2.11). Applying the information distortion measure to neural data, which is equivalent to maximizing the *information transmission* between the stimulus space and quantized neural responses, has theoretical justification [9, 20, 22, 37, 51, 59, 64, 72, 83, 84]. Such a $q^*(Y_N|Y)$ for a fixed $N$ produces the Gaussian distribution $p(X|Y_N)$, which is an approximation to the decoder $p(X|Y)$ (see (2.26)). Recall that the choice of a Gaussian is justified by Remark 3. These approximations $p(X|Y_N)$ can be refined by increasing $N$, which increases the size of the reproduction $\mathcal{Y}_N$. There is a critical size, $N_{\max}$, beyond which further refinements do not significantly decrease the distortion $D_I(Y, Y_{N_{\max}})$ given the amount of data. Thus, given sufficient data, one chooses the optimal quantization $q^*(Y_{N_{\max}}|Y)$ at this size $N_{\max}$, which in turn gives the Gaussian $p(X|Y_{N_{\max}})$, an estimate of the decoder $p(X|Y)$.

## Outline of Thesis

The goal of this thesis is to solve problems of the form (1.9),

$$\max_{q \in \Delta}(G(q) + \beta D(q)),$$

where Assumption 2 is satisfied, and $q$ is a clustering or quantization of some objects $Y$ to some objects $Y_N$. To motivate why we are interested in the problem, we require the language of information theory. To study solution behavior of the problem, we need ideas from optimization theory, bifurcation theory, and group theory. The purpose of this section is to further elucidate the details of how the chapters that follow present these ideas.

In chapter 2, we introduce the notation and develop the mathematical tools that will be used throughout the rest of this thesis. The tools we develop here include the rudiments of Information Theory, a formal introduction to instances of the functions $D(q)$ and $G(q)$ which compose the terms of (1.9), and finally a formal exposition of the information distortion measure which we introduced earlier in this chapter. The latter objective is necessary since optimizing this measure is a key ingredient to both the Information Distortion [22, 20, 29] and the Information Bottleneck [70, 78, 69] methods, our two main problems of interest.

In chapter 3, we use tools from constrained optimization theory to rewrite (1.9) in terms of its Lagrangian

$$\mathcal{L}(q, \lambda, \beta) : \Re^{NK} \times \Re^{K} \times \Re \to \Re. \tag{1.21}$$

Later, in chapter 9, we examine optimization schemes, such as the implicit solution [22, 29] and projected Augmented Lagrangian [29, 50] methods, which exploit the structure of (1.21) to find local solutions to (1.9) for step 3 of algorithm 1.

We wish to pose (1.9) as a dynamical system in order to study the *bifurcation structure* of these local solutions for $\beta \in [0, \mathcal{B}]$. To this end, we consider the equilibria of the flow

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda} \mathcal{L}(q, \lambda, \beta) \tag{1.22}$$

for $\beta \in [0, \mathcal{B}]$ and some $\mathcal{B} < 0$. These are points $\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix}$ where $\nabla_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta) = 0$ for some $\beta$. The Jacobian of this system is the Hessian $\Delta_{q,\lambda} \mathcal{L}(q, \lambda, \beta)$. Equilibria, $(q^*, \lambda^*)$, of (1.22), for which $\Delta F(q^*, \beta)$ is negative definite on the kernel of the Jacobian of the constraints, are local solutions of (1.9) (Remark 27).

In chapter 4 we explore the pivotal role that the kernel of $\Delta_{q,\lambda} \mathcal{L}$ plays determining the bifurcation structure of solutions to (1.9). This is due to the fact that bifurcation of a branch of equilibria $(q^*, \lambda^*, \beta)$ of (1.22) at $\beta = \beta^*$ happens when $\ker \Delta_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta^*)$ is nontrivial (Theorem 24). Furthermore, the bifurcating branches are tangent to certain linear subspaces of $\ker \Delta_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta^*)$ (Theorem 110). More surprisingly perhaps is that the block diagonal Hessian $\Delta F$ (Assumption 2.3) plays a crucial role as well. We will derive explicit relationships between these Hessians in this chapter, and we will show that, generically, there are only three types of singularities of $\Delta_{q,\lambda} \mathcal{L}$ and $\Delta F$ which can occur. Furthermore, we explain how these singularities dictate the bifurcation structure of equilibria of (1.22) (Figure 12). In particular, the singularity types show that, generically, only two different types of bifurcation can occur: symmetry breaking bifurcation and saddle-node bifurcation.

In chapter 5, we present the general theory of bifurcations in the presence of symmetries, which includes the Equivariant Branching Lemma (Theorem 47) and the Smoller-Wasserman Theorem (Theorem 49). We are able to extend some of the results of Golubitsky [33, 34] to determine the bifurcation structure of pitchfork-like bifurcations for equilibria of a general dynamical system with symmetries.

In chapter 6 we apply the general theory of bifurcations in the presence of symmetries to the dynamical system (1.22). When an equilibrium $(q^*, \lambda^*, \beta^*)$, which is fixed by the action of the group $S_M$, undergoes bifurcation, then the Equivariant Branching Lemma ascertains the existence of explicit bifurcating solutions in one dimensional subspaces of $\ker \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*)$ which are fixed by special subgroups of $S_M$ (Theorem 110). Such symmetry breaking bifurcations are always pitchfork-like (Theorem 120). Further information about the bifurcation structure of solutions to (1.9) can be garnered using the symmetry of $F$. In the sequel, we show that every singularity of $\Delta_{q,\lambda}\mathcal{L}$ yields bifurcating branches when $G$ is strictly concave (Corollary 108), which is the case for the Information Distortion problem (1.4). We also provide conditions which determine the location (Theorem 80), type (Theorem 120), orientation (i.e. supercritical or subcritical), and stability (Theorems 127 and 128) of bifurcating branches from certain solutions to (1.9). In some instances, unstable branches can not contain solutions to (1.9) (Theorem 129).

In chapter 7, we introduce continuation techniques which allow us to confirm the theory of chapter 6 by numerically computing the bifurcation structure of stationary points of the Information Distortion problem (2.34). There are two types of bifurcations which we observe numerically: symmetry breaking bifurcations and saddle-node bifurcations. See Figures 16–24 and 25.

In chapter 8 we show that bifurcations that are not symmetry breaking bifurcations are generically saddle-node bifurcations. We also give necessary and sufficient conditions for the existence of saddle-node bifurcations (Theorems 135 and 141).

In chapter 9, we introduce two numerical optimization schemes [40, 50] which can be used in step 3 of the annealing algorithm (Algorithm 1) to find *solutions* of the problem (1.9): the Augmented Lagrangian Method (Algorithm 149) and an implicit solution method (9.20). Another optimization scheme, which does not use the method of annealing, can be used to solve (1.9) when $D(q)$ is convex and $\mathcal{B} = \infty$, as is the case for the Information Distortion method. This vertex search algorithm is a greedy search over the vertices of $\Delta$ (Algorithm 155). Each of these algorithms has its advantages and disadvantages, and we rate their performance on synthetic and physiological data sets (Tables 4–5 and Figure 27).

One of the purposes of this thesis is to introduce methodology to improve Algorithm 1 and to minimize the arbitrariness of the choice of the algorithm's parameters. Thus, we conclude with an algorithm (Algorithm 157) which shows how continuation and bifurcation theory in the presence of symmetries can be used to aid in the implementation of Algorithm 1.

# CHAPTER 2

## MATHEMATICAL PRELIMINARIES

In this chapter we introduce the notation and develop the mathematical tools that will be used throughout the rest of this thesis as we study solutions of (1.9),

$$\max_{q\in\Delta}(G(q)+\beta D(q)),$$

where $q$ is a clustering or quantization of some objects $Y$ to some objects $Y_N$. To motivate why we are interested in problems of this form, we present the rudiments of Information Theory, introduce the functions $D(q)$ and $G(q)$ which compose the terms of (1.9), and give a formal exposition of the information distortion measure which we introduced in chapter 1. The latter objective is necessary since optimizing this measure is a key ingredient to both the Information Distortion [22, 20, 29] and the Information Bottleneck [70, 78, 69] methods, our two main problems of interest.

### Notation and Definitions

The following notation will be used throughout the sequel:

$|H|$ := the number of elements of the set $H$, differentiated from "the absolute value of" when the argument is a set.

$Y$ := a random variable with realizations from a finite set $\mathcal{Y} := \{y_1, y_2, ..., y_K\}$.

$K := |\mathcal{Y}| < \infty$, the number of elements of $\mathcal{Y}$, the realizations of the random variable $Y$.

$Y_N$ := a random variable with realizations from the *set of classes* $\mathcal{Y}_N := \{1, 2, ..., N\}$.

$N := |\mathcal{Y}_N|$, the total number of classes.

$p(X)$ := the probability mass function of $X$ if $X$ is a discrete random variable. If $X$ is a continuous random variable, then $p(X)$ is the probability density function of $X$.

$q(Y_N|Y)$ := the $K \times N$ matrix, $p(Y_N|Y)$, defining the conditional probability mass function of the random variable $Y_N|Y$, written explicitly as

$$\begin{pmatrix} q(1|y_1) & q(1|y_2) & q(1|y_3) & ... & q(1|y_K) \\ q(2|y_1) & q(2|y_2) & q(2|y_3) & ... & q(2|y_K) \\ \vdots & \vdots & \vdots & & \vdots \\ q(N|y_1) & q(N|y_2) & q(N|y_3) & ... & q(N|y_K) \end{pmatrix} = \begin{pmatrix} q(1|Y)^T \\ q(2|Y)^T \\ \vdots \\ q(N|Y)^T \end{pmatrix}.$$

$q^\nu$ := $q(\nu|Y)$, the transpose of the $1 \times K$ row of $q(Y_N|Y)$ corresponding to the class $\nu \in Y_N$.

$q$ := the vectorized form of $q(Y_N|Y)^T$, written as

$$q = ((q^1)^T \ \ (q^2)^T \ \ ... \ \ (q^N)^T)^T.$$

$q_{\nu k}$ := $q(Y_N = \nu | Y = y_k)$, the component of $q$ corresponding to the class $\nu \in Y_N$ and the element $y_k \in Y$.

$\delta_{a_1 \ldots a_m}$ := a scalar function on the natural numbers $\{a_i\}_{i=1}^m$ with range $\begin{cases} 1 \text{ if } a_i = a_j \ \ \forall i,j \\ 0 \text{ otherwise} \end{cases}$

$\log \boldsymbol{x}$ :=$\log_2 \boldsymbol{x}$, the component-wise log base 2 operator of the vector $\boldsymbol{x}$.

$\ln \boldsymbol{x}$ :=$\log_e \boldsymbol{x}$, the component-wise natural log operator of the vector $\boldsymbol{x}$.

$[\boldsymbol{x}]_i$ := $i^{th}$ component of the vector $\boldsymbol{x}$

$[A]_{ij}$ := the $(i,j)^{th}$ component of the matrix $A$

$A^-$ := the Moore-Penrose generalized inverse of the $k \times m$ matrix $A$.

$\det A$ := the determinant of the matrix $A$.

**peigenspace(A)** := the vector space spanned by the eigenvectors corresponding to the positive eigenvalues of the square matrix $A$.

$A \otimes B$ := the Kronecker product of the $p \times q$ matrix $A$ and the $r \times s$ matrix $B$ is defined as the $pr \times qs$ matrix $C$, such that the $(i,j)^{th}$ block of $C$ is $[C]_{ij} = A \otimes B = a_{ij}B$.

$< \boldsymbol{v}, \boldsymbol{w} >_A$ := $\boldsymbol{v}^T A \boldsymbol{w} = \sum_{i,j} [\boldsymbol{v}]_i A_{ij} [\boldsymbol{w}]_j$, an inner product with respect to $A$ if $A$ is positive definite.

$< \boldsymbol{v}, \boldsymbol{w} >$ := $< \boldsymbol{v}, \boldsymbol{w} >_I = \sum_{i,j} [\boldsymbol{v}]_i [\boldsymbol{w}]_j$, the Euclidean inner product.

$||\boldsymbol{v}||$ := $\sqrt{< \boldsymbol{v}, \boldsymbol{v} >}$, the Euclidean norm.

$\angle(\boldsymbol{v}, \boldsymbol{w})$ := the angle between the vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, measured in radians.

$I_k$ := the $k \times k$ identity matrix.

$\boldsymbol{e}_i$ :=$i^{th}$ column of the identity $I$.

$E_X f(X)$ := $\int_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$, the expected value of scalar function $f(X)$ with respect to the distribution $p(X)$.

$\psi(\boldsymbol{x})|_\Omega$ := the vector valued function $\psi$ restricted to the space $\Omega$.

$\partial_{\boldsymbol{x}}\psi$ := Jacobian of the vector valued function $\psi$ with respect to the vector $\boldsymbol{x}$

$\partial^2_{\boldsymbol{x}}\psi$ := three dimensional array of second derivatives of the vector valued function $\psi$ with respect to the vector $\boldsymbol{x}$

$\partial^2_{\boldsymbol{x}}\psi(\boldsymbol{x}_0)[\boldsymbol{v},\boldsymbol{w}]$ := the vector defined by the multilinear form $\sum_{i,j} \frac{\partial^2\psi}{\partial[\boldsymbol{x}]_i\partial[\boldsymbol{x}]_j}(\boldsymbol{x}_0)[\boldsymbol{v}]_i[\boldsymbol{w}]_j$, where $\psi(\boldsymbol{x})$ is a vector valued function.

$\partial^3_{\boldsymbol{x}}\psi(\boldsymbol{x}_0)[\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}]$ := the vector defined by the multilinear form

$$\sum_{i,j,k} \frac{\partial^2\psi}{\partial[\boldsymbol{x}]_i\partial[\boldsymbol{x}]_j\partial[\boldsymbol{x}]_k}(\boldsymbol{x}_0)[\boldsymbol{u}]_i[\boldsymbol{v}]_j[\boldsymbol{w}]_k,$$

where $\psi(\boldsymbol{x})$ is a vector valued function.

$\nabla_{\boldsymbol{x}} f$ := gradient of the scalar function $f$ with respect to the vector $\boldsymbol{x}$.

$\nabla f(\boldsymbol{x},\beta)$ := $\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\beta)$.

$\Delta_{\boldsymbol{x}} f$ := Hessian of the scalar function $f$ with respect to the vector $\boldsymbol{x}$.

$\Delta f(\boldsymbol{x},\beta)$ := $\Delta_{\boldsymbol{x}} f(\boldsymbol{x},\beta)$.

$$\textbf{sgn } f(x) := \begin{cases} 1 & \text{if } f(x) > 0 \\ 0 & \text{if } f(x) = 0 \\ -1 & \text{if } f(x) < 0 \end{cases}$$

$\mathcal{O}(\boldsymbol{x}^m)$ := "big oh" of $||\boldsymbol{x}||^m$. By definition, if $f(\boldsymbol{x}) = \mathcal{O}(\boldsymbol{x}^m)$, then there exists $n > 0$ such that $\|f(\boldsymbol{x})\| \leq n\|\boldsymbol{x}\|^m$ if $\|\boldsymbol{x}\|$ is sufficiently small.

$\leq$ := is a subgroup of, differentiated from "is less than or equal to" when the arguments being compared are sets.

$<$ := is a proper subgroup of, differentiated from "is strictly less than" when the arguments being compared are sets

$[G : H]$ := $\frac{|G|}{|H|}$, the index of $H$ in $G$, when $H \leq G$ and $|G| < \infty$.

$\cong$ := is isomorphic as a group to

$\langle g \rangle$ := the cyclic group generated by $g$, where $g$ is an element of some group $G$

$|g|$ := the order of the element $g$ in the group $G$, which is equivalent to $|\langle g \rangle|$.

$S_M$ := the abstract group of $M!$ elements of all permutations on $M$ objects.

An $n \times n$ symmetric matrix $A$ is *positive definite* if $\boldsymbol{x}^T A\boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \Re^n$ and is *negative definite* if $\boldsymbol{x}^T A\boldsymbol{x} < 0$ for all $\boldsymbol{x} \in \Re^n$. The symmetric matrix $A$ is *non-positive definite* if $\boldsymbol{x}^T A\boldsymbol{x} \leq 0$ for all $\boldsymbol{x} \in \Re^n$ and is *non-negative definite* if $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \Re^n$.

A square matrix $A$ has a *singularity*, or is *singular*, if at least one of its eigenvalues is zero. The space spanned by the eigenvectors corresponding to the zero eigenvalues of $A$ is called the *kernel* or *nullspace* of $A$, denoted by $\ker A$. Thus, $A$ is singular if and only if $\ker A \neq \emptyset$ if and only if $\det A = 0$.

A vector space $B$ is called a normed vector space if there a norm defined on the elements of $B$. The vector space $B$ is said to be *complete* if every Cauchy sequence converges to a point in $B$. A complete normed vector space is a called a *Banach* space. A vector space $B$ is called an *inner product space* if there is an inner product (or dot product) defined on the elements of $B$. A complete normed inner product space is called a *Hilbert* space.

A stationary point $\boldsymbol{x}^*$ of a differentiable function $f(\boldsymbol{x})$ is a point where

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*) = \mathbf{0}.$$

A *Lie group* is any continuous group. In this thesis, if $G$ is a Lie group, then we use the matrix representation of $G$, which has the form

$$G = \{g \in \Re^m \times \Re^m | g \text{ is invertible}\},$$

together with the binary operation of matrix multiplication.

## Information Theory

The basic object in information theory is an *information source* or a random variable (measurable function)

$$X : (\Omega, \mathcal{O}) \rightarrow (\mathcal{X}, \mathcal{B}), \tag{2.1}$$

where $\mathcal{X}$ is the probability space of symbols produced by $X$, a representation of the elements of the probability space $\Omega$. $\mathcal{O}$ and $\mathcal{B}$ are the respective $\sigma$-algebras. A source $X$ is a mathematical model for a physical system that produces a succession of symbols $\{X_1, X_2, \ldots, X_n\}$ in a manner which is unknown to us and is treated as random [17, 35]. The sequence $\{X_i\}_{i=1}^n$ is said to be *i.i.d* or *identically and independently distributed* if $X_i$ are mutually independent

$$p(X_i, X_j) = p(X_i)p(X_j)$$

for $i \neq j$, and if the probability density of $X_i$, is the same for every $i$ and $j$,

$$p(X_i) = p(X_j).$$

The sequence $\{X_i\}$ is *stationary* if for each $m$ and $k$, $(X_0, ..., X_m)$ and $(X_k, ..., X_{k+m})$ have the same probability density. In other words, $\{X_i\}$ is stationary if no matter when one starts observing the sequence of random variables, the resulting observation has the same probabilistic structure.

A measurable transformation $\varphi : \Omega \to \Omega$ is *measure preserving* if $p(\varphi^{-1}A) = p(A)$ for all $A \in \mathcal{O}$. A set $A \in \mathcal{O}$ is $\varphi$-*invariant* if $\varphi^{-1}A = A$. Let $\mathcal{I} = \{A | A \text{ is } \varphi\text{-invariant}\}$. The measurable transformation $\varphi$ is *ergodic* if for every $A \in \mathcal{I}, p(A) \in \{0, 1\}$. The source $X_i = X \circ \varphi^i$ is said to be ergodic if $\varphi$ is ergodic.

An *information channel* is a pair of information sources $(X, Y)$, an input

$$X : (\Omega_X, \mathcal{O}_X) \to (\mathcal{X}, \mathcal{B}_X), \tag{2.2}$$

and an output

$$Y : (\Omega_Y, \mathcal{O}_Y) \to (\mathcal{Y}, \mathcal{B}_Y) \tag{2.3}$$

where the spaces and $\sigma$-algebras are defined as in (2.1).

The basic concepts of information theory are *entropy* and *mutual information* [17]. In information theory, entropy is described as a measure of the uncertainty, or of the self information, of a source, and is defined as

$$H(X) = -E_X \log p(X).$$

The *conditional* and *joint* entropy respectively given an information channel $(X, Y)$ are defined respectively as

$$\begin{aligned} H(Y|X) &= -E_{X,Y} \log p(Y|X) \\ H(X, Y) &= -E_{X,Y} \log p(X, Y). \end{aligned}$$

It is easy to show that

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

The notion of *mutual information* $I(X; Y)$ is introduced as a measure of the degree of dependence between a pair of sources in an information channel $(X, Y)$:

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \tag{2.4}$$

$$= E_{X,Y} \log \frac{p(X, Y)}{p(X)p(Y)} \tag{2.5}$$

Both entropy and mutual information are special cases of a more general quantity – the *Kullback-Leibler directed divergence* or *relative entropy* [43] between two probability measures, $p$ and $r$, on the same discrete probability space $\mathcal{X}$,

$$KL(p\|r) = E_X \log \left( \frac{p(X)}{r(X)} \right). \tag{2.6}$$

The Kullback-Leibler divergence is always nonnegative and it is zero if and only if $p(X) = r(X)$ almost everywhere. However, it is not symmetric and so it is not a proper distance on a set of probability measures. In spite of this it provides a sense of how different two probability measures are.

The information quantities $H$, $I$ and $KL$ depend only on the underlying probability distributions and not on the structure of $X$ and $Y$. This allows us to evaluate them in cases where more traditional statistical measures (e.g. variance, correlation, etc.) do not exist.

Why are entropy and mutual information valid measures to use when analyzing an information channel between $X$ and $Y$? Let $\{Y_1, Y_2, \ldots, Y_n\}$ be i.i.d. observations from an information source $Y$. Then the Strong Law of Large Numbers provides theoretical justification for making inference about population parameters (such as the mean and variance) from data collected experimentally [28]. In particular, the Shannon Entropy Theorem [17, 28, 68] in this case assures that the entropy (and hence the mutual information) calculated from data taken experimentally converges to the true population entropy as the amount of data available increases.

THEOREM 5 (SHANNON ENTROPY THEOREM). *([68]) If $\{Y_i\}$ are i.i.d. then*

$$\lim_{n\to\infty} -\frac{1}{n}\log p(Y_1, Y_2, ..., Y_n) = H(Y) \text{ a.s.}$$

*Proof.* The random variables $\{\log p(Y_i)\}_{i=1}^n$ are i.i.d. and so by the Strong Law of Large Numbers

$$\begin{aligned} E(\log(p(Y)) &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log p(Y_i) \\ &= \lim_{n\to\infty} \frac{1}{n}\log \Pi_{i=1}^{n} p(Y_i) \\ &= \lim_{n\to\infty} \frac{1}{n} p(Y_1, Y_2, ..., Y_n) \end{aligned}$$

almost surely. □

In many instances, as in the case of physiological recordings from a biological sensory system, the data $\{Y_1, Y_2, \ldots, Y_n\}$ are not i.i.d.. For example, in the data presented in this thesis, a single, "long" recording of a neural response is partitioned into observations of length, say, 10 ms. Inference made about population parameters from data collected this way is justified if we can assume that $Y$ is stationary ergodic. Now we may appeal to the Ergodic Theorem [10, 28] and the Shannon-McMillan-Breiman Theorem [17, 28] to justify the use of information theoretic quantities.

THEOREM 6 (ERGODIC THEOREM). *(Birkhoff, 1931, p. 113-5 [10], p. 341-3 [28]) If $\varphi$ is a measure preserving transformation on $(\Omega, \mathcal{O})$ and $Y$ is a source with $E(Y) < \infty$.*

*Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} Y(\varphi^i \omega) = E(Y|\mathcal{I}) \text{ a.s.}$$

REMARK 7. *If $\varphi$ is ergodic, then $E(Y|\mathcal{I}) = E(Y)$. The Ergodic Theorem in this instance can be interpreted as a Strong Law of Large Numbers for ergodic processes.*

THEOREM 8 (SHANNON-MCMILLAN-BREIMAN THEOREM). *([17] p.474-479 , [28] p.356-360) If $Y_n$ for an integer $n$ is an ergodic stationary sequence taking values in a finite set $\mathcal{Y}$, then*

$$\lim_{n\to\infty} -\frac{1}{n} \log p(Y_0, Y_1, ..., Y_{n-1}) = H$$

*where $H \equiv \lim_{n\to\infty} E(-\log p(Y_n|Y_{n-1}, ..., Y_0))$ is the* entropy rate *of $\{Y_i\}$.*

REMARK 9. *Theorem 5 is a special case of Theorem 8 when $\{Y_i\}$ are i.i.d..*

Instead of considering the full space $\mathcal{Y}$ of all of the symbols elicited by $Y$, Theorem 8 gives justification for considering only a subset of $\mathcal{Y}$ which one "typically observes." This set is defined rigourously in the following way. Each element of the output space $\mathcal{Y}$ can be modelled as a sequence of symbols of a random variable

$$Z : (\Omega_Z, \mathcal{O}_Z) \to (\mathcal{Z}, \mathcal{B}_Z)$$

where $\Omega_Z$ and $\mathcal{B}_Z$ are defined as in (2.1). Hence $Y = Z^k$, the *k-th* extension of $Z$, can be thought of as the set of all sequences of length $k$ of symbols from $Z \in \mathcal{Z}$. There is a limited number of distinct messages which can be transmitted with sequences of length $k$ from the source $Z$. These are the typical sequences of $Z$ [17].

DEFINITION 10. *The* typical set $A_\epsilon^k$ *with respect to probability density $p(Z)$ on $Z$ is the set of sequences $(z_1, z_2, \ldots, z_k) \in Z^k$ for which*

$$2^{-k(H(Z)+\epsilon)} \leq p(z_1, z_2, \ldots, z_k) \leq 2^{-k(H(Z)-\epsilon)}.$$

*$(z_1, z_2, ..., z_n) \in A_\epsilon^k$ is called a* typical sequence.

A reformulation of Theorem 8 shows that the typical set has the following properties:

THEOREM 11 (ASYMPTOTIC EQUIPARTITION PROPERTY). *(p. 360 [28], p. 51 [17]) If $Z$ is stationary ergodic, then*

1. *If $(z_1, z_2, ..., z_k) \in A_\epsilon^k$ then $H(Z) - \epsilon \leq -\frac{1}{k} p(z_1, z_2, ..., z_k) \leq H(Z) + \epsilon$*
2. *$p(A_\epsilon^k) > 1 - \epsilon$ for $k$ sufficiently large*

3. $(1-\epsilon)2^{k(H(Z)-\epsilon)} \leq |A_\epsilon^k| \leq 2^{k(H(Z)+\epsilon)}$ *for* $k$ *sufficiently large. Here* $|A|$ *is the number of elements in set* $A$.

Thus a typical set $A_\epsilon^k$ has probability nearly 1, typical sequences are nearly equiprobable (with probability nearly $2^{-kH(Z)}$), and the number of typical sequences is nearly $2^{kH(Z)}$.

Now we rewrite $X$ as a sequence of $k$ symbols of a random variable

$$W : (\Omega_W, \mathcal{O}_W) \to (\mathcal{W}, \mathcal{B}_W),$$

so that $X = W^k$. The next theorem considers the behavior of the pair $(W, Z)$.

Definition 12. *The set* $A_\epsilon^k$ *of* jointly typical *sequences* $\{(w^k, z^k)\}$ *with respect to the joint distribution* $p(w, z)$ *on* $W \times Z$ *is the set*

$$\begin{aligned} A_\epsilon^k \;=\; & \left\{(w^k, z^k) \in W^k \times Z^k : \right. \\ & 2^{-k(H(W)+\epsilon)} \leq p(w^k) \leq 2^{-k(H(W)-\epsilon)}, \\ & 2^{-k(H(Z)+\epsilon)} \leq p(z^k) \leq 2^{-k(H(Z)-\epsilon)}, \\ & \left. 2^{-k(H(W,Z)+\epsilon)} \leq p(w^k, z^k) \leq 2^{-k(H(W,Z)-\epsilon)} \right\}, \end{aligned}$$

Theorem 13 (Asymptotic Equipartition Property for jointly typical sequences). *(p. 195 of [17]) Let* $(W^k, Z^k)$ *be a pair of i.i.d. sources. Then*

1. $p(A_\epsilon^k) > 1 - \epsilon$.
2. $(1-\epsilon)2^{k(H(W,Z)-\epsilon)} \leq |A_\epsilon^k| \leq 2^{k(H(W,Z)+\epsilon)}$ *for* $n$ *sufficiently large.*
3. *If* $(\tilde{W}^k, \tilde{Z}^k)$ *are a pair of random variables with joint probability* $p(w^k, z^k) = p(w^k)p(z^k)$ *(i.e.* $\tilde{W}^k$ *and* $\tilde{Z}^k$ *are independent with the same marginal distributions as* $W^k$ *and* $Z^k$*), then for sufficiently large* $k$*,*

$$(1-\epsilon)2^{-k(I(W;Z)+3\epsilon)} \leq p\left((\tilde{W}^k, \tilde{Z}^k) \in A_\epsilon^k\right) \leq 2^{-k(I(W;Z)-3\epsilon)}.$$

Thus, a jointly typical set $A_\epsilon^k$ has probability close to 1. The number of jointly typical sequences is nearly $2^{kH(W,Z)}$ and they are each nearly equiprobable (with probability close to $2^{-kI(W;Z)}$). Cover and Thomas (p. 197 of [17]) give the following argument to ascertain the number of distinguishable signals $W^k$ given a signal $Z^k$. Observe that there are about $2^{kH(W)}$ typical $W$ sequences and about $2^{kH(Z)}$ typical $Z$ sequences. However, as pointed out above, there are only about $2^{kH(W,Z)}$ jointly typical sequences. Since a jointly typical sequence has probability close to $2^{-kI(W;Z)}$, then, for a fixed $Z^k$, we can consider about $2^{kI(W;Z)}$ such pairs before we are likely

to find a jointly typical pair. This suggests that the set of jointly typical sequences can be divided into $2^{kI(W,Z)}$ disjoint sets, such that projections of these sets to $W^k$ as well as to $Z^k$ are almost disjoint. This justifies Figure 5B for spaces $X = W^k$ and $Y = Z^k$.

A source $Y$ can be related to another random variable $Y_N$ through the process of *quantization* or *lossy compression* [17, 35]. $Y_N$ is referred to as the *reproduction* of $Y$. The process is defined by a conditional probability map

$$q(Y_N|Y) : \mathcal{Y} \to \mathcal{Y}_N,$$

called a *quantizer* as in (1.20). Without loss of generality, and for simplification of the notation, we assume that the elements or *classes* of $Y_N$ are the natural numbers,

$$\mathcal{Y}_N = \{1, 2, ..., N\}.$$

We will use Greek letters such as $\nu, \delta, \omega, \mu$ and $\eta$ when referring to the classes of $Y_N$. As we point out in the Notation and Definition section of this chapter, we will write

$$q(Y_N = \nu|Y = y_k) = q(\nu|y_k) = q_{\nu k}.$$

If we assume that $|\mathcal{Y}| = K$, then $q(Y_N|Y)$ is defined by an $N \times K$ matrix, given by

$$\begin{pmatrix} q(1|y_1) & q(1|y_2) & q(1|y_3) & ... & q(1|y_K) \\ q(2|y_1) & q(2|y_2) & q(2|y_3) & ... & q(2|y_K) \\ \vdots & \vdots & \vdots & & \vdots \\ q(N|y_1) & q(N|y_2) & q(N|y_3) & ... & q(N|y_K) \end{pmatrix}.$$

In general, quantizers are stochastic: $q$ assigns to each $y \in \mathcal{Y}$ the probability that the response $y$ belongs to an abstract class $\nu \in Y_N$. A *deterministic quantizer* is a special case in which $q_{\nu k}$ takes the values of 0 or 1 for every $\nu$ and $k$. The *uniform quantizer*, which we denote by $q_{\frac{1}{N}}$, is the special case when

$$q_{\frac{1}{N}}(\nu|y_k) = \frac{1}{N} \tag{2.7}$$

for every $\nu$ and $k$. The constraint space $\Delta$ from (1.11),

$$\Delta := \left\{ q(Y_N|Y) \mid \sum_{\nu \in \mathcal{Y}_N} q(\nu|y) = 1 \text{ and } q(\nu|y) \geq 0 \ \forall y \in \mathcal{Y} \right\},$$

is the space of valid quantizers in $\Re^{NK}$.

It can be shown [35] that the mutual information $I(X;Y)$ is the least upper bound of $I(X;Y_N)$ over all possible reproductions $Y_N$ of $Y$. Hence, the original mutual information can be approximated with arbitrary precision using carefully chosen reproduction spaces.

The new random variable $Y_N$ produced by a quantization $q(Y_N|Y)$ has associated probabilities $p(Y_N)$, computed by

$$p(Y_N = \nu) = \sum_y q(\nu|y)p(y).$$

Given an information channel $(X, Y)$, the random variables $X, Y, Y_N$ form a *Markov chain* [22]

$$X \leftrightarrow Y \leftrightarrow Y_N,$$

which means that

$$p(X = x, Y = y, Y_N = \nu) = p(x)p(y|x)q(\nu|y)$$

and that

$$\begin{aligned} p(X = x, Y = y, Y_N = \nu) &= p(\nu)p(y|\nu)p(x|y) \\ &= p(y)q(\nu|y)p(x|y). \end{aligned} \tag{2.8}$$

The Distortion Function $D(q)$

The class of problems (1.9)

$$\max_{q\in\Delta}(G(q) + \beta D(q))$$

which we analyze in this thesis contain the cost functions used in Rate Distortion Theory [17, 35], Deterministic Annealing [61], the Information Distortion [22, 20, 29] and the Information Bottleneck methods [78, 70, 69]. We discuss the explicit form of the function $D(q)$, called a *distortion function*, for each of these scenarios.

Rate Distortion Theory is the information theoretic approach to the study of optimal source coding systems, including systems for quantization and data compression [35]. To define how well a source, the random variable $Y$, is represented by a particular representation using $N$ symbols, which we call $Y_N$, one introduces a *distortion function* between $Y$ and $Y_N$

$$D(q(Y_N|Y)) = D(Y, Y_N) = E_{Y,Y_N} d(Y, Y_N) = \sum_y \sum_{\nu\in\mathcal{Y}_N} q(\nu|y)p(y)d(y, \nu)$$

where $d(Y, Y_N)$ is the *pointwise distortion function* on the individual elements of $\mathcal{Y}$ and $\mathcal{Y}_N$. $q(Y_N|Y)$ is the quantization of $\mathcal{Y}$ into the representation $\mathcal{Y}_N$. A representation $Y_N$ is said to be optimal if there is a quantizer $q^*(Y_N|Y)$ such that

$$D(q^*) = \min_{q\in\Delta} D(q). \tag{2.9}$$

In engineering and imaging applications, the distortion function is usually chosen as the *mean squared error* [17, 61, 31],

$$\hat{D}(Y, Y_N) = E_{Y,Y_N}\hat{d}(Y, Y_N) = \sum_{y}\sum_{\nu\in\mathcal{Y}_N} q(\nu|y)p(y)\hat{d}(y,\nu), \tag{2.10}$$

where the pointwise distortion function $\hat{d}(Y, Y_N)$ is the Euclidean squared distance,

$$\hat{d}(Y = y, Y_N = \nu) = \|y - \nu\|^2.$$

This requires that $\mathcal{Y}, \mathcal{Y}_N \subset \Re^{NK}$. In this case, $\hat{D}(Y, Y_N)$ is a linear function of the quantizer $q$.

## The Information Distortion Problem

In neural coding, as we have seen in chapter 1, one can model the neural decoder by $p(X|Y)$, the stochastic map from the space of neural responses $\mathcal{Y}$ to the stimulus space $\mathcal{X}$. The Information Distortion method examined in chapter 1 determines an approximation to $p(X|Y)$ by quantizing the neural responses $\mathcal{Y}$ into a reproduction space $\mathcal{Y}_N$ by minimizing a distortion function as in (2.9). We now determine the explicit form of the distortion function used by the Information Distortion method, which we call the *information distortion measure*, then show how one optimizes this function.

### The Information Distortion Measure

Since the metric between spike trains may not coincide with Euclidean distance [83, 84] (see (1.19)), the Information Distortion method does not impose $\hat{D}(q)$ from (2.10) as the distortion function when searching for a neural decoder.

The natural measure of closeness between two probability distributions is the Kullback-Leibler divergence (see (2.6)) [22]. For each fixed $y \in \mathcal{Y}$ and $\nu \in \mathcal{Y}_N$, $p(X|Y = y)$ and $p(X|Y_N = \nu)$ are a pair of distributions on the space $X$. As a pointwise distortion function, consider

$$d(Y, Y_N) = KL(p(X|y)||p(X|\nu)).$$

Unlike the pointwise distortion functions usually investigated in information theory [17, 61], $D_I$ explicitly considers a third space, $\mathcal{X}$, of inputs, and it is a nonlinear function of the quantizer $q(Y_N|Y)$ through

$$\begin{aligned} p(X = x|Y_N = \nu) &= \sum_{y} \frac{p(x, y, \nu)}{p(\nu)} \\ &= \sum_{y} \frac{q(\nu|y)p(y)p(x|y)}{p(\nu)}, \end{aligned}$$

where the last equality follows from (2.8) . The *information distortion measure* is defined as the expected Kullback-Leibler divergence over all pairs $(y, \nu)$

$$D_I(q(Y_N|Y)) = D_I(Y, Y_N) := E_{Y,Y_N} KL(p(X|Y = y)||p(X|Y_N = \nu)). \tag{2.11}$$

We derive an alternate expression for $D_I$. Starting from the definition

$$\begin{aligned} D_I &= \sum_{y\in\mathcal{Y},\nu\in\mathcal{Y}_N} p(y,\nu)\ KL\big(p(X|y)\|p(X|\nu)\big) \\ &= \sum_{y,\nu} p(y,\nu) \sum_x p(x|y) \log \frac{p(x|y)}{p(x|\nu)} \\ &= \sum_{x,y,\nu} p(x,y,\nu)\Big(\log p(x|y) - \log p(x|\nu)\Big) \qquad (2.12) \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} - \sum_{x,\nu} p(x,\nu) \log \frac{p(x,\nu)}{p(x)p(\nu)} \qquad (2.13) \\ &= I(X;Y) - I(X;Y_N) \end{aligned}$$

In (2.12) we used the Markov property (2.8), and (2.13) is justified by using the identities $p(x, y) = \sum_\nu p(x, y, \nu)$, $p(x, \nu) = \sum_y p(x, y, \nu)$ and the Bayes property $p(x, y)/p(y) = p(x|y)$. This shows that the information distortion measure can be written as

$$D_I = I(X;Y) - I(X;Y_N).$$

Recall from (2.9) that the goal is to find a quantization $q(\nu|y)$ for a fixed reproduction size $N$ that minimizes the information distortion measure $D_I$

$$\min_{q\in\Delta} D_I. \tag{2.14}$$

Since the only term in $D_I$ that depends on the quantizer is $I(X; Y_N)$, we can replace $D_I$ with the effective distortion

$$D_{eff} := I(X;Y_N)$$

in the optimization problem. Thus, the minimizer of (2.14) is the maximizer of

$$\max_{q\in\Delta} D_{eff}. \tag{2.15}$$

Applying the information distortion measure to neural data, which, as we have just seen, is equivalent to maximizing the mutual information between the stimulus and the quantized neural responses, has theoretical justification [9, 20, 22, 37, 51, 59, 64, 72, 83, 84].

The Information Bottleneck method is another unsupervised non-parametric data clustering technique [78, 70, 69] which has been applied to document classification,

gene expression, neural coding [64] and spectral analysis. It also uses $D_I(q)$ as the distortion function.

The Maximal Entropy Problem

Solving (2.15) directly is difficult using many numerical optimization techniques since there are many local, suboptimal maxima on the boundary of $\Delta$ [61, 22]. This is not surprising since $D_{eff}$ is convex and $\Delta$ is a convex domain. To deal with this issue, the Information Distortion method introduces a strictly concave function, the entropy $H(Y_N|Y)$, to maximize simultaneously with $D_{eff}$, which serves to regularize the problem (2.15) [61],

$$\begin{aligned} &\max_{q\in\Delta} H(Y_N|Y) \quad \text{constrained by} \\ &D_{eff}(q) \geq I_0 \end{aligned} \tag{2.16}$$

In other words, of all the local solutions $q^*$ to (2.15), the method seeks the one that maximizes the entropy. Using the entropy as a regularizer is justified by Jayne's maximum entropy principle (see Remark 3), since among all quantizers that satisfy a given set of constraints, the maximum entropy quantizer does not implicitly introduce additional constraints in the problem [36]. Thus, the problem of optimal quantization (2.15) is reformulated [22] as a maximum entropy problem with a distortion constraint (2.16). The goal is to find the maximal entropy solution for a maximal possible value of $D_{eff}$.

Tishby et al. use the concave function $I(Y;Y_N)$ as a regularizer [70, 78]. The fact that $I(Y;Y_N)$ is concave (and not strictly concave) causes some difficulties for numerics, which we discuss in chapter 4.

The conditional entropy $H(Y_N|Y)$ and the function $D_{eff}$, can be written explicitly in terms of $q_{\nu k} = q(\nu \mid y_k)$

$$\begin{aligned} H(Y_N \mid Y) &= -E_{Y,Y_N} \log q(Y_N|Y) \\ &= -\sum_{\nu,k} p(y_k) q_{\nu k} \log\left(q_{\nu k}\right) \end{aligned} \tag{2.17}$$

and

$$\begin{aligned} D_{eff} = I(X;Y_N) &= E_{X,Y_N} \log \frac{p(X,Y_N)}{p(X)p(Y_N)} \\ &= \sum_{\nu,k,i} q_{\nu k} p(x_i,y_k) \log\left(\frac{\sum_k q_{\nu k} p(x_i,y_k)}{p(x_i)\sum_k p(y_k) q_{\nu k}}\right). \end{aligned} \tag{2.18}$$

Derivatives

To find local solutions of (2.16) (see chapter 9), we compute the first and second derivatives of $H(Y_N|Y)$ and $D_{eff}$. To determine the bifurcation structure of these solutions (see chapter 6), we compute the third and fourth derivatives.

The gradient of $H(Y_N|Y)$ with respect to $q$ is [22]

$$\begin{aligned}(\nabla H)_{\nu k} &\equiv -\frac{\partial H(Y_N|Y)}{\partial q_{\nu k}} \\ &= -p(y_k)\Big(\log q_{\nu k} + \frac{1}{\ln 2}\Big).\end{aligned} \tag{2.19}$$

The Hessian of $H(Y_N|Y)$ is [22]

$$\begin{aligned}\frac{\partial^2 H(Y_N|Y)}{\partial q_{\eta l}\partial q_{\nu k}} &= -\frac{\partial}{\partial q_{\eta l}} p(y_k)\Big(\log q_{\nu k} + \frac{1}{\ln 2}\Big) \\ &= -\frac{p(y_k)}{(\ln 2) q_{\nu k}}\delta_{\nu\eta}\delta_{kl}.\end{aligned} \tag{2.20}$$

The three dimensional array of third derivatives is

$$\begin{aligned}\frac{\partial^3 H(Y_N|Y)}{\partial q_{\delta m}\partial q_{\eta l}\partial q_{\nu k}} &= -\frac{\partial}{\partial q_{\delta m}}\frac{p(y_k)}{(\ln 2) q_{\nu k}}\delta_{\nu\eta}\delta_{kl} \\ &= \frac{p(y_k)}{(\ln 2) q_{\nu k}^2}\delta_{\nu\eta\delta}\delta_{klm}.\end{aligned} \tag{2.21}$$

The four dimensional array of fourth derivatives is

$$\begin{aligned}\frac{\partial^4 H(Y_N|Y)}{\partial q_{\mu p}\partial q_{\delta m}\partial q_{\eta l}\partial q_{\nu k}} &= \frac{\partial}{\partial q_{\mu p}}\frac{p(y_k)}{(\ln 2) q_{\nu k}^2}\delta_{\nu\eta\delta}\delta_{klm} \\ &= -\frac{2}{(\ln 2)}\frac{p(y_k)}{q_{\nu k}^3}\delta_{\nu\eta\delta\mu}\delta_{klmp}.\end{aligned} \tag{2.22}$$

The gradient of $D_{eff}$ is [22]

$$\begin{aligned}(\nabla D_{eff})_{\nu k} &\equiv \frac{\partial D_{eff}}{\partial q_{\nu k}} \\ &= \sum_i p(x_i, y_k)\log\frac{\sum_k q_{\nu k} p(x_i, y_k)}{p(x_i)\sum_k q_{\nu k} p(y_k)}.\end{aligned}$$

The Hessian of $D_{eff}$ is [22]

$$\begin{aligned}\frac{\partial^2 D_{eff}}{\partial q_{\eta l}\partial q_{\nu k}} &= \frac{\partial}{\partial q_{\eta l}}\sum_i p(x_i, y_k)\log\frac{\sum_k q_{\nu k} p(x_i, y_k)}{p(x_i)\sum_k q_{\nu k} p(y_k)} \\ &= \frac{\delta_{\nu\eta}}{\ln 2}\left(\sum_i \frac{p(x_i, y_k)\ p(x_i, y_l)}{\sum_k q_{\nu k} p(x_i, y_k)} - \frac{p(y_k)p(y_l)}{\sum_k q_{\nu k} p(y_k)}\right).\end{aligned} \tag{2.23}$$

The three dimensional array of third derivatives $\frac{\partial^3 D_{eff}}{\partial q_{\delta m}\partial q_{\eta l}\partial q_{\nu k}}$ is

$$\frac{\delta_{\nu\eta\delta}}{\ln 2}\left(\frac{p(y_k)p(y_l)p(y_m)}{\left(\sum_k q_{\nu k} p(y_k)\right)^2} - \sum_i \frac{p(x_i, y_k)\ p(x_i, y_l)p(x_i, y_m)}{\left(\sum_k q_{\nu k} p(x_i, y_k)\right)^2}\right). \tag{2.24}$$

The four dimensional array of fourth derivatives $\frac{\partial^4 D_{eff}}{\partial q_{\mu p}\partial q_{\delta m}\partial q_{\eta l}\partial q_{\nu k}}$ is

$$\frac{2\delta_{\nu\eta\delta\mu}}{\ln 2}\left(\sum_i \frac{p(x_i,y_k)\ p(x_i,y_l)p(x_i,y_m)p(x_i,y_p)}{\left(\sum_k q_{\nu k}p(x_i,y_k)\right)^3} - \frac{p(y_k)p(y_l)p(y_m)p(y_p)}{\left(\sum_k q_{\nu k}p(y_k)\right)^3}\right). \quad (2.25)$$

Dealing with Complex Inputs

To successfully apply the Information Distortion method to physiological data, we need to estimate the information distortion $D_{eff}$, which in turn depends on the joint stimulus/response probability $p(X,Y)$. If the stimuli are sufficiently simple, $p(X,Y)$ can be estimated directly as a joint histogram, and the method applied by solving (2.16). In general, we want to analyze conditions close to the natural for the particular sensory system, which usually entails observing stimulus sets of high dimensionality. Characterizing such a relationship non-parametrically is extremely difficult, since usually one cannot provide the large amounts of data this procedure needs [51]. To cope with this regime, we model the stimulus/response relationship [23, 25]. The formulation as an optimization problem suggests certain classes of models which are better suited for this approach. We shall look for models that give us strict lower bounds $\tilde{D}_{eff}$ of the information distortion function $D_{eff}$. In this case, when we maximize the lower bound $\tilde{D}_{eff}$, the actual value of $D_{eff}$ is also increased, since $I(X;Y) \geq D_{eff} \geq \tilde{D}_{eff} \geq 0$. This also gives us a quantitative measure of the quality of a model: a model with a larger $\tilde{D}_{eff}$ is better.

In [24, 25, 29] the authors modelled the class conditioned stimulus $p(X|Y_N = \nu)$ with the Gaussian:

$$p(X|Y_N = \nu) = N(x_\nu, C_{X|\nu}). \quad (2.26)$$

The class conditioned stimulus mean $x_\nu$ and covariance matrix $C_{X|\nu}$ can be estimated from data. The stimulus estimate obtained in this manner is effectively a Gaussian mixture model [18]

$$p(X) = \sum_\nu p(\nu)N(x_\nu, C_{X|\nu})$$

with weights $p(\nu)$ and Gaussian parameters $(x_\nu, C_{X|\nu})$. This model produces an upper bound [59] $\tilde{H}(X|Y_N)$ of $H(X|Y_N)$:

$$\tilde{H}(X|Y_N = \nu) = \sum_\nu p(\nu)\frac{1}{2}\log(2\pi e)^{|X|}\det\Big[\sum_y p(y|\nu)(C_{X|y} + x_y^2) - (\sum_y p(y|\nu)x_y)^2\Big]. \quad (2.27)$$

Here $x_y^2$ is the matrix $x_y x_y^T$.

Since $\tilde{H}(X|Y_N)$ is an upper bound on $H(X|Y_N)$ and

$$D_{eff} = I(X;Y_N) = H(X) - H(X|Y_N),$$

the quantity

$$\tilde{D}_{eff}(q(Y_N|Y) := H(X) - \tilde{H}(X|Y_N) \tag{2.28}$$

is the lower bound to $D_{eff}$. This transforms the optimization problem (2.16) for physiological data to

$$\begin{aligned} &\max_{q\in\Delta} H(Y_N|Y) && \text{constrained by} \\ &\tilde{D}_{eff}(q(\nu|y)) \geq I_0 && \text{and} \\ &\sum_{\nu\in\mathcal{Y}_N} q(\nu|y) = 1 \quad \text{and} \quad q(\nu|y) \geq 0 \quad \forall y \in Y. \end{aligned} \tag{2.29}$$

It is not immediately obvious that solutions to (2.29) have properties similar to the solutions of (2.16). Gedeon et al. [29] showed that $\tilde{D}_{eff}$ is convex in $q(Y_N|Y)$. This implies that for the problem (2.29), the optimal quantizer $q^*(Y_N|Y)$ will be generically deterministic (Theorems 153 and 154). This means that $\tilde{D}_{eff}$ can be used in place of $D_{eff}$ in the problem (2.34).

## The Function $G(q)$

The class of problems (1.9)

$$\max_{q\in\Delta} G(q) + \beta D(q)$$

which we analyze in this thesis contain similar cost functions used in Rate Distortion Theory [17, 35], Deterministic Annealing [61], the Information Distortion [22, 20, 29] and the Information Bottleneck methods [78, 70, 69]. In this section we discuss the explicit form of the function $G(q)$ for each of these scenarios.

There are two related methods used to analyze communication systems at a distortion $D(q) \leq D_0$ for some given $D_0 \geq 0$ [17, 35, 61]. In rate distortion theory [17, 35], the problem of finding a minimum rate at a given distortion is posed as a *minimal information rate* distortion problem (as in (1.5)):

$$\begin{matrix} \min_{q\in\Delta} I(Y;Y_N) \\ D(Y,Y_N) \leq D_0 \end{matrix} . \tag{2.30}$$

This formulation is justified for i.i.d. sources by the Rate Distortion Theorem [17].

A similar exposition using the Deterministic Annealing approach [61] is a *maximal entropy* problem (as in (1.2))

$$\begin{matrix} \max_{q\in\Delta} H(Y_N|Y) \\ D(Y;Y_N) \leq D_0 \end{matrix} . \tag{2.31}$$

The justification for using (2.31) is Jayne's maximum entropy principle [36] (see Remark 3). The formulations (2.30) and (2.31) are related since

$$I(Y;Y_N) = H(Y_N) - H(Y_N|Y).$$

Let $I_0 > 0$ be some given information rate. In (2.16), the neural coding problem is formulated as an entropy problem as in (2.31)

$$\begin{array}{c} \max_{q\in\Delta} H(Y_N|Y) \\ D_{eff}(q) \geq I_0 \end{array} \tag{2.32}$$

which uses the nonlinear effective information distortion measure $D_{eff}$. Tishby et. al. [78, 70] pose an information rate distortion problem as in (2.30)

$$\begin{array}{c} \min_{q\in\Delta} I(Y;Y_N) \\ D_{eff}(q) \geq I_0 \end{array} . \tag{2.33}$$

Using the method of Lagrange multipliers, the rate distortion problems (2.30), (2.31), (2.32), (2.33) can be reformulated as finding the maxima of

$$\max_{q\in\Delta} F(q,\beta) = \max_{q\in\Delta} \left(G(q) + \beta D(q)\right)$$

as in (1.9) where $\beta \in [0,\infty)$. This construction removes the nonlinear constraint from the problem and replaces it with a parametric search in $\beta(I_0)$. For the maximal entropy problem (2.32),

$$F(q,\beta) = H(Y_N|Y) + \beta D_{eff}(q) \tag{2.34}$$

and so in this case $G(q)$ from (1.9) is the conditional entropy $H(Y_N|Y)$ (compare with (1.4)). For the minimal information rate distortion problem (2.33),

$$F(q,\beta) = -I(Y;Y_N) + \beta D_{eff}(q) \tag{2.35}$$

and so here $G(q) = -I(Y;Y_N)$ (compare with (1.6)).

We now compare the two formulations (2.32) and (2.34). In [22, 29, 61], one explicitly considers (2.34) for $\beta = \infty$. This involves taking

$$\lim_{\beta\to\infty} \max_{q\in\Delta} F(q,\beta) = \max_{q\in\Delta} D_{eff}(q)$$

which in turn gives $\min_{q\in\Delta} D_I$. This observation can be made rigorous by noting that $D_{eff}$, as a continuous function on a compact domain $\Delta$, has a maximal value $I^*$. Therefore, for values of the parameter $I_0 > I^*$ problem (2.32) has no solution. On the other hand, problem (2.34) has a solution for all values of $\beta$, since $F$ is a continuous function on a compact set $\Delta$. We have the following result

LEMMA 14. *[29] Let $q^*$ be a solution of (2.32) with $I_0 = I^*$. Let $q(\beta)$ be a solution of problem (2.34) as a function of the annealing parameter $\beta$. Then*

$$\lim_{\beta\to\infty} D_{eff}(q(\beta)) \to I^*.$$

*Proof.* As $\beta \to \infty$ the solution $q(\beta)$ converges to the solution of the problem

$$\max_{q \in \Delta} D_{eff}.$$

The maximum of $D_{eff}$ on $\Delta$ is $I^*$. □

In the Information Bottleneck method, one may only be interested in solutions to (2.35) for finite $\mathcal{B}$ which takes into account a tradeoff between $I(Y;Y_N)$ and $D_{eff}$.

# CHAPTER 3

## THE DYNAMICAL SYSTEM

When using the method of annealing, Algorithm 1, to solve (1.9),

$$\max_{q\in\Delta}(G(q)+\beta D(q)),$$

one obtains a sequence of solutions $\{(q_k,\beta_k)\}$ that converge to $(q^*,\mathcal{B})$, where $\mathcal{B}\in(0,\infty)$, and

$$q^*=\text{argmax}_{q\in\Delta}(G(q)+\mathcal{B}D(q)).$$

As we explained in chapter 1, it has been observed that the solution branch which contains $\{(q_k,\beta_k)\}$ undergoes bifurcations or phase transitions. The purpose of this chapter is to formulate a dynamical system so that we may study the bifurcation structure of these solutions. First, we must present the rudiments of Constrained Optimization Theory. Then we present the formulation of the dynamical system, whose equilibria are the stationary points of (1.9).

### The Optimization Problem

The objective of this thesis is to solve the problem (1.9). We now pose a slightly different optimization problem, one which does not explicitly enforce the nonnegativity constraints of $\Delta$, which will help us to understand the bifurcation structure of solutions to (1.9) (see Remarks 19 and 28).

Consider the optimization problem

$$\max_{q\in\Delta_{\mathcal{E}}}F(q,\beta) \tag{3.1}$$

for fixed $\beta=\mathcal{B}\in[0,\infty)$, where

$$F(q,\beta)=G(q)+\beta D(q) \tag{3.2}$$

as in (1.9) and (1.10), and

$$\Delta_{\mathcal{E}}:=\left\{q\in\Re^{NK}|\sum_{\nu\in\mathcal{Y}_N}q_{\nu k}=1\ \ \forall\ y_k\in\mathcal{Y}\right\}$$

(compare with (1.11)). As with Assumptions 2 on (1.9), we assume that

Assumption 15.

1. *$G$ and $D$ are real valued functions of $q(Y_N|Y)$, which depend on $Y_N$ only through $q$, are invariant to relabelling of the elements or* classes *$\nu$ of $Y_N$ That is, $G$ and $D$ are $S_N$-invariant, with the explicit group action defined in (6.6).*

2. *$G$ and $D$ are sufficiently smooth in $q$ and $\beta$ on the interior of $\Delta$.*

3. *The Hessians of $G$ and $D$ are block diagonal.*

Assumption 15 holds for the Information Distortion and the Information Bottleneck cost functions (2.34) and (2.35). We prove this claim in the former case in Theorem 73.

We rewrite (3.1) using its Lagrangian

$$\mathcal{L}(q, \lambda, \beta) = F(q, \beta) + \sum_{k=1}^{K} \lambda_k \left( \sum_{\nu=1}^{N} q_{\nu k} - 1 \right), \tag{3.3}$$

where the scalar $\lambda_k$ is the Lagrange multiplier for the constraint $\sum_{\nu=1}^{N} q_{\nu k} - 1 = 0$, and $\lambda$ is the $K \times 1$ vector of Lagrange multipliers

$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_K \end{pmatrix}.$$

The gradient of (3.3) is

$$\nabla_{q,\lambda} \mathcal{L}(q, \lambda, \beta) = \begin{pmatrix} \nabla_q \mathcal{L} \\ \nabla_\lambda \mathcal{L} \end{pmatrix}, \tag{3.4}$$

where

$$\nabla_q \mathcal{L} = \nabla F(q, \beta) + \Lambda, \tag{3.5}$$

and $\Lambda = \begin{pmatrix} \lambda^T & \lambda^T & \dots & \lambda^T \end{pmatrix}^T$, an $NK \times 1$ vector. The gradient $\nabla_\lambda \mathcal{L}$ is the vector of $K$ constraints

$$\nabla_\lambda \mathcal{L} = \begin{pmatrix} \sum_\nu q_{\nu 1} - 1 \\ \sum_\nu q_{\nu 2} - 1 \\ \vdots \\ \sum_\nu q_{\nu K} - 1 \end{pmatrix} \tag{3.6}$$

imposed by $\Delta_{\mathcal{E}}$. Let $J$ be the $K \times NK$ Jacobian of (3.6)

$$J := \partial_q \nabla_\lambda \mathcal{L} = \partial_q \begin{pmatrix} \sum_\nu q_{\nu 1} - 1 \\ \sum_\nu q_{\nu 2} - 1 \\ \vdots \\ \sum_\nu q_{\nu K} - 1 \end{pmatrix} = \underbrace{\begin{pmatrix} I_K & I_K & \dots & I_K \end{pmatrix}}_{N \text{ blocks}}. \tag{3.7}$$

Observe that $J$ has full row rank. The $(NK+K)\times(NK+K)$ Hessian of (3.3) is

$$\Delta_{q,\lambda}\mathcal{L}(q,\lambda,\beta)=\begin{pmatrix}\Delta F(q,\beta) & J^T\\ J & \mathbf{0}\end{pmatrix}, \tag{3.8}$$

where $\mathbf{0}$ is $K\times K$. The $NK\times NK$ matrix $\Delta F$ is the block diagonal Hessian of $F$ (Assumption 15.3),

$$\Delta F=\begin{pmatrix}B_1 & \mathbf{0} & ... & \mathbf{0}\\ \mathbf{0} & B_2 & ... & \mathbf{0}\\ \vdots & \vdots & & \vdots\\ \mathbf{0} & \mathbf{0} & ... & B_N\end{pmatrix}, \tag{3.9}$$

where $\mathbf{0}$ and $B_i$ are $K\times K$ matrices for $i=1,...,N$.

There are optimization schemes, such as the implicit solution (see (9.20)) and projected Augmented Lagrangian methods (Algorithm 149), which exploit the structure of (3.3) and (3.4) to find local solutions to (3.1). This exploitation depends on the following *first order* necessary conditions:

THEOREM 16 (KARUSH-KUHN-TUCKER CONDITIONS). *([50] p328) Let $x^*$ be a local solution of*

$$\max_{x\in\Omega} f(x)$$

*where the constraint space $\Omega$ is defined by some equality constraints, $c_i(x)=0, i\in\mathcal{E}$, and some inequality constraints, $c_i(x)\geq 0, i\in\mathcal{I}$. Suppose that the Jacobian of the constraints has full row rank. Then there exists a vector of Lagrange multipliers, $\lambda^*$, with components $\lambda_i$, $i\in\mathcal{E}\cup\mathcal{I}$ such that*

$$\begin{aligned}\nabla_x f(x^*) &= -\sum_{i\in\mathcal{E}\cup\mathcal{I}}\lambda_i\nabla_x c_i(x^*)\\ c_i(x^*) &= 0 \text{ , for all } i\in\mathcal{E}\\ c_i(x^*) &\geq 0 \text{ , for all } i\in\mathcal{I}\\ \lambda^* &\geq 0 \text{ , for all } i\in\mathcal{I}\\ \lambda^* c_i(x^*) &= 0 \text{ , for all } i\in\mathcal{E}\cup\mathcal{I}\end{aligned} \tag{3.10}$$

REMARK 17. *Using the notation from Theorem 16, the equality constraints from (1.9) and (3.1) are represented as*

$$\{c_i(q)\}_{i\in\mathcal{E}}=\left\{\sum_\nu q_{\nu k}-1\right\}_{k=1}^{K}. \tag{3.11}$$

*Thus, if $q\in\Delta_{\mathcal{E}}$, then $c_i(q)=0$ for every $i\in\mathcal{E}$. For the inequality constraints which are present only in the problem (1.9), we have that*

$$\{c_i(q)\}_{i\in\mathcal{I}}=\{q_{\nu k}\}_{\nu\in Y_N,1\leq k\leq K}. \tag{3.12}$$

*In this case then, $q\in\Delta$ implies that $c_i(q)\geq 0$ for every $i\in\mathcal{I}$.*

The *Karush-Kuhn-Tucker* or KKT conditions for solutions of (3.1) only entail equality constraints. Furthermore, the Jacobian of these equality constraints is the matrix with full row rank given in (3.7). We have the following corollary.

COROLLARY 18. *Let $q^*$ be a local solution of (3.1) for some fixed $\beta$. Then there exists a vector of Lagrange multipliers, $\lambda^* \in \Re^K$, such that*

$$\begin{aligned} \nabla_q \mathcal{L}(q^*, \lambda^*, \beta) &= \mathbf{0} \\ [\nabla_\lambda \mathcal{L}(q^*, \lambda^*, \beta)]_k &= \sum_\nu q_{\nu k} - 1 = 0. \end{aligned}$$

Recall that a stationary point of a differentiable function $f(\boldsymbol{x})$ is a point where $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*) = \mathbf{0}$. A stationary point of a constrained system such as (3.1) is a point where $\nabla_{q,\lambda}\mathcal{L} = 0$. In other words, it is a point where the KKT conditions are satisfied.

REMARK 19. *One reason we consider the problem (3.1) instead of (1.9) is the following. The Lagrangian for the latter maximization problem is*

$$\hat{\mathcal{L}}(q, \lambda, \xi, \beta) = F(q, \beta) + \sum_{k=1}^{K} \lambda_k (\sum_{\nu=1}^{N} q_{\nu k} - 1) + \sum_{k=1}^{K} \sum_{\nu=1}^{N} \xi_{\nu k} q_{\nu k}, \tag{3.13}$$

*where $\{\lambda_k\}$ are the Lagrange multipliers for the equality constraints (3.11) and $\{\xi_{\nu k}\}$ are the Lagrange multipliers for the inequality constraints (3.12). Thus, $[\nabla_\xi \hat{\mathcal{L}}]_{\nu k} = q_{\nu k}$. From this, (3.6), and (3.7), we see that the Jacobian of the constraints in this case is*

$$\partial_q \nabla_{\lambda,\xi} \hat{\mathcal{L}} = \begin{pmatrix} J \\ \boldsymbol{e}_1^T \\ \boldsymbol{e}_2^T \\ \vdots \\ \boldsymbol{e}_{NK}^T \end{pmatrix},$$

*which does not have full row rank as required by Theorem 16 since the row space of $J$ is a subspace of* $\mathrm{span}(\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_{NK})$.

*If $(q, \beta)$ is a stationary point of (1.9) in the interior of $\Delta$, then the inequality constraints (3.12) are* inactive*: $c_i(q^*) > 0$ for $i \in \mathcal{I}$. By requirement (3.10) of Theorem 16 and the fact that $\{c_i\}_{i\in\mathcal{I}} = \{q_{\nu k}\}_{\nu\in\mathcal{Y}_N, y_k\in\mathcal{Y}}$, then for the vector of Lagrange multipliers $\xi$ from (3.13), $\xi_{\nu k} = 0$ for every $\nu$ and $k$. Thus,*

$$\nabla_{q,\lambda} \hat{\mathcal{L}} = \nabla_{q,\lambda} \mathcal{L} = \mathbf{0} \tag{3.14}$$

*by Theorem 16, which shows that a stationary point to (1.9) in the interior of $\Delta$ is a stationary point of (3.1).*

For a general optimization problem, the best that any optimization scheme can accomplish is to procure a stationary point ([50] p.45). To determine whether a given

stationary point $q \in \Re^{NK}$ is truly a local solution of (3.1), one appeals to the following theorem:

THEOREM 20. *([50], p 345 and 348) Assume that the Jacobian of the constraints, $J$, has full row rank and that for some $q^* \in \Delta_{\mathcal{E}}$ there is a vector of Lagrange multipliers $\lambda^*$ such that the KKT conditions (Theorem 16) are satisfied. If*

$$\boldsymbol{w}^T \Delta_q \mathcal{L}(q^*, \lambda^*, \beta)\boldsymbol{w} < 0$$

*for all $\boldsymbol{w} \in \ker J$ then $q^*$ is a local solution for (3.1). Conversely, if $q^*$ is a local solution for (3.1), then*

$$\boldsymbol{w}^T \Delta_q \mathcal{L}(q^*, \lambda^*, \beta)\boldsymbol{w} \leq 0$$

*for all $\boldsymbol{w} \in \ker J$.*

Hence, to find a local solution of (3.1) for some $\beta$, we need to find $q^*$ such that $\nabla_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta) = \mathbf{0}$ and that $\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta)$ is negative definite on $\ker J$.

REMARK 21.

1. *The constraints of (3.1) are linear. It follows that $\Delta_q \mathcal{L}(q, \lambda, \beta) = \Delta F(q, \beta)$. Therefore, if we track $q^*$ where the KKT conditions are satisfied and where $\Delta F(q^*, \beta)$ is negative definite on $\ker J$, then we satisfy the assumptions of Theorem 20 which shows that $q^*$ is a local solution to (3.1).*

2. *Let $d := \dim \ker J$ and let $Z$ be the $NK \times d$ matrix with full column rank whose columns span $\ker J$. Thus, any $\boldsymbol{w} \in \ker J$ can be written as $Z\boldsymbol{u}$ for some $\boldsymbol{u} \in \Re^d$. The condition*

$$\boldsymbol{w}^T \Delta F(q^*, \beta)\boldsymbol{w} \leq 0 \quad \forall \boldsymbol{w} \in \ker J$$

*can be restated as*

$$\boldsymbol{u}^T Z^T \Delta F(q^*, \beta) Z\boldsymbol{u} \leq 0 \quad \forall \boldsymbol{u} \in \Re^d.$$

*Hence, the conditions of Theorem 20 become that $Z^T \Delta F(q^*, \beta) Z$ must be (non)-negative definite.*

## The Gradient Flow

We wish to pose (3.1) as a dynamical system in order to study bifurcations of its local solutions. This section provides the explicit dynamical system which we will study. First, some terminology is introduced. Let

$$\dot{\boldsymbol{x}} = \psi(\boldsymbol{x}, \beta), \tag{3.15}$$

where $\boldsymbol{x}$ is in some Banach space $B_2$ and $\beta \in \Re$, so that

$$\psi : B_2 \times \Re \to B_0 \tag{3.16}$$

for some Banach space $B_0$. The solutions $(\boldsymbol{x}, \beta) \in B_2 \times \Re$ which satisfy

$$\psi(\boldsymbol{x}, \beta) = \mathbf{0} \tag{3.17}$$

are *equilibria* of the system. Such a continuum of solutions is called a *solution branch* or a *branch of equilibria* of (3.15). The Jacobian of $\psi$ is $\partial_{\boldsymbol{x}}\psi$. Let $n(\beta)$ be the number of $\boldsymbol{x}$'s for which $(\boldsymbol{x}, \beta)$ is a solution of (3.17).

DEFINITION 22. $(\boldsymbol{x}^*, \beta^*)$ *is a* bifurcation point *if* $n(\beta)$ *changes as* $\beta$ *varies in a neighborhood of* $\beta^*$.

REMARK 23. *This definition of bifurcation, as used in [33], may seem too restrictive. However, the class of systems we study are gradient systems,* $\psi = \nabla_{\boldsymbol{x}} f$ *(compare with (3.15)), where* $f$ *is some scalar function. Thus, the bifurcations allowed by Definition 22 are the only ones that can occur. This is because the Jacobian,* $\partial_{\boldsymbol{x}}\psi = \Delta_{\boldsymbol{x}} f$*, is a symmetric matrix , and so it has only real eigenvalues [65]. Bifurcations not considered in Definition 22, such as Hopf bifurcations, require purely imaginary eigenvalues [6].*

THEOREM 24. *If* $(\boldsymbol{x}^*, \beta^*)$ *is a bifurcation of (3.17) then* $\partial_{\boldsymbol{x}}\psi(\boldsymbol{x}^*, \beta^*)$ *is singular.*

*Proof.* If $\partial_{\boldsymbol{x}}\psi(\boldsymbol{x}^*, \beta^*)$ is not singular then the Implicit Function Theorem gives that $\boldsymbol{x}^* = \boldsymbol{x}(\beta)$ is the unique solution of (3.17) about $(\boldsymbol{x}^*, \beta^*)$. Therefore, $(\boldsymbol{x}^*, \beta^*)$ cannot be a bifurcation point. □

DEFINITION 25. *If* $\partial_{\boldsymbol{x}}\psi(\boldsymbol{x}^*, \beta^*)$ *is singular, but* $(\boldsymbol{x}^*, \beta^*)$ *is not a bifurcation point of (3.17), then* $(\boldsymbol{x}^*, \beta^*)$ *is a* degenerate singularity.

Now back to our purpose stated at the beginning of this section: We wish to pose (3.1) as a dynamical system in order to study bifurcations of its local solutions. To this end, consider the equilibria of the *gradient flow*

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda}\mathcal{L}(q, \lambda, \beta) \tag{3.18}$$

for $\mathcal{L}$ as defined in (3.3) and $\beta \in [0, \infty)$. The equilibria of (3.18) are points $\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix}$ where

$$\nabla_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta) = 0.$$

The Jacobian of this system is the Hessian $\Delta_{q,\lambda}\mathcal{L}(q, \lambda, \beta)$ from (3.8).

DEFINITION 26. *An equilibrium* $(q^*, \lambda^*)$ *of (3.18) is* stable *if* $\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta)$ *is negative definite. The equilibrium* $(q^*, \lambda^*)$ *is* unstable *if* $\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta)$ *is not negative definite.*

REMARK 27. *By Theorem 20 and Remark 21.1, the equilibria* $(q^*, \beta)$ *of (3.18) where* $\Delta F(q^*, \beta)$ *is negative definite on* $\ker J$ *are local solutions of (3.1). Conversely local solutions* $(q^*, \beta)$ *of (3.1) are equilibria of (3.18) such that* $\Delta F(q^*, \beta)$ *is non-positive definite on* $\ker J$.

By Remark 27, we determine the *bifurcation structure* of equilibria of (3.18), $q^*$, such that $\Delta F(q^*, \beta)$ is non-positive definite on $\ker J$ for each $\beta \in [0, \infty)$. A note of caution is in order: these equilibria need not be stable in the flow (3.18). In fact, $\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta)$ need not be negative definite even when $\Delta F(q^*, \beta^*)$ is negative definite. For example, for the Information Distortion in the case of the Four Blob problem presented in chapter 1,where $N = 4$ and $K = 52$, the $260 \times 260$ Hessian $\Delta_{q,\lambda}\mathcal{L}$ always has at least 52 positive eigenvalues along the solution branch $(q_{\frac{1}{N}}, \beta)$ for every beta.

REMARK 28. *We now point out another reason why we choose to solve (3.1) instead of (1.9). The gradient flow associated with (1.9) may be given as*

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \\ \dot{\xi} \end{pmatrix} = \nabla_{q,\lambda,\xi}\hat{\mathcal{L}}(q, \lambda, \xi, \beta),$$

*where* $\hat{\mathcal{L}}$ *is defined as in (3.13)*

$$\hat{\mathcal{L}}(q, \lambda, \xi, \beta) = F(q, \beta) + \sum_{k=1}^{K} \lambda_k \Big(\sum_{\nu=1}^{N} q_{\nu k} - 1\Big) + \sum_{k=1}^{K}\sum_{\nu=1}^{N} \xi_{\nu k} q_{\nu k}.$$

*There are no equilibria of this system for any* $\beta$ *since if* $\nabla_{q,\lambda,\xi}\hat{\mathcal{L}}(q^*, \lambda^*, \xi^*, \beta) = \mathbf{0}$, *then the equality constraints must be satisfied,* $\nabla_{\lambda}\hat{\mathcal{L}}(q^*, \lambda^*, \xi^*, \beta) = \mathbf{0}$ *(see (3.6)), and all of the inequality constraints are* active*:* $\nabla_{\xi}\hat{\mathcal{L}}(q^*, \lambda^*, \xi^*, \beta) = q^* = \mathbf{0}$. *These conditions clearly cannot both be satisfied. One could instead define the flow*

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda}\hat{\mathcal{L}}(q, \lambda, \xi, \beta). \tag{3.19}$$

*As we point out in (3.14), for an equilibrium* $(q^*, \lambda^*, \xi^*, \beta)$ *of (3.19) in the interior of* $\Delta$,

$$\nabla_{q,\lambda}\hat{\mathcal{L}}(q^*, \lambda^*, \xi, \beta) = \nabla_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta) = \mathbf{0}$$

*if (3.10) holds, which shows that* $(q^*, \lambda^*, \beta)$ *is an equilibrium of (3.18).*

# CHAPTER 4

## KERNEL OF THE HESSIAN

The kernel of $\Delta_{q,\lambda}\mathcal{L}$ plays a pivotal role in the analysis that follows. This is due to the fact that a bifurcation of equilibria of (3.18)

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda}\mathcal{L}(q, \lambda, \beta)$$

at $\beta = \beta^*$ happens when $\ker \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*)$ is nontrivial (Theorem 24). In this chapter, we determine some properties which any vector $k \in \ker \Delta_{q,\lambda}\mathcal{L}$ must satisfy. We then derive a way to evaluate $\det \Delta_{q,\lambda}\mathcal{L}$, which depends only on the blocks $\{B_i\}$ of $\Delta F$. We describe the three types of generic singularities of $\Delta_{q,\lambda}\mathcal{L}$ and $\Delta F$ which can occur, and we also provide an overview of how the singularities of $\Delta_{q,\lambda}\mathcal{L}$ and $\Delta F$ dictate the bifurcation structure of equilibria of (3.18) (Figure 12). We conclude the chapter by analyzing the singularities of $\Delta_{q,\lambda}\mathcal{L}$ and $\Delta F$ for the Information Bottleneck problem (2.35).

### General Form of a Vector in the Kernel

Consider an element $\boldsymbol{k} \in \ker \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*)$. In this section, we determine some properties which any vector $k \in \ker \Delta_{q,\lambda}\mathcal{L}$ must satisfy, which will prove useful in the sequel. Decompose $\boldsymbol{k}$ as

$$\boldsymbol{k} = \begin{pmatrix} \boldsymbol{k}_F \\ \boldsymbol{k}_J \end{pmatrix} \tag{4.1}$$

where $\boldsymbol{k}_F$ is $NK \times 1$ and $\boldsymbol{k}_J$ is $K \times 1$. Hence

$$\begin{aligned} \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta)\boldsymbol{k} = \begin{pmatrix} \Delta F(q^*, \beta^*) & J^T \\ J & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{k}_F \\ \boldsymbol{k}_J \end{pmatrix} = \mathbf{0} \\ \implies \begin{pmatrix} \Delta F(q^*, \beta)\boldsymbol{k}_F + J^T\boldsymbol{k}_J \\ J\boldsymbol{k}_F \end{pmatrix} = \mathbf{0} \qquad (4.2) \\ \implies \Delta F(q^*, \beta)\boldsymbol{k}_F = -J^T\boldsymbol{k}_J \qquad (4.3) \\ J\boldsymbol{k}_F = \mathbf{0} \qquad . \qquad (4.4) \end{aligned}$$

From (3.9), (3.7), and (4.3) we have

$$\begin{pmatrix} B_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & B_2 & \dots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & B_N \end{pmatrix} \boldsymbol{k}_F = - \begin{pmatrix} \boldsymbol{k}_J \\ \boldsymbol{k}_J \\ \vdots \\ \boldsymbol{k}_J \end{pmatrix}. \tag{4.5}$$

We set

$$\boldsymbol{k}_F = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{pmatrix}, \tag{4.6}$$

where $\boldsymbol{x}_i$ is $K \times 1$, so that (4.5) becomes

$$\begin{pmatrix} B_1\boldsymbol{x}_1 \\ B_2\boldsymbol{x}_2 \\ \vdots \\ B_N\boldsymbol{x}_N \end{pmatrix} = - \begin{pmatrix} \boldsymbol{k}_J \\ \boldsymbol{k}_J \\ \vdots \\ \boldsymbol{k}_J \end{pmatrix}. \tag{4.7}$$

From (4.4), $J\boldsymbol{k}_F = \boldsymbol{0}$, and so (3.7) implies that

$$\sum_\nu \boldsymbol{x}_\nu = 0. \tag{4.8}$$

THEOREM 29. *Let $(q^*, \beta^*)$ be a local solution to (3.1) such that $\Delta F(q^*, \beta^*)$ is negative definite on* $\ker J$*, and let $\lambda^*$ be the vector of Lagrange multipliers such that the* KKT *conditions hold (Theorem 16). Then $\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*)$ is nonsingular.*

*Proof.* Let $d$ and $Z$ be defined as in Remark 21.2. Let $\boldsymbol{w} \in \ker J$ which implies $Z\boldsymbol{u} = \boldsymbol{w}$ for some $\boldsymbol{u} \in \Re^d$. Thus

$$\boldsymbol{w}^T \Delta F \boldsymbol{w} = \boldsymbol{u}^T Z^T \Delta F Z \boldsymbol{u} < 0 \text{ for every nontrivial } \boldsymbol{u} \in \Re^d \tag{4.9}$$

by the assumption on $\Delta F(q^*)$. Now let $\boldsymbol{k} \in \ker \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*)$ and decompose it as in (4.1). By (4.4), $\boldsymbol{k}_F \in \ker J$. From (4.3), we see that

$$\boldsymbol{k}_F^T \Delta F \boldsymbol{k}_F = -\boldsymbol{k}_F^T J^T \boldsymbol{k}_J = -(J\boldsymbol{k}_F)^T \boldsymbol{k}_J = \boldsymbol{0}.$$

By (4.9), $\boldsymbol{k}_F = \boldsymbol{0}$. Substituting this into (4.3) shows that $J^T\boldsymbol{k}_J = \boldsymbol{0}$, and so $\boldsymbol{k}_J = \boldsymbol{0}$ since $J^T$ has full column rank (by (3.7)). Therefore $\ker \Delta_{q,\lambda}\mathcal{L} = \{\boldsymbol{0}\}$ and we are done. □

REMARK 30.

1. *The proof to Theorem 29 does not depend on the particular form of the Lagrangian (3.3). The theorem holds for general optimization problems as long as the constraints of the optimization problem are linear (from which it follows that $\Delta F = \Delta_q \mathcal{L}$) and the Jacobian of the constraints has full row rank (assumption of Theorem 20) so that Theorem 20 and Remark 21.2 can be applied.*

2. *The proof to Theorem 29 gives an interesting result. Assuming the hypotheses of the theorem and that $\Delta F$ is negative definite, then (4.7) holds if and only if*

$$\boldsymbol{x}_\nu = B_\nu^{-1}\boldsymbol{k}_J \quad \forall \nu : 1 \leq \nu \leq N.$$

*It follows from (4.8) that $(\sum_\nu B_\nu^{-1})\boldsymbol{k}_J = \mathbf{0}$, which has $\boldsymbol{k}_J = \mathbf{0}$ as the unique solution if and only if $\sum_\nu B_\nu^{-1}$ is nonsingular. Since the proof to the theorem shows the former, then $\sum_\nu B_\nu^{-1}$ must be nonsingular.*

For some equilibria of (3.18) such that $\Delta F(q^*, \beta)$ is negative definite on $\ker J$, Theorem 29 shows a relationship between $\Delta F(q^*, \lambda^*, \beta)$ and $\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta)$: $\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta)$ is nonsingular. In fact, a much more complex relationship is shown later in this chapter.

## Determinant Forms of the Hessian

We now provide explicit forms of the determinant of $\Delta_{q,\lambda}\mathcal{L}$, which, of course, determines whether $\Delta_{q,\lambda}\mathcal{L}$ is singular. The interesting fact is that it depends only on the blocks $\{B_i\}$ of $\Delta F$. In particular, Theorem 33 shows that

$$\det \Delta_{q,\lambda}\mathcal{L} = (-1)^K \det \begin{pmatrix} (B_1 + B_N) & B_N & \dots & B_N & B_N \\ B_N & (B_2 + B_N) & \dots & B_N & B_N \\ B_N & B_N & & B_N & B_N \\ \vdots & \vdots & & \vdots & \vdots \\ B_N & B_N & \dots & B_N & (B_{N-1} + B_N) \end{pmatrix},$$

and Corollary 35 shows that when every block of $\Delta F$ is identically $B$, then

$$\det \Delta_{q,\lambda}\mathcal{L} = (-N)^K (\det B)^{N-1}.$$

Before proving these results, we present the following general theorem.

PROPOSITION 31. *([65] p.250) Let $A$ be a square matrix that can be partitioned as*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

*where $A_{11}$ and $A_{22}$ are square matrices. Then*

$$\det A = \det A_{11} \det(A_{22} - A_{21}A_{11}^{-1}A_{12})$$

*if $A_{11}$ is nonsingular, and*

$$\det A = \det A_{22} \det(A_{11} - A_{12}A_{22}^{-1}A_{21})$$

*if $A_{22}$ is nonsingular.*

An immediate consequence of Proposition 31 is the following theorem.

THEOREM 32. *If $\Delta F$ is nonsingular with blocks $\{B_i\}_{i=1}^N$, then*

$$\det \Delta_{q,\lambda}\mathcal{L} = -\det\left(\sum_i B_i^{-1}\right)\Pi_{i=1}^N \det B_i.$$

*Proof.* By (3.8),

$$\det \Delta_{q,\lambda}\mathcal{L} = \det\begin{pmatrix} \Delta F & J^T \\ J & \mathbf{0} \end{pmatrix}.$$

Applying Proposition 31 with $A_{11} = \Delta F$, we have that

$$\det \Delta_{q,\lambda}\mathcal{L} \;=\; \det \Delta F \det(\mathbf{0} - J\Delta F^{-1}J^T).$$

Since $\Delta F$ is block diagonal as in (3.9), then $\det \Delta F = \Pi_{i=1}^N \det B_i$ and

$$\Delta F^{-1} = \begin{pmatrix} B_1^{-1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & B_2^{-1} & \dots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & B_N^{-1} \end{pmatrix}.$$

This and the fact that $J = \begin{pmatrix} I_K & I_K & \dots & I_K \end{pmatrix}$ (see (3.7)) prove the theorem. □

The following theorem is more general since it does not require the condition that $\Delta F$ be nonsingular.

THEOREM 33.

$$\det \Delta_{q,\lambda}\mathcal{L} = (-1)^K \det\begin{pmatrix} (B_1 + B_N) & B_N & \dots & B_N & B_N \\ B_N & (B_2 + B_N) & \dots & B_N & B_N \\ B_N & B_N & & B_N & B_N \\ \vdots & \vdots & & \vdots & \vdots \\ B_N & B_N & \dots & B_N & (B_{N-1} + B_N) \end{pmatrix}$$

*Proof.* From (3.7), (3.8), and (3.9), we have that the determinant of the $(NK+K)\times(NK+K)$ matrix $\Delta_{q,\lambda}\mathcal{L}$ is given by

$$\det \Delta_{q,\lambda}\mathcal{L} = \det\begin{pmatrix} B_1 & \mathbf{0} & \dots & \mathbf{0} & I_K \\ \mathbf{0} & B_2 & \dots & \mathbf{0} & I_K \\ \mathbf{0} & \mathbf{0} & & \mathbf{0} & I_K \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & B_N & I_K \\ I_K & I_K & \dots & I_K & \mathbf{0} \end{pmatrix}$$

where $\mathbf{0}$ is a $K \times K$ matrices of zeros. Moving the last $K$ rows of the determinant on the right hand side $NK$ rows up gives

$$\det \Delta_{q,\lambda}\mathcal{L} = (-1)^{NK^2} \det \begin{pmatrix} I_K & I_K & \dots & I_K & \mathbf{0} \\ B_1 & \mathbf{0} & \dots & \mathbf{0} & I_K \\ \mathbf{0} & B_2 & \dots & \mathbf{0} & I_K \\ \mathbf{0} & \mathbf{0} & & \mathbf{0} & I_K \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & B_N & I_K \end{pmatrix}.$$

Applying Proposition 31 with $A_{22} = I_K$, we see that the right hand side becomes the determinant of an $NK \times NK$ matrix,

$$\det \Delta_{q,\lambda}\mathcal{L} = (-1)^{NK^2} \det \begin{pmatrix} I_K & I_K & \dots & I_K & I_K \\ B_1 & \mathbf{0} & \dots & \mathbf{0} & -B_N \\ \mathbf{0} & B_2 & \dots & \mathbf{0} & -B_N \\ \mathbf{0} & \mathbf{0} & & \mathbf{0} & -B_N \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & B_{N-1} & -B_N \end{pmatrix}.$$

Moving the first $K$ rows of the determinant on the right hand side $NK - K$ rows down shows that

$$\det \Delta_{q,\lambda}\mathcal{L} = (-1)^{2NK^2-K^2} \det \begin{pmatrix} B_1 & \mathbf{0} & \dots & \mathbf{0} & -B_N \\ \mathbf{0} & B_2 & \dots & \mathbf{0} & -B_N \\ \mathbf{0} & \mathbf{0} & & \mathbf{0} & -B_N \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & B_{N-1} & -B_N \\ I_K & I_K & \dots & I_K & I_K \end{pmatrix}.$$

Now applying Proposition 31 with $A_{22} = I_K$ yields

$$\det \Delta_{q,\lambda}\mathcal{L} = (-1)^{(2N-1)K^2} \det \begin{pmatrix} (B_1 + B_N) & B_N & \dots & B_N & B_N \\ B_N & (B_2 + B_N) & \dots & B_N & B_N \\ B_N & B_N & & B_N & B_N \\ \vdots & \vdots & & \vdots & \vdots \\ B_N & B_N & \dots & B_N & (B_{N-1} + B_N) \end{pmatrix}.$$

Since $2N-1$ is always odd, and $K^2$ is odd if and only if $K$ is odd, then the coefficient $(-1)^{(2N-1)K^2} = (-1)^K$. □

A special case of this result occurs when $\Delta F(q, \beta)$ has $N$ identical blocks, $B_i = B$, for every $i$. We will see in chapter 6 that this occurs if $q$ is fixed by the symmetry defined by the relabelling of the classes of $Y_N$ (Theorem 72). Before we can present the result for this special case, we need the following Lemma.

LEMMA 34. *The $m \times m$ matrix* $\begin{pmatrix} 2 & 1 & \dots & 1 \\ 1 & 2 & \dots & 1 \\ 1 & 1 & & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 2 \end{pmatrix}$ *has determinant equal to $m+1$ and its inverse is the $m \times m$ matrix* $\begin{pmatrix} \frac{m}{m+1} & \frac{-1}{m+1} & \cdots & \frac{-1}{m+1} \\ \frac{-1}{m+1} & \frac{m}{m+1} & \cdots & \frac{-1}{m+1} \\ \frac{-1}{m+1} & \frac{-1}{m+1} & & \frac{-1}{m+1} \\ \vdots & \vdots & & \vdots \\ \frac{-1}{m+1} & \frac{-1}{m+1} & \cdots & \frac{m}{m+1} \end{pmatrix}$.

*Proof.* It is trivial to confirm the inverse. To compute the determinant, we multiply the last row of the matrix by $-1$, then add it to each of the first $m-1$ rows, which shows that

$$\det \begin{pmatrix} 2 & 1 & \dots & 1 \\ 1 & 2 & \dots & 1 \\ 1 & 1 & & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 2 \end{pmatrix} = \det \begin{pmatrix} 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & \dots & 0 & -1 \\ 0 & 0 & & 0 & -1 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -1 \\ 1 & 1 & \dots & 1 & 2 \end{pmatrix}.$$

Multiplying each of the first $m-1$ rows of the determinant on the right by $-1$, and adding it to the last row shows that

$$\det \begin{pmatrix} 2 & 1 & \dots & 1 \\ 1 & 2 & \dots & 1 \\ 1 & 1 & & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 2 \end{pmatrix} = \det \begin{pmatrix} 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & \dots & 0 & -1 \\ 0 & 0 & & 0 & -1 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & \dots & 0 & m+1 \end{pmatrix}.$$

□

COROLLARY 35. *If the blocks, $\{B_i\}_{i=1}^N$, of $\Delta F$ are identical so that $B_i = B$ for every $i$, then* $\det \Delta_{q,\lambda}\mathcal{L} = (-N)^K (\det B)^{N-1}$.

*Proof.* By Theorem 33,

$$\det \Delta_{q,\lambda}\mathcal{L} = (-1)^K \det \begin{pmatrix} 2B & B & \dots & B & B \\ B & 2B & \dots & B & B \\ B & B & & B & B \\ \vdots & \vdots & & \vdots & \vdots \\ B & B & \dots & B & 2B \end{pmatrix}$$

where the matrix on the right is $(NK-K)\times(NK-K)$. Using the Kronecker product, this equation can be rewritten as

$$\det\Delta_{q,\lambda}\mathcal{L} \;=\; (-1)^K\det\left(\begin{pmatrix} 2 & 1 & \dots & 1\\ 1 & 2 & \dots & 1\\ 1 & 1 & & 1\\ \vdots & \vdots & & \vdots\\ 1 & 1 & \dots & 2\end{pmatrix}\otimes B\right).$$

Since the matrix $\begin{pmatrix} 2 & 1 & \dots & 1\\ 1 & 2 & \dots & 1\\ 1 & 1 & & 1\\ \vdots & \vdots & & \vdots\\ 1 & 1 & \dots & 2\end{pmatrix}$ in the last equation is $(N-1)\times(N-1)$, then

$$\det\Delta_{q,\lambda}\mathcal{L} \;=\; (-1)^K(\det B)^{N-1}\det\begin{pmatrix} 2 & 1 & \dots & 1\\ 1 & 2 & \dots & 1\\ 1 & 1 & & 1\\ \vdots & \vdots & & \vdots\\ 1 & 1 & \dots & 2\end{pmatrix}^K.$$

The last equality follows from the fact that if a matrix $A$ is $m\times m$ and a matrix $B$ is $k\times k$, then $\det(A\otimes B)=(\det A)^k(\det B)^m$ ([65] p.256). Now the desired result follows by Lemma 34. □

When $\Delta F$ has $M$ identical blocks which are nonsingular, we can further simplify the determinant given in Theorem 33.

THEOREM 36. *If there exists an $M$ with $1<M<N$ such that $\Delta F$ has $M$ identical blocks, $B$, which are nonsingular, and $N-M$ other blocks, $\{R_i\}_{i-1}^{N-M}$, then* $\det(\Delta_{q,\lambda}\mathcal{L})$ *is equal to*

$$(-M)^K(\det B)^{M-1}\det\begin{pmatrix} (R_1+\frac{1}{M}B) & \frac{1}{M}B & \dots & \frac{1}{M}B & \frac{1}{M}B\\ \frac{1}{M}B & (R_2+\frac{1}{M}B) & \dots & \frac{1}{M}B & \frac{1}{M}B\\ \frac{1}{M}B & \frac{1}{M}B & & \frac{1}{M}B & \frac{1}{M}B\\ \vdots & \vdots & & \vdots & \vdots\\ \frac{1}{M}B & \frac{1}{M}B & \dots & \frac{1}{M}B & (R_{N-M}+\frac{1}{M}B)\end{pmatrix} \tag{4.10}$$

*Proof.* Observe that if $B_N\neq B$, we can perform elementary row and column operations on $\Delta_{q,\lambda}\mathcal{L}$, so that Theorem 33 shows that $\det\Delta_{q,\lambda}\mathcal{L}$ is equal to the determinant

of an $(NK-K)\times(NK-K)$ matrix

$$(-1)^K \det \begin{pmatrix} \begin{pmatrix} (R_1+B) & B & ... & B \\ B & (R_2+B) & ... & B \\ B & B & & B \\ \vdots & \vdots & & \vdots \\ B & B & ... & (R_{N-M}+B) \end{pmatrix} & \mathbf{1}\otimes B \\ \mathbf{1}^T \otimes B & T\otimes B \end{pmatrix}, \quad (4.11)$$

where $\mathbf{1}$ is the $(N-M)\times(M-1)$ matrix of ones and $T$ is the $(M-1)\times(M-1)$ matrix

$$T = \begin{pmatrix} 2 & 1 & ... & 1 \\ 1 & 2 & ... & 1 \\ 1 & 1 & & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & ... & 2 \end{pmatrix}, \text{ with } T^{-1} = \begin{pmatrix} \frac{M-1}{M} & \frac{-1}{M} & ... & \frac{-1}{M} \\ \frac{-1}{M} & \frac{M-1}{M} & ... & \frac{-1}{M} \\ \frac{-1}{M} & \frac{-1}{M} & & \frac{-1}{M} \\ \vdots & \vdots & & \vdots \\ \frac{-1}{M} & \frac{-1}{M} & ... & \frac{M-1}{M} \end{pmatrix},$$

and the inverse is from Lemma 34. We denote the $(N-M)K\times(N-M)K$ matrix in the upper left block of (4.11) by $S$. Now applying Proposition 31 with $A_{22}=T\otimes B$, gives

$$\det \Delta_{q,\lambda}\mathcal{L} = (-1)^K \det(T\otimes B)\det(S-(\mathbf{1}\otimes B)(T\otimes B)^{-1}(\mathbf{1}^T\otimes B)). \quad (4.12)$$

From the proof to Corollary 35, we saw taking determinants of Kronecker products yields $\det(T\otimes B)=(\det T)^K(\det B)^{M-1}$, and so Lemma 34 shows that

$$\det(T\otimes B) = M(\det B)^{M-1}.$$

We proceed by using two more properties of Kronecker products: $(A\otimes B)(C\otimes D)=(AC\otimes BD)$ if the matrices $A,B,C,D$ can be multiplied respectively, and $(A\otimes B)^{-1}=(A^{-1}\otimes B^{-1})$ if $A$ and $B$ are invertible [65]. Thus, (4.12) becomes

$$\begin{aligned} \det \Delta_{q,\lambda}\mathcal{L} &= (-M)^K(\det B)^{M-1}\det(S-(\mathbf{1}\otimes B)(T^{-1}\otimes B^{-1})(\mathbf{1}^T\otimes B)) \\ &= (-M)^K(\det B)^{M-1}\det(S-(\mathbf{1}\otimes B)\left(\frac{1}{M}\mathbf{1}^T\otimes I_K\right)) \\ &= (-M)^K(\det B)^{M-1}\det(S-\left(\frac{M-1}{M}I_{N-M}\otimes B\right)), \end{aligned}$$

which gives the desired result. □

If $\Delta F$ is nonsingular, then its identical blocks must be nonsingular. Thus, Theorem 36 shows that if $\Delta F$ is nonsingular, then $\Delta_{q,\lambda}\mathcal{L}$ is singular if and only if the $(N-M)K\times(N-M)K$ matrix in (4.10) is singular. We wait until chapter 8 to explore this relationship more fully (Theorem 135). We now prove a slightly different version of Theorem 36.

COROLLARY 37. *Let $(q^*, \beta^*)$ be an isolated singularity of $B$ and let $\mathcal{M}(q, \beta)$ be the $(N-M)K \times (N-M)K$ matrix in (4.10) evaluated at $(q, \beta)$. Suppose that there exists an $m > 0$ such that $|\det(\mathcal{M}(q, \beta))| < m$ for all $(q, \beta)$ in some neighborhood about $(q^*, \beta^*)$. Then*

$$\det \Delta_{q,\lambda}\mathcal{L} = (-M)^K (\det B)^{M-1} \det \mathcal{M}(q, \beta)$$

*for all $(q, \beta)$ about $(q^*, \beta^*)$.*

*Proof.* Since $(q^*, \beta^*)$ is an isolated singularity of $B$, then in some neighborhood of $(q^*, \beta^*)$, Theorem 36 shows that,

$$\lim_{(q,\beta)\to(q^*,\beta^*)} |\det \Delta_{q,\lambda}\mathcal{L}| \leq \lim_{(q,\beta)\to(q^*,\beta^*)} mM |\det B(q, \beta)|^{M-1}.$$

Thus, if we define $\det \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*) = 0$, then

$$\det \Delta_{q,\lambda}\mathcal{L} = (-M)^K (\det B)^{M-1} \det \mathcal{M}(q, \beta)$$

for all $(q, \beta)$ in a neighborhood of $(q^*, \beta^*)$, and we can dispense with the assumption in Theorem 36 that $B$ is nonsingular. □

We next give a necessary condition when $\mathcal{M}$, the $(N-M)K \times (N-M)K$ matrix given in (4.10), is singular. This condition is related to a pivotal requirement that we must make in Assumptions 81 in chapter 6.

LEMMA 38. *Suppose that there exists $1 < M < N$ such that $\Delta F$ has $M$ identical blocks, $B$, which are nonsingular, and $N - M$ other blocks, $\{R_i\}_{i-1}^{N-M}$, which are also nonsingular. Then if the matrix $\mathcal{M}$, the $(N-M)K \times (N-M)K$ matrix given in (4.10), is singular, then $B\sum_i R_i^{-1} + MI_K$ is singular.*

*Proof.* Let $\boldsymbol{u} \in \ker \mathcal{M}$ and decompose it as

$$\boldsymbol{u} = \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_{N-M} \end{pmatrix}$$

where $\boldsymbol{u}_i$ is $K \times 1$ for every $i$. Then the equation $S\boldsymbol{u} = \boldsymbol{0}$ can be rewritten as the system of equations

$$\begin{aligned} R_1\boldsymbol{u}_1 + \frac{1}{M}\sum_{i=1}^{M} B\boldsymbol{u}_i &= \boldsymbol{0} \\ R_2\boldsymbol{u}_2 + \frac{1}{M}\sum_{i=1}^{M} B\boldsymbol{u}_i &= \boldsymbol{0} \\ &\vdots \\ R_{N-M}\boldsymbol{u}_{N-M} + \frac{1}{M}\sum_{i=1}^{M} B\boldsymbol{u}_i &= \boldsymbol{0}. \end{aligned}$$

Thus,

$$\boldsymbol{u}_j = -\frac{1}{M} R_j^{-1} B \sum_i \boldsymbol{u}_i$$

from which it follows that

$$\sum_j \boldsymbol{u}_j = -\frac{1}{M} \sum_j R_j^{-1} B \sum_i \boldsymbol{u}_i.$$

The substitution $\boldsymbol{v} = \sum_i \boldsymbol{u}_i$ shows that

$$(\sum_j R_j^{-1} B + M I_K)\boldsymbol{v} = \boldsymbol{0}.$$

We observe that since $B$ is nonsingular, then multiplying this equation on the right by $B^{-1}$ and on the left by $B$ completes the proof. □

The converse of this lemma holds as well, which we will prove in chapter 8 (Theorem 135). For now, we state the result.

THEOREM 39. *Suppose that $\Delta F$ is nonsingular. Then $\Delta_{q,\lambda}\mathcal{L}$ is singular if and only if $B\sum_\nu R_\nu^{-1} + M I_K$ is singular.*

## Generic Singularities

In this chapter, we have considered the case where $\Delta F$ has $M > 1$ blocks that are identical. As we have seen in the last section, these identical blocks can simplify the form of the determinant of $\Delta_{q,\lambda}\mathcal{L}$. In fact, much more is true. In this section we show that, generically, there are three types of singularities of $\Delta_{q,\lambda}\mathcal{L}$ which can occur, one of which gives rise to the symmetry breaking bifurcations we will study in chapter 6, and another which gives rise to the saddle-node bifurcations which we study in chapter 8.

First, we introduce some terminology. We will call the classes of $Y_N$ which correspond to the identical blocks of $\Delta F$ *unresolved* classes. The classes of $Y_N$ which are not unresolved will be called *resolved* classes (this terminology is consistent with Definition 69 in chapter 6). We now partition the set $\mathcal{Y}_N$ into two disjoint sets. Let

$$\mathcal{U} \text{ be the set of } M \text{ unresolved classes}$$

and let

$$\mathcal{R} \text{ be the set of } N - M \text{ resolved classes.}$$

Thus $\mathcal{U} \cap \mathcal{R} = \emptyset$ and $\mathcal{U} \cup \mathcal{R} = \{1, ..., N\} = \mathcal{Y}_N$.

Let $B_\nu$ be the block of $\Delta F$ corresponding to class $\nu$. For clarity, we denote

$$B = B_\nu \text{ for } \nu \in \mathcal{U}$$

and

$$R_\nu = B_\nu \text{ for } \nu \in \mathcal{R}.$$

Now we define *genericity*.

DEFINITION 40. *Let $\mathcal{T}$ be a topological space. A set $\mathcal{W} \subseteq \mathcal{T}$ is* generic *if $\mathcal{W}$ is open and dense in $\mathcal{T}$.*

REMARK 41. *Let $\Delta F^\nu(q,\beta)$ denote the $\nu^{th}$ block of the Hessian $\Delta F(q,\beta)$. Consider the class $\mathcal{T}_\mathcal{U}$ of* singular *$NK \times NK$ block diagonal matrices of the form*

$$\Delta F^\nu(q,\beta) = \begin{cases} B(q,\beta) & \text{if } \nu \in \mathcal{U} \\ R_\nu(q,\beta) & \text{otherwise (i.e. if } \nu \in \mathcal{R}) \end{cases}$$

*over all $(q,\beta) \in \Delta \times \Re$. Let $\mathcal{W} \subseteq \mathcal{T}_\mathcal{U}$ such that a matrix $\Delta F \in \mathcal{W}$ if and only if at most one of the matrices $B$, $\{R_\nu\}$, and $B\sum_\nu R_\nu^{-1} + MI_K$ is singular. We assume that $\mathcal{W}$ is generic in $\mathcal{T}_\mathcal{U}$. Thus, by generic, we mean that only one of the matrices $B$, $\{R_\nu\}_{\nu\in\mathcal{R}}$, or $B\sum_\nu R_\nu^{-1} + MI_K$ is singular at a given point $(q,\beta) \in \Delta \times \Re$.*

We are now ready to discuss the three types of generic singularities, which we have depicted in Figure 12. We will cite the relevant results in the text which support these claims.

The first type of singularity is when the $M$ unresolved blocks of $\Delta F$ are singular. A generic assumption in this instance is that the $N - M$ resolved blocks, $\{R_\nu\}$, are nonsingular at $(q^*,\beta)$. By Corollary 89, $\Delta_{q,\lambda}\mathcal{L}$ must be singular. Conversely, suppose that $\Delta_{q,\lambda}\mathcal{L}$ is singular. Generically, the resolved blocks of $\Delta F$ are nonsingular, and $B\sum_\nu R_\nu^{-1} + MI_K$ is nonsingular. Then Corollary 89 shows that $\Delta F$ is singular. We will see in chapter 6 that this is the type of singularity that exhibits symmetry breaking bifurcation (Theorem 110).

The second type of singularity is a special case in which no bifurcation occurs. If only a single block, $R_\nu$, of $\Delta F$ is singular, and if the generic condition that $B\sum_\nu R_\nu^{-1} + MI_K$ is nonsingular holds, then we will show in chapter 6 (Theorem 114) that $\Delta_{q,\lambda}\mathcal{L}$ is nonsingular. Thus, generically, no bifurcation occurs for this case.

The third type of singularity is when $\Delta_{q,\lambda}\mathcal{L}$ is singular, but when $\Delta F$ is nonsingular. By Theorem 39, it must be that $B\sum_\nu R_\nu^{-1} + MI_K$ is singular. This singularity type manifests itself as saddle-node bifurcations in the numerical results of chapter 7. In chapter 8 (Theorem 138), we prove that $\Delta F$ is generically nonsingular at any bifurcation that is not a symmetry breaking bifurcation, which includes saddle-node bifurcations. Observe that if $\Delta F$ were singular, then, generically, we would be in one of the first two cases of singularity just described.

Figure 12, which summarizes the preceding discussion, indicates how the singularities of $\Delta_{q,\lambda}\mathcal{L}$ and $\Delta F$ affect the bifurcation structure of equilibria of (3.18). At the top of the diagram, we have the assumption that $\Delta_{q,\lambda}\mathcal{L}$ is singular, which is a

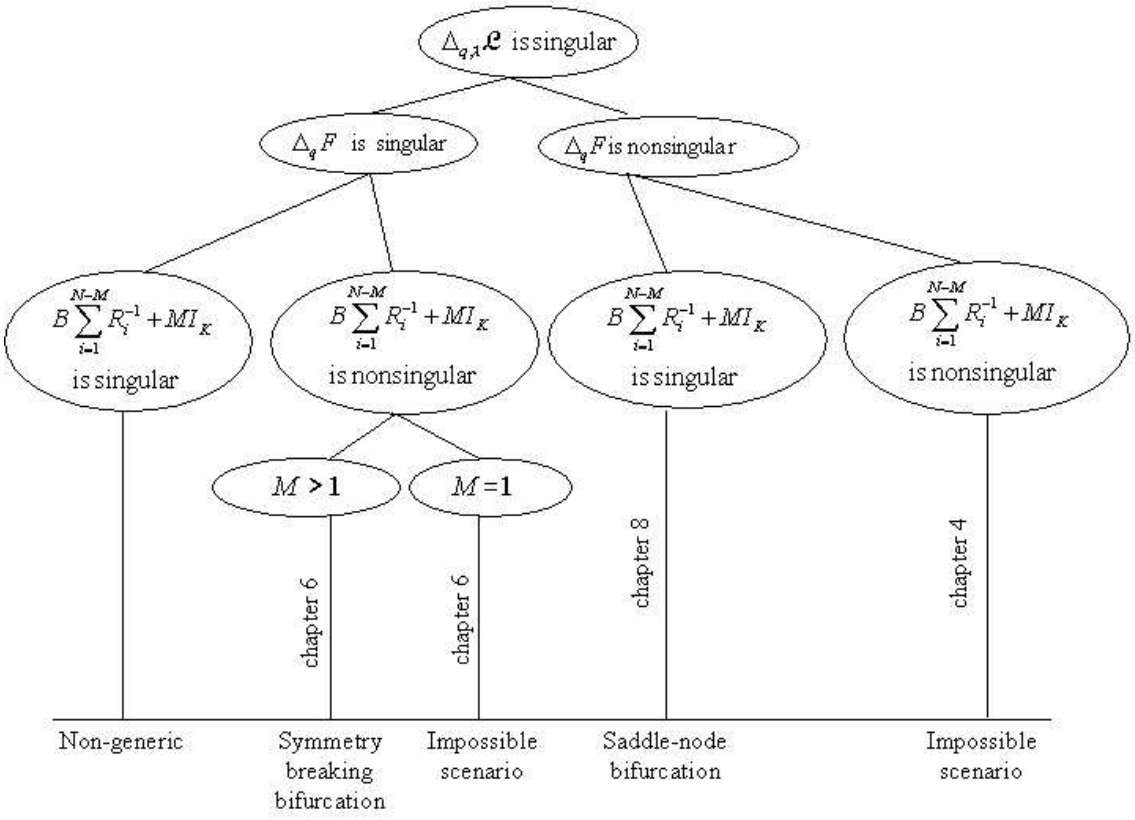



Figure 12. A hierarchical diagram showing how the singularities of $\Delta_{q,\lambda}\mathcal{L}$ and $\Delta F$ affect the bifurcation structure of equilibria of (3.18).

necessary condition given that a bifurcation occurs (Theorem 24). To proceed to the second level of the of the diagram, one must further assume that either $\Delta F$ is singular or nonsingular. To get to the third level, one must add to the list of assumptions that either $B\sum_i R_i^{-1} + MI_K$ is either singular or nonsingular. At the base level of the diagram, we have indicated the type of bifurcation possible given the assumptions on $\Delta_{q,\lambda}\mathcal{L}$ and $\Delta F$ above. We have indicated the chapter which justifies the different conclusions. In particular, see Theorem 36 and Lemma 38 in chapter 4; see Corollary 111 and Theorems 110 and 114 in chapter 6; and see Theorems 135 and 141 in chapter 8.

## Singularities of the Information Bottleneck

For the Information Bottleneck problem (2.35),

$$\max_{q\in\Delta} F_B(q,\beta) = \max_{q\in\Delta}(-I(Y;Y_N) + \beta I(X,Y_N)),$$

the $NK \times 1$ vector $q$ is always in the kernel of $\Delta F_B(q, \beta)$ for every value of $\beta$ (Lemma 42). This implies, for example, that the $K \times 1$ vector of $\frac{1}{N}$'s is in the kernel of each block of $\Delta F_B(q_{\frac{1}{N}}, \beta)$, for every $\beta$. We prove this observation in this section, which shows that $\Delta F_B$ is highly degenerate (Theorem 43).

First, we need to compute the quantities $\Delta I(Y, Y_N)$ and $\Delta I(X, Y_N)$. The second quantity was computed in (2.23). To compute the first quantity, we notice that [17]

$$-I(Y; Y_N) = H(Y_N|Y) - H(Y_N). \tag{4.13}$$

Since we know the Hessian of the first term (2.20), we only need to compute $\Delta H(Y_N)$. By definition

$$-H(Y_N) = \sum_{\mu \in \mathcal{Y}_N} p(\mu) \log p(\mu).$$

Using the fact that $\frac{\partial p(\mu)}{\partial q_{\nu k}} = \delta_{\mu\nu} p(y_k)$, the gradient of $H(Y_N)$ is

$$\begin{aligned}
(-\nabla H(Y_N))_{\nu k} &\equiv -\frac{\partial H(Y_N)}{\partial q_{\nu k}} \\
&= \frac{\partial}{\partial q_{\nu k}} \sum_{\mu \in \mathcal{Y}_N} p(\mu) \log p(\mu) \\
&= \sum_{\mu} \delta_{\nu\mu} p(y_k) \log p(\mu) + p(\mu) \frac{\delta_{\nu\mu} p(y_k)}{(\ln 2) p(\mu)} \\
&= p(y_k) \left( \log p(\nu) + \frac{1}{\ln 2} \right).
\end{aligned}$$

Thus, the Hessian is given by

$$\begin{aligned}
\frac{-\partial^2 H(Y_N)}{\partial q_{\eta l} \partial q_{\nu k}} &= \frac{\partial}{\partial q_{\eta l}} p(y_k) \left( \log p(\nu) + \frac{1}{\ln 2} \right) \\
&= p(y_k) \frac{\delta_{\nu\eta} p(y_l)}{(\ln 2) p(\nu)}.
\end{aligned}$$

From this calculation, (4.13) and (2.20), we get

$$\frac{-\partial^2 I(Y; Y_N)}{\partial q_{\eta l} \partial q_{\nu k}} = \frac{\delta_{\nu\eta}}{\ln 2} \left( \frac{p(y_k) p(y_l)}{p(\nu)} - \frac{\delta_{lk} p(y_k)}{q_{\nu k}} \right). \tag{4.14}$$

Equation (4.14) shows that $\delta_{\nu\eta}$ can be factored out of

$$\Delta F_B = -\Delta I(Y; Y_N) + \beta \Delta I(X; Y_N). \tag{4.15}$$

This implies that $\Delta F_B$ is block diagonal, with each block corresponding to a particular class of $Y_N$.

Before proving the main theorem, we first show that each block of $\Delta F_B$ is singular.

LEMMA 42. *Fix an arbitrary quantizer $q$ and arbitrary class $\nu$. Then the vector $q^\nu$ is in the kernel of the $\nu^{th}$ block of $\Delta F_B(q,\beta)$ for each value of $\beta$.*

*Proof.* To show that the vector $q^\nu$ is in the kernel of $\Delta F_B^\nu(q,\beta)$, the $\nu^{th}$-block of $\Delta F_B(q)$, we compute the $l^{th}$ row of this matrix. From (4.15), (4.14), and (2.23), we see that

$$\begin{aligned}
[\Delta F_B^\nu(q)q^\nu]_l &= \frac{1}{\ln 2}\left(\sum_k \frac{p(y_l)p(y_k)q_{\nu k}}{p(\nu)} - \sum_k \delta_{lk}\frac{q_{\nu k}p(y_k)}{q_{\nu k}}\right) \\
&\quad + \frac{\beta}{\ln 2}\sum_k\left(\sum_i \frac{p(x_i,y_k)p(x_i,y_l)q_{\nu k}}{p(x_i,\nu)} - \frac{p(y_k)p(y_l)q_{\nu k}}{p(\nu)}\right) \\
&= \frac{1}{\ln 2}(p(y_l)-p(y_l)) + \frac{\beta}{\ln 2}(\sum_i \frac{p(x_i,y_l)}{p(x_i,\nu)}\sum_k q_{\nu k}p(y_k,x_i) \\
&\quad - \frac{p(y_l)}{p(\nu)}\sum_k q_{\nu k}p(y_k)) \\
&= \frac{\beta}{\ln 2}\left(\sum_i p(x_i,y_l) - p(y_l)\right) \\
&= \mathbf{0}.
\end{aligned}$$

This shows that $q^\nu$ is in the kernel of the $\nu^{th}$ block $\Delta F_B$.

THEOREM 43. *For an arbitrary pair $(q,\beta)$, the dimension of the kernel of the matrix $\Delta F_B$ is at least $N$.*

*Proof.* Define the vectors $\{\boldsymbol{v}_i\}_{i=1}^N$ by

$$\boldsymbol{v}_1 = \begin{pmatrix} q^1 \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \boldsymbol{v}_2 = \begin{pmatrix} \mathbf{0} \\ q^2 \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \ldots, \boldsymbol{v}_N = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ q^N \end{pmatrix}.$$

By Lemma 42, $\{\boldsymbol{v}_i\}_{i=1}^N$ are in $\ker \Delta F_B(q,\beta)$. Clearly, these vectors are linearly independent. □

# CHAPTER 5

## GENERAL BIFURCATION THEORY WITH SYMMETRIES

This chapter introduces the rudiments of bifurcation theory in the presence of symmetries, which includes the Equivariant Branching Lemma (Theorem 47) and the Smoller-Wasserman Theorem (Theorem 49). This theory shows the existence of branches from symmetry breaking bifurcation of equilibria of systems such as (3.15)

$$\dot{\boldsymbol{x}} = \psi(\boldsymbol{x}, \beta)$$

which have symmetry. We obtain results which can ascertain the structure of the bifurcating branches. These results enable us to answer questions about equilibria of (3.15) such as: Are symmetry breaking bifurcations pitchfork-like or transcritical? Are the bifurcating branches subcritical or supercritical? Are the bifurcating branches stable or unstable?

In order to apply the bifurcation theory to a system such as (3.15) in the presence of symmetries, it is first necessary to determine the Liapunov-Schmidt reduction, $\phi(\boldsymbol{w}, \beta)$, of the system. We present the mechanics of this reduction, as well as the symmetries of the reduction.

This theory is required so that later, in chapter 6, we may show the bifurcation structure of equilibria of the gradient flow (3.18)

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda} \mathcal{L}(q, \lambda, \beta),$$

which we introduced in chapter 3. This will yield information about solutions to the constrained optimization problem (1.9)

$$\max_{q \in \Delta} (G(q) + \beta D(q)).$$

We begin by introducing the required terminology and some preliminary results which will prove useful in the sequel. Let

$$\dot{\boldsymbol{w}} = \phi(\boldsymbol{w}, \beta) \tag{5.1}$$

where $\boldsymbol{w}$ is in some Banach space $V$ and $\beta \in \Re$, so that

$$\phi : V \times \Re \to V.$$

Let $G$ be a compact Lie Group acting on $V$. The vector valued function $\phi$ is *$G$-invariant* if

$$\phi(g\boldsymbol{w}) = \phi(\boldsymbol{w})$$

for every $\boldsymbol{w} \in V$ and every $g \in G$. The function $\phi$ is *G-equivariant* if

$$\phi(g\boldsymbol{w}) = g\phi(\boldsymbol{w})$$

for every $\boldsymbol{w} \in V$ and every $g \in G$. Let $H \leq G$ and let $W$ be a subspace of $V$. For the vectors $\boldsymbol{w} \in V$ such that $\phi(\boldsymbol{w}) = \mathbf{0}$, the amount of symmetry present in $\boldsymbol{w}$ is measured by its *isotropy subgroup*

$$H = H_{\boldsymbol{w}} = \{h \in G | h\boldsymbol{w} = \boldsymbol{w}\}.$$

An isotropy subgroup of $H < G$ is a *maximal isotropy subgroup* if there does not exist any isotropy subgroup $K < G$ that contains $H$,

$$H < K < G.$$

The *fixed point space* of any subgroup $H \leq G$ is

$$\text{Fix}(H) = \{\boldsymbol{v} \in V | h\boldsymbol{v} = \boldsymbol{v} \text{ for every } h \in H\}.$$

The subspace $W$ is *G-invariant* if $g\boldsymbol{w} \in W$ for all $\boldsymbol{w} \in W$. The subspace $W$ is *H-irreducible* if the only $H$-invariant subspaces of $W$ are $\{\mathbf{0}\}$ and $W$. The action of the group $G$ on $V$ is *absolutely irreducible* if the only linear mappings on $V$ that commute with every $g \in G$ are scalar multiples of the identity.

The following results will prove useful in the sequel.

LEMMA 44. *([34] p.74) Let $\phi : V \times \Re \to V$ be a G-equivariant function for some Banach space $V$ and let $H \leq G$. Then*

$$\phi(\text{Fix}(H) \times \Re) \subseteq \text{Fix}(H).$$

PROPOSITION 45. *([34] p.75) Let $G$ be a compact Lie group acting on a Banach space $V$. The following are equivalent:*

1. $\text{Fix}(G) = \{\mathbf{0}\}$.
2. *Every G-equivariant map $\phi : V \times \Re \to V$ satisfies $\phi(\mathbf{0}, \beta) = \mathbf{0}$ for all $\beta$.*
3. *The only G-equivariant linear function is the zero function.*

PROPOSITION 46. *Let $G$ be a compact Lie group such that $\phi : V \times \Re \to V$ is G-equivariant. Further suppose that $\phi(\mathbf{0}, 0) = \mathbf{0}$, and that $\partial_{\boldsymbol{w}}\phi(\mathbf{0}, 0)$ is singular. Then*

1. *([33] p.304) The Jacobian $\partial_{\boldsymbol{w}}\phi(\mathbf{0}, \beta)$ commutes with every $g \in G$.*
2. *([34] p.82 or [33] p. 304) The spaces* $\ker \partial_{\boldsymbol{w}}\phi(\mathbf{0}, 0)$ *and* $\text{range } \partial_{\boldsymbol{w}}\phi(\mathbf{0}, 0)$ *are each G-invariant.*

3. *([34] p.69) Let $g \in G$. The vector $\boldsymbol{w} \in V$ has isotropy subgroup $H \leq G$ if and only if $g\boldsymbol{w}$ has isotropy subgroup $gHg^{-1} \leq G$.*

4. *(Trace Formula) ([34] p.76) Let $H \leq G$ where $|H| < \infty$. Then*

$$\dim \mathrm{Fix}(H) = \frac{1}{|H|} \sum_{h \in H} \mathrm{tr}(h).$$

5. *([34] p.40) If the action of $G$ on a vector space $V$ is absolutely irreducible then $V$ is $G$-irreducible.*

6. *If $V$ is $G$-irreducible with $\dim(V) \geq 1$, then $\mathrm{Fix}(G) = \{\mathbf{0}\}$.*

*Proof.* We prove 1, 2, and 6. Let

$$\Phi := \partial_{\boldsymbol{w}} \phi(\mathbf{0}, 0).$$

For $g \in G$, we have $\phi(g\boldsymbol{w}, \beta) = g\phi(\boldsymbol{w}, \beta)$, giving $\partial_{\boldsymbol{w}} \phi(g\boldsymbol{w}, \beta) g = g \partial_{\boldsymbol{w}} \phi(\boldsymbol{w}, \beta)$. Evaluating at $(\mathbf{0}, 0)$ gives

$$\begin{aligned} &\partial_{\boldsymbol{w}} \phi(\mathbf{0}, 0) g = g \partial_{\boldsymbol{w}} \phi(\mathbf{0}, 0) \\ \implies &\ g \text{ commutes with } \Phi = \partial_{\boldsymbol{w}} \phi(\mathbf{0}, 0). \end{aligned}$$

This proves 1.

If $\boldsymbol{k} \in \ker \Phi$ then $\Phi g \boldsymbol{k} = g \Phi \boldsymbol{k} = g\mathbf{0} = \mathbf{0}$. Furthermore, if $\boldsymbol{r} \in \mathrm{range}\Phi$, then there exists $\boldsymbol{w} \in B_2$ such that $\Phi \boldsymbol{w} = \boldsymbol{r}$. Then $g\boldsymbol{r} = g\Phi\boldsymbol{w} = \Phi g \boldsymbol{w}$ from which it follows that $g\boldsymbol{r} \in \mathrm{range}\Phi$. This proves 2.

To prove 6, we show the contrapositive. Suppose that $\mathrm{Fix}(G) \neq \{\mathbf{0}\}$. Then $g\boldsymbol{v} = \boldsymbol{v}$ for some $\boldsymbol{v} \in V$, which implies that $\mathrm{span}(\boldsymbol{v})$ is an invariant subspace of $V$. Thus, $V$ is not irreducible. □

## Existence Theorems for Bifurcating Branches

We are interested in bifurcations of equilibria of the dynamical system (5.1),

$$\dot{\boldsymbol{w}} = \phi(\boldsymbol{w}, \beta),$$

where $\phi : V \times \Re \to V$ for some Banach space $V$. If $\phi$ is $G$-equivariant for some compact Lie group $G$, then the next three theorems are the main results which relate the subgroup structure of $G$ with the existence of bifurcating branches of equilibria of (5.1). We first introduce the theorem attributed to Vanderbauwhede [82] and Cicogna [12, 13].

Theorem 47 (Equivariant Branching Lemma). *([34] p.83) Assume that*

1. *The sufficiently smooth function* $\phi : V \times \Re \to V$ *from (5.1) is* $G$ *equivariant for a compact Lie group* $G$*, and a Banach space* $V$*.*

2. *The Jacobian* $\partial_{\boldsymbol{w}}\phi(\mathbf{0}, 0) = \mathbf{0}$*.*

3. *The group* $G$ *acts absolutely irreducibly on* $\ker \partial_{\boldsymbol{w}}\phi(\mathbf{0}, 0)$ *so that* $\partial_{\boldsymbol{w}}\phi(\mathbf{0}, \beta) = c(\beta)I$ *for some scalar valued function* $c(\beta)$*.*

4. *The scalar function* $c'(0) \neq 0$*.*

5. *The subgroup* $H$ *is an isotropy subgroup of* $G$ *with* $\dim \mathrm{Fix}(H) = 1$*.*

*Then there exists a unique smooth solution branch* $(t\boldsymbol{w}_0, \beta(t))$ *to* $\phi = \mathbf{0}$ *such that* $\boldsymbol{w}_0 \in \mathrm{Fix}(H)$*, and the isotropy subgroup of each solution is* $H$*.*

*Proof.* Let $\hat{\phi} := \phi|\mathrm{Fix}(H) \times \Re$ and let $\boldsymbol{w}_0 \in \mathrm{Fix}(H)$. By Lemma 44

$$\hat{\phi} : \mathrm{Fix}(H) \times \Re \to \mathrm{Fix}(H) \tag{5.2}$$

and so $\dim \mathrm{Fix}(H) = 1$ implies that

$$\hat{\phi}(\boldsymbol{w}, \beta) = \phi(t\boldsymbol{w}_0, \beta) = h(t, \beta)\boldsymbol{w}_0$$

for some scalar function $h(t, \beta)$. Since $G$ acts absolutely irreducibly on $\ker \partial_{\boldsymbol{w}}\phi$, then $\mathrm{Fix}(G) = \{\mathbf{0}\}$ (Proposition 46.6) which implies

$$\phi(\mathbf{0}, \beta) = \mathbf{0} \tag{5.3}$$

by Proposition 45. Hence, $h(0, \beta) = 0$. Therefore, the Taylor series for $h$ is

$$\begin{aligned} h(t, \beta) &= h'(0, \beta)t + \frac{h''(0, \beta)}{2}t^2 + ... \\ &= tk(t, \beta) \end{aligned}$$

where

$$k(t, \beta) := \sum_{n=1}^{\infty} \frac{\partial^n h(0, \beta)}{n!} t^{n-1} \tag{5.4}$$

and the $n^{th}$ derivative $\partial^n h(0, \beta)$ is with respect to $t$. Hence

$$\hat{\phi}(\boldsymbol{w}, \beta) = \phi(t\boldsymbol{w}_0, \beta) = tk(t, \beta)\boldsymbol{w}_0. \tag{5.5}$$

Differentiating this equation yields

$$\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0, \beta)\boldsymbol{w}_0 = (k(t, \beta) + t\partial_t k(t, \beta))\boldsymbol{w}_0 \tag{5.6}$$

and so

$$k(t,\beta)\boldsymbol{w}_0 = \partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta)\boldsymbol{w}_0 - t\partial_t k(t,\beta)\boldsymbol{w}_0 \tag{5.7}$$

from which it follows that

$$k(0,0) = 0 \tag{5.8}$$

since $\partial_{\boldsymbol{w}}\phi(\mathbf{0},0) = \mathbf{0}$ by assumption. From (5.7) we compute

$$\partial_\beta k(t,\beta)\boldsymbol{w}_0 = \partial_\beta\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta)\boldsymbol{w}_0 - t\partial_\beta\partial_t k(t,\beta)\boldsymbol{w}_0. \tag{5.9}$$

Thus

$$\partial_\beta k(0,0)\boldsymbol{w}_0 = \partial_\beta\partial_{\boldsymbol{w}}\phi(0,0)\boldsymbol{w}_0.$$

Now, the absolute irreducibility of $G$ on $\ker\partial_{\boldsymbol{w}}\phi(\mathbf{0},0)$ shows that

$$\partial_\beta k(0,0) = c'(0). \tag{5.10}$$

By assumption, $c'(0) \neq 0$ giving

$$\partial_\beta k(0,0) \neq 0. \tag{5.11}$$

By (5.8) and (5.11), the Implicit Function Theorem can be applied to solve

$$k(t,\beta) = 0 \tag{5.12}$$

uniquely for $\beta = \beta(t)$ in $\mathrm{Fix}(H)$, which shows that $(t\boldsymbol{w}_0, \beta(t))$ is a bifurcating solution from $(0,0)$ of $\phi(\boldsymbol{w},\beta) = \mathbf{0}$.

By assumption, $\boldsymbol{w}_0 \in \mathrm{Fix}(H)$, from which it follows that the isotropy group of the branch $(t\boldsymbol{w}_0, \beta(t))$ is $H$. □

Cicogna [12, 13, 14] has generalized the Equivariant Branching Lemma to show the existence of bifurcating branches for every maximal isotropy subgroup where the dimension of the fixed point space is odd.

We now present the theorem which deals with dynamical systems (5.1) that are gradient flows, such as (3.18), where

$$\dot{\boldsymbol{w}} = \phi(\boldsymbol{w},\beta) = \nabla_{\boldsymbol{w}} f(\boldsymbol{w},\beta).$$

First we present the theorem as posed by Smoller and Wasserman in [71]. We restate a weaker form of this result in Theorem 49, which presupposes a bifurcation point, so that the eigenvalue crossing condition is not required.

THEOREM 48. *([71] p.85) Let $G$ be a compact Lie group. Assume the following:*

1. *Let $B_2$ and $B_0$ be Banach spaces, and let $\mathcal{H}$ be a $G$-invariant Hilbert space, such that*

   $$B_2 \subseteq B_0 \subseteq \mathcal{H},$$

   *where the embeddings are all continuous.*

2. *There is a twice continuously differentiable function $f$ on $B_2 \times \Re$,*

$$\nabla_{\boldsymbol{w}} f : B_2 \times \Re \to B_0,$$

*such that $\nabla_{\boldsymbol{w}} f$ is $G$-equivariant.*

3. *The equation $\nabla_{\boldsymbol{w}} f(\mathbf{0}, \beta) = \mathbf{0}$ holds for every $\beta \in I$ where $I$ is some interval in $\Re$.*

4. *The matrices $\Delta_{\boldsymbol{w}} f(\mathbf{0}, \beta_1)$ and $\Delta_{\boldsymbol{w}} f(0, \beta_2)$ are nonsingular for some $\beta_1, \beta_2 \in I$.*

5. *The compact Lie group $G$ acts on $\boldsymbol{w} \in B_2$ such that the only $G$-invariant solution of $\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \beta) = \mathbf{0}$ is $(\mathbf{0}, \beta)$ for every $\beta \in I$.*

6. *The kernel $\ker \Delta_{\boldsymbol{w}} f(\mathbf{0}, \beta)$ contains no nontrivial $G$-invariant subspaces.*

7. *There exists subgroups $H, L < G$ such that*

$$\dim(\text{peigenspace}(\Delta_{\boldsymbol{w}} f(0, \beta_1)) \cap \text{Fix}(H)) \neq \dim(\text{peigenspace}(\Delta_{\boldsymbol{w}} f(0, \beta_2)) \cap \text{Fix}(H)),$$

*and that*

$$\dim(\text{peigenspace}(\Delta_{\boldsymbol{w}} f(0, \beta_1)) \cap \text{Fix}(L)) \neq \dim(\text{peigenspace}(\Delta_{\boldsymbol{w}} f(0, \beta_2)) \cap \text{Fix}(L)).$$

8. *The group generated by $H$ and $L$, $HL$, is the full group, $HL = G$.*

*Then there exists $\beta_H, \beta_L \in (\beta_1, \beta_2)$ such that the solutions $(\boldsymbol{w} = \mathbf{0}, \beta_H)$ and $(\boldsymbol{w} = \mathbf{0}, \beta_L)$ are bifurcation points of solutions with isotropy groups $H$ and $L$ respectively. The bifurcating solutions do not coincide.*

The condition on the dimensionality of the peigenspaces in Theorem 48 assures that an eigenvalue of $\partial_{\boldsymbol{w}} \phi(\mathbf{0}, \beta)$ changes sign for some $\beta^*$ in the interval $I \subset \Re$, which guarantees that bifurcation occurs at $\beta = \beta^*$. If we assume a priori that bifurcation occurs at $(\mathbf{0}, \beta^*)$, then we may dispense with the assumption on the peigenspaces, as well as the assumption that $\partial_{\boldsymbol{w}} \phi(\mathbf{0}, \beta)$ is nonsingular at $\beta = \beta_1$ and at $\beta = \beta_2$.

The condition that the group, $HL$, generated by the subgroups $H, L < G$, be equal to the full group $G$, is satisfied if we require that $H$ and $L$ are maximal isotropy subgroups ([34] p.138).

Using these observations, as well as the terminology which we have developed thus far, we have the following theorem.

THEOREM 49 (SMOLLER-WASSERMAN THEOREM). *([71] p.85, [33] p.138) Let $G$ be a compact Lie group. Assume the following:*

1. *Let $B_2$ and $B_0$ be Banach spaces, and let $\mathcal{H}$ be a $G$-invariant Hilbert space, such that*

$$B_2 \subseteq B_0 \subseteq \mathcal{H},$$

*where the embeddings are all continuous.*

2. *There is a twice continuously differentiable function* $f$ *on* $B_2 \times \Re$,

$$\nabla_{\boldsymbol{w}} f : B_2 \times \Re \to B_0,$$

*such that* $\nabla_{\boldsymbol{w}} f$ *is* $G$*-equivariant.*

3. *The equation* $\nabla_{\boldsymbol{w}} f(\mathbf{0}, \beta) = \mathbf{0}$ *holds for every* $\beta \in I$ *where* $I$ *is some interval in* $\Re$.

4. *Bifurcation of solutions to* $\nabla_{\boldsymbol{w}} f(\mathbf{0}, \beta) = \mathbf{0}$ *occurs at* $\beta = \beta^*$.

5. *The fixed point space* $\mathrm{Fix}(G) = \{\mathbf{0}\}$.

6. *The kernel* $\ker \Delta_{\boldsymbol{w}} f(\mathbf{0}, \beta)$ *is* $G$*-irreducible.*

7. *Let* $H$ *be a maximal isotropy subgroup of* $G$.

*Then there exists bifurcating solutions to*

$$\nabla_{\boldsymbol{w}} f(\mathbf{0}, \beta) = \mathbf{0}$$

*with isotropy subgroup* $H$.

The advantage of using the Smoller-Wasserman Theorem over the Equivariant Branching Lemma for a gradient system such as (3.18) is that we get the existence of bifurcating branches for each and every maximal isotropy subgroup, not merely the ones where the dimension of the fixed point space of the isotropy group is 1.

## Bifurcation Structure

In this section, the bifurcation structure of the solution branches $(\boldsymbol{w}^*, \beta^*)$ to (5.1),

$$\phi(\boldsymbol{w}, \beta) = \mathbf{0},$$

whose existence is guaranteed by the Equivariant Branching Lemma, is considered. The independent variable $\boldsymbol{w}$ is in some Banach space $V$ and $\beta \in \Re$, so that

$$\phi : V \times \Re \to V. \tag{5.13}$$

We explicitly derive a condition (Lemma 53) which determines whether a bifurcation is pitchfork-like or transcritical.

In the transcritical case, we present the results of Golubitsky [34] which ascertain whether bifurcating branches are subcritical or supercritical (Remarks 54.1 and 54.3). In the transcritical case, bifurcating branches are always unstable (Proposition 58 and Theorem 60).

To determine whether bifurcating branches are subcritical or supercritical when the bifurcation is pitchfork-like, we have further developed the theory of Golubitsky (Remark 54.4 and Lemma 63). Subcritical solutions are always unstable (Proposition 55). We have derived a condition (Proposition 65) which determines the stability of the supercritical branches.

We begin by outlining the assumptions that are required to apply the theory developed in this section.

ASSUMPTION 50. *As in Theorem 47 we consider the bifurcation branch* $(t\boldsymbol{w}_0, \beta(t))$ *from* $(\mathbf{0}, 0)$ *of the flow (5.1) where* $\boldsymbol{w}_0 \in \mathrm{Fix}(H)$ *for an isotropy group* $H \leq G$*. The assumptions we make throughout this section are that*

1. $\phi$ *is* $G$*-equivariant and infinitely differentiable in* $\boldsymbol{w}$ *and* $\beta$*, with* $\partial_{\boldsymbol{w}}\phi(\mathbf{0}, 0) = \mathbf{0}$*.*
2. $G$ *acts absolutely irreducibly on* $\ker \partial_{\boldsymbol{w}}\phi(\mathbf{0}, 0)$ *so that* $\partial_{\boldsymbol{w}}\phi(\mathbf{0}, \beta) = c(\beta)I$ *for some scalar function* $c(\beta)$*.*
3. $c(0) = 0$ *and* $c'(0) > 0$*.*
4. $H \leq G$ *with* $\dim \mathrm{Fix}(H) = 1$*.*

The prudent reader will note that the Equivariant Branching Lemma (Theorem 47) requires the Assumptions 50.1, 50.2, and 50.4. Instead of requiring Assumption 50.3, the Equivariant Branching Lemma requires that $c(0) = 0$ and that $c'(0) \neq 0$, which guarantees that bifurcation occurs at $(\mathbf{0}, 0)$ (see (5.10) and (5.11)). The additional assumption that $c'(0) > 0$ is the basis for all of the results that we introduce in this section. In the case where $c'(0) < 0$, similar results hold, as we point out in Remarks 56 and 59.

DEFINITION 51. *The branch* $(t\boldsymbol{w}_0, \beta(t))$ *is* subcritical *if for all nonzero* $t$ *such that* $|t| < \epsilon$ *for some* $\epsilon > 0$*,* $t\beta(t)' < 0$*. The branch is* supercritical *if* $t\beta'(t) > 0$*.*

DEFINITION 52. *The branch* $(t\boldsymbol{w}_0, \beta(t))$ *is* transcritical *if* $\beta'(0) \neq 0$*. If* $\beta'(0) = 0$*, then the branch is called* pitchfork-like*.*

Golubitsky ([34] p.90) shows that

$$\mathrm{sgn}\beta'(0) = -\mathrm{sgn}c'(0)\mathrm{sgn} < \boldsymbol{w}_0, \partial^2_{\boldsymbol{ww}}\phi(\mathbf{0}, 0)[\boldsymbol{w}_0, \boldsymbol{w}_0] > .$$

We now prove the following generalization.

LEMMA 53. *If Assumption 50 holds, then*

$$\beta'(0) = \frac{- < \boldsymbol{w}_0, \partial^2_{\boldsymbol{ww}}\phi(\mathbf{0}, 0)[\boldsymbol{w}_0, \boldsymbol{w}_0] >}{2||\boldsymbol{w}_0||^2 c'(0)}.$$

*Proof.* As in (5.5), we write

$$\phi(t\boldsymbol{w}_0, \beta(t)) = tk(t, \beta)\boldsymbol{w}_0$$

where $k(t, \beta)$ is defined in (5.4). Differentiating (5.12) shows that

$$\partial_t k(t, \beta(t)) + \partial_\beta k(t, \beta(t))\beta'(t) = 0 \tag{5.14}$$

$$\implies \beta'(t) = -\frac{\partial_t k(t, \beta(t))}{\partial_\beta k(t, \beta(t))}. \tag{5.15}$$

By (5.10), $\partial_\beta k(0, 0) = c'(0)$. Differentiating (5.6) yields

$$\partial^2_{\boldsymbol{ww}}\phi(t\boldsymbol{w}_0, \beta)[\boldsymbol{w}_0, \boldsymbol{w}_0] = (2\partial_t k(t, \beta) + t\partial^2_{tt}k(t, \beta))\boldsymbol{w}_0 \tag{5.16}$$

showing that

$$\partial_t k(0, 0) = \frac{< \boldsymbol{w}_0, \partial^2_{\boldsymbol{ww}}\phi(\boldsymbol{0}, 0)[\boldsymbol{w}_0, \boldsymbol{w}_0] >}{2||\boldsymbol{w}_0||^2}.$$

Substituting this and $\partial_\beta k(0, 0) = c'(0)$ into (5.15) gives the desired result. □

Remark 54.

1. *([34] p.90) By Assumption 50.3,* $\text{sgn}\beta'(0) = -\text{sgn} < \boldsymbol{w}_0, \partial^2_{\boldsymbol{ww}}\phi(\boldsymbol{0}, 0)[\boldsymbol{w}_0, \boldsymbol{w}_0] >$. *This simplification of Lemma 53 proves useful when one is interested in determining whether bifurcating branches are subcritical or supercritical when the bifurcation is transcritical.*

2. *If one were interested in $\beta$ as a function of $t$ about $t = 0$, then equations (5.15) and (5.16) show that*

$$\beta'(t) = \frac{< \boldsymbol{w}_0, \partial^2_{\boldsymbol{ww}}\phi(t\boldsymbol{w}_0, \beta)[\boldsymbol{w}_0, \boldsymbol{w}_0] > \|\boldsymbol{w}_0\|^{-2} - t\partial^2_{tt}k(t, \beta)}{2\partial_\beta k(t, \beta)}. \tag{5.17}$$

3. *Assumptions 50.1, 50.3 and equations (5.10),(5.17) imply that $\beta'(t)$ is continuous at $t = 0$. Hence, for $t > 0$, $\beta'(0) < 0$ implies that the branch $(t\boldsymbol{w}_0, \beta(t))$ is subcritical. If $\beta'(0) > 0$, then the branch is supercritical for $t > 0$.*

4. *To determine whether a branch $(t\boldsymbol{w}_0, \beta(t))$ is supercritical or subcritical when $\beta'(0) = 0$, we consider $\beta''(0)$. $\beta''(0) > 0$ implies that for small $t < 0$, $\beta'(t) < 0$, and that for small $t > 0$, $\beta'(t) > 0$. Thus, when $\beta''(0) > 0$, the branch is supercritical. Similarly, if $\beta''(0) < 0$, then the branch is subcritical.*

Proposition 55. *([34] p.91) Suppose that Assumption 50 holds. If, for $t > 0$, the unique branch of bifurcating solutions $(t\boldsymbol{w}_0, \beta(t))$ to $\phi(\boldsymbol{w}, \beta)$, as guaranteed by Theorem 47, is subcritical, then it consists of unstable solutions.*

*Proof.* Write $\phi$ as in (5.5),

$$\phi(t\boldsymbol{w}_0, \beta(t)) = tk(t, \beta).$$

Note that (5.6) shows that $\boldsymbol{w}_0$ is an eigenvector of $\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0, \beta)$, with eigenvalue

$$\zeta(t, \beta) = k(t, \beta) + t\partial_t k(t, \beta). \tag{5.18}$$

Along a branch of solutions, $k(t, \beta) = 0$ (see (5.12)). From (5.14), we see that

$$\partial_t k(t, \beta) \;\; = \;\; -\partial_\beta k(t, \beta)\beta'(t).$$

Substituting this and $k(t, \beta) = 0$ into (5.18), we have that

$$\zeta(t, \beta) = -t\partial_\beta k(t, \beta)\beta'(t). \tag{5.19}$$

By (5.10),

$$\partial_\beta k(0, 0) = c'(0)$$

which is positive by Assumption 50.3. By Assumption 50.1, $\partial_\beta k(t, \beta)$ is continuous, and so $\partial_\beta k(t, \beta(t))$ is positive for all sufficiently small $t > 0$. Furthermore, by the assumption of subcriticality, we have that $t\beta'(t) < 0$ for small $t$. Hence the eigenvalue

$$\zeta(t, \beta) > 0. \tag{5.20}$$

for small $t$ and $\beta$. Thus, this branch is unstable for sufficiently small $t$. □

Remark 56. *If Assumptions 50.1, 50.2, and 50.4 hold, if $c(0) = 0$, and if $c'(0) < 0$, then the argument above shows that supercritical branches are unstable.*

To prove a result regarding supercritical branches from transcritical bifurcation, we first need to prove the following claim.

Claim 57. *([34] p.93) If Assumption 50 holds, then*

$$\operatorname{trace}(\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0, \beta)) = \dim(V)c'(0)\beta'(0)t + \mathcal{O}(t^2)$$

*where $V$ is the Banach space defined in (5.13).*

*Proof.* The Taylor series for $\phi(\boldsymbol{w}, \beta)$ about $\boldsymbol{w} = \boldsymbol{0}$ is

$$\phi(\boldsymbol{w}, \beta) = \phi(\boldsymbol{0}, \beta) + \partial_{\boldsymbol{w}}\phi(\boldsymbol{0}, \beta)\boldsymbol{w} + \partial^2_{\boldsymbol{ww}}\phi(0, \beta)[\boldsymbol{w}, \boldsymbol{w}] + \mathcal{O}(\boldsymbol{w}^3). \tag{5.21}$$

Equation (5.3) shows that $\phi(\boldsymbol{0}, \beta) = \boldsymbol{0}$, and by Assumption 50.2, $\partial_{\boldsymbol{w}}\phi(\boldsymbol{0}, \beta) = c(\beta)I$. Letting

$$Q(\boldsymbol{w}, \beta) = \frac{1}{2}\partial^2_{\boldsymbol{ww}}\phi(0, \beta)[\boldsymbol{w}, \boldsymbol{w}] \tag{5.22}$$

gives

$$\phi(\boldsymbol{w},\beta) = c(\beta)\boldsymbol{w} + Q(\boldsymbol{w},\beta) + \mathcal{O}(\boldsymbol{w}^3). \tag{5.23}$$

Hence,

$$\partial_{\boldsymbol{w}}\phi(\boldsymbol{w},\beta) = c(\beta)I + \partial_{\boldsymbol{w}}Q(\boldsymbol{w},\beta) + \mathcal{O}(\boldsymbol{w}^2)$$

from which it follows that

$$\operatorname{trace}(\partial_{\boldsymbol{w}}\phi(\boldsymbol{w},\beta)) = \dim(V)c(\beta) + \operatorname{trace}(\partial_{\boldsymbol{w}}Q(\boldsymbol{w},\beta)) + \mathcal{O}(\boldsymbol{w}^2).$$

Observe that $Q$ is $G$-equivariant by the equivariance of $\phi$, from which we get $Q(g\boldsymbol{w},\beta) = gQ(\boldsymbol{w},\beta)$ and so

$$\partial_{\boldsymbol{w}}Q(g\boldsymbol{w},\beta) = g\partial_{\boldsymbol{w}}Q(\boldsymbol{w},\beta)g^{-1}$$

giving

$$\operatorname{trace}(\partial_{\boldsymbol{w}}Q(g\boldsymbol{w},\beta)) = \operatorname{trace}(\partial_{\boldsymbol{w}}Q(\boldsymbol{w},\beta)).$$

Thus, $\operatorname{trace}(\partial_{\boldsymbol{w}}Q(\boldsymbol{w},\beta))$ is a $G$-invariant function. Furthermore, $\operatorname{trace}(\partial_{\boldsymbol{w}}Q(\boldsymbol{w},\beta))$ is linear in $\boldsymbol{w}$ since $Q$ is quadratic. Therefore, Propositions 45 and 46.6 assure that

$$\operatorname{trace}(\partial_{\boldsymbol{w}}Q(\boldsymbol{w},\beta)) = 0.$$

Finally, we see that

$$\operatorname{trace}(\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta(t))) = \dim(V)c(\beta(t)) + \mathcal{O}(t^2),$$

which can be rewritten using the Taylor expansion of $c(\beta(t))$ about $t = 0$, showing that

$$\begin{aligned}\operatorname{trace}(\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta(t))) &= \dim(V)\left(c(0) + c'(0)\beta'(0)t + \mathcal{O}(t^2)\right)\\ &= \dim(V)c'(0)\beta'(0)t + \mathcal{O}(t^2),\end{aligned} \tag{5.24}$$

where the last equality follows from Assumption 50.3. □

Proposition 58. *([34] p.93) Suppose that Assumption 50 holds. If $\beta'(0) > 0$, then for $t > 0$, the unique branch of bifurcating solutions $(t\boldsymbol{w}_0,\beta(t))$ to $\phi(\boldsymbol{w},\beta)$, as guaranteed by Theorem 47, is supercritical and consists of unstable solutions.*

*Proof.* Remark 54.3 implies that $(t\boldsymbol{w}_0,\beta(t))$ is supercritical. Claim 57 shows that

$$\operatorname{trace}(\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta)) = \dim(V)c'(0)\beta'(0)t + \mathcal{O}(t^2).$$

from which it follows that $\operatorname{trace}(\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta))$ is positive for sufficiently small $t$. Thus, some eigenvalue of $\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta)$ has positive real part. □

Remark 59. *If Assumptions 50.1, 50.2, and 50.4 hold, if $c(0) = 0$, and if $c'(0) < 0$, then the argument above shows that subcritical branches are unstable.*

We summarize Propositions 55 and 58 in the following theorem.

THEOREM 60. *([34] p.90) Suppose that Assumptions 50.1, 50.2, and 50.4 hold, that* $c(0) = 0$*, and that* $c'(0) \neq 0$*. Then at a transcritical bifurcation, each branch of bifurcating solutions to* $\phi(\boldsymbol{w}, \beta) = \mathbf{0}$*, as guaranteed by Theorem 47, consists of unstable solutions.*

*Proof.* The theorem follows from Propositions 55 and 58, and Remarks 56 and 59. □

We now examine the pitchfork-like case when $\beta'(0) = 0$.

THEOREM 61. *([34] p.93) Suppose that* $\beta'(0) = 0$*. In addition to Assumption 50, we further assume that some term in the Taylor expansion of* $\hat{\phi}$ *from (5.2) is non-zero and that* $\partial_{\boldsymbol{w}} Q(\boldsymbol{w}_0, \beta)$ *has an eigenvalue with nonzero real part, where* $Q(\boldsymbol{w}, \beta)$ *is the quadratic part of* $\phi$ *as in (5.22). Then the unique branch of bifurcating solutions* $(t\boldsymbol{w}_0, \beta(t))$ *to* $\phi(\boldsymbol{w}, \beta)$*, as guaranteed by Theorem 47, consists of unstable solutions.*

REMARK 62. *In addition to Assumption 50, Theorem 61 also requires that some term in the Taylor expansion of* $\hat{\phi}$ *from (5.2) is non-zero and that* $\partial_{\boldsymbol{w}} Q(\boldsymbol{w}_0, \beta)$ *has an eigenvalue with nonzero real part. These hypotheses are automatically satisfied when the bifurcation is transcritical,* $\beta'(0) \neq 0$ *[34].*

To determine whether solution branches from a pitchfork-like bifurcation are either subcritical or supercritical is to compute $\beta''(0)$ (see Remark 54.4).

LEMMA 63. *Suppose that Assumption 50 holds. If* $\beta'(0) = 0$*, then*

$$\beta''(0) = \frac{-\langle \boldsymbol{w}_0, \partial^3_{\boldsymbol{www}} \phi(\mathbf{0}, 0)[\boldsymbol{w}_0, \boldsymbol{w}_0, \boldsymbol{w}_0] \rangle}{3\|\boldsymbol{w}_0\|^2 c'(0)}.$$

*Proof.* As in (5.5), we write

$$\phi(t\boldsymbol{w}_0, \beta(t)) = tk(t, \beta)\boldsymbol{w}_0$$

where $k(t, \beta)$ is defined in (5.4). Twice differentiating (5.12) (or, equivalently, once differentiating (5.14)) shows that

$$\partial^2_{tt} k + \partial_\beta \partial_t k \beta'(t) + \left(\partial_t \partial_\beta k + \partial^2_{\beta\beta} k \beta'(t)\right) \beta'(t) + \partial_\beta k \beta''(t) = 0.$$

Thus

$$\beta''(t) = \frac{-\partial^2_{tt} k - 2\partial_\beta \partial_t k \beta'(t) - \partial^2_{\beta\beta} k \beta'(t)^2}{\partial_\beta k}$$

and so

$$\beta''(0) = \frac{-\partial^2_{tt} k(\mathbf{0}, 0)}{\partial_\beta k(\mathbf{0}, 0)}. \tag{5.25}$$

By (5.10), $\partial_\beta k(0,0) = c'(0)$. Differentiating (5.16) with respect to $t$ gives

$$\partial^3_{\boldsymbol{www}}\phi(t\boldsymbol{w}_0,\beta)[\boldsymbol{w}_0,\boldsymbol{w}_0,\boldsymbol{w}_0] = (3\partial^2_{tt}k(t,\beta) + t\partial^3_{ttt}k(t,\beta))\boldsymbol{w}_0$$

from which it follows that

$$\partial^2_{tt}k(0,0) = \frac{\langle \boldsymbol{w}_0, \partial^3_{\boldsymbol{www}}\phi(\boldsymbol{0},0)[\boldsymbol{w}_0,\boldsymbol{w}_0,\boldsymbol{w}_0]\rangle}{3\|\boldsymbol{w}_0\|^2}.$$

Substituting this and $\partial_\beta k(0,0) = c'(0)$ into (5.25) gives the desired result. □

The following corollary is a consequence of Lemma 63, Definition 51, and Assumption 50.3.

Corollary 64. *If Assumption 50 holds, then at a pitchfork-like bifurcation,*

$$\operatorname{sgn}(\beta''(0)) = -\operatorname{sgn}(\langle \boldsymbol{w}_0, \partial^3_{\boldsymbol{www}}\phi(\boldsymbol{0},0)[\boldsymbol{w}_0,\boldsymbol{w}_0,\boldsymbol{w}_0]\rangle).$$

We conclude this section with a result which addresses the stability of supercritical branches from pitchfork-like bifurcations.

Proposition 65. *Suppose Assumption 50 holds. If the unique branch of bifurcating solutions $(t\boldsymbol{w}_0,\beta(t))$, as guaranteed by Theorem 47, is pitchfork-like with $\beta''(0) > 0$, and if*

$$\sum_{i,j,m} \frac{\partial^3\phi_m(\boldsymbol{0},0)}{\partial x_i \partial x_j \partial x_m}[\boldsymbol{w}_0]_i[\boldsymbol{w}_0]_j > 0,$$

*then the branch is supercritical and consists of unstable solutions.*

*Proof.* The branch is supercritical by Remark 54.4. To show instability, we determine $\operatorname{trace}(\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta))$ as in (5.24). Since $\beta'(0) = 0$, it is necessary to compute the quadratic term in the Taylor series given in each of (5.23) and (5.24). Letting

$$T(\boldsymbol{w},\beta) = \frac{1}{6}\partial^3_{\boldsymbol{www}}\phi(\boldsymbol{0},\beta)[\boldsymbol{w},\boldsymbol{w},\boldsymbol{w}], \tag{5.26}$$

then (5.23) can be rewritten as

$$\phi(\boldsymbol{w},\beta) = c(\beta)\boldsymbol{w} + Q(\boldsymbol{w},\beta) + T(\boldsymbol{w},\beta) + \mathcal{O}(\boldsymbol{w}^4)$$

and from the proof to Proposition 58 it follows that

$$\operatorname{trace}(\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta(t))) = \dim(V)c(\beta(t)) + \operatorname{trace}(\partial_{\boldsymbol{w}}T(t\boldsymbol{w}_0,\beta)) + \mathcal{O}(t^3).$$

The Taylor expansion for $c(\beta(t))$ about $t = 0$ given in (5.24) becomes

$$c(\beta(t)) = c'(0)\beta'(0)t + (c''(0)\beta'(0)^2 + c'(0)\beta''(0))\frac{t^2}{2} + \mathcal{O}(t^3).$$

Thus, $\mathrm{trace}(\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta(t)))$ is equal to

$$\dim(V)\left(c'(0)\beta'(0)t+(c''(0)\beta'(0)^2+c'(0)\beta''(0))\frac{t^2}{2}\right)+\mathrm{trace}(\partial_{\boldsymbol{w}}T(t\boldsymbol{w}_0,\beta))+\mathcal{O}(t^3).$$

This and Assumption 50.3 show that when $\beta'(0)=0$ and $\beta''(0)>0$,

$$\mathrm{trace}(\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta(t)))>0$$

if

$$\mathrm{trace}(\partial_{\boldsymbol{w}}T(t\boldsymbol{w}_0,\beta))>0$$

for sufficiently small $t$. Thus, if $\mathrm{trace}(\partial_{\boldsymbol{w}}T(t\boldsymbol{w}_0,\beta))>0$ for sufficiently small $t$, then some eigenvalue of $\partial_{\boldsymbol{w}}\phi(t\boldsymbol{w}_0,\beta)$ is positive, which implies that the supercritical branch $(t\boldsymbol{w}_0,\beta(t))$ is unstable .

We now show that $\mathrm{sgn}(\mathrm{trace}(\partial_{\boldsymbol{w}}T(t\boldsymbol{w}_0,\beta)))$ for small $t$ is determined by

$$\mathrm{sgn}\left(\sum_{i,j,k}\frac{\partial^3\phi_k(\mathbf{0},0)}{\partial x_i\partial x_j\partial x_k}[\boldsymbol{w}_0]_i[\boldsymbol{w}_0]_j\right).$$

The function $[T(\boldsymbol{w},\beta)]_l$ from (5.26) can be written as

$$\begin{aligned}\frac{1}{6}\quad&\Big(\sum_{i\neq m,j\neq m,k\neq m}\frac{\partial^3\phi_l(\mathbf{0},\beta)}{\partial x_i\partial x_j\partial x_k}x_ix_jx_k+3\sum_{i\neq m,j\neq m}\frac{\partial^3\phi_l(\mathbf{0},\beta)}{\partial x_i\partial x_j\partial x_m}x_ix_jx_m\\&+\ 3\sum_{i\neq m}\frac{\partial^3\phi_l(\mathbf{0},\beta)}{\partial x_i\partial x_m\partial x_m}x_ix_m^2+\frac{\partial^3\phi_l(\mathbf{0},\beta)}{\partial x_m^3}x_m^3\Big).\end{aligned}\tag{5.27}$$

Thus, $\partial_{x_m}[T(t\boldsymbol{w}_0,\beta)]_l$ is

$$\frac{1}{6}t^2\left(3\sum_{i\neq m,j\neq m}\frac{\partial^3\phi_l(\mathbf{0},\beta)}{\partial x_i\partial x_j\partial x_m}[\boldsymbol{w}_0]_i[\boldsymbol{w}_0]_j+6\sum_{i\neq m}\frac{\partial^3\phi_l(\mathbf{0},\beta)}{\partial x_i\partial x_m\partial x_m}[\boldsymbol{w}_0]_i[\boldsymbol{w}_0]_m+3\frac{\partial^3\phi_l(\mathbf{0},\beta)}{\partial x_m^3}[\boldsymbol{w}_0]_m^2\right)$$

which shows that

$$[\partial_{\boldsymbol{w}}T(t\boldsymbol{w}_0,\beta)]_{lm}=\frac{1}{2}t^2\sum_{i,j}\frac{\partial^3\phi_l(\mathbf{0},\beta)}{\partial x_i\partial x_j\partial x_m}[\boldsymbol{w}_0]_i[\boldsymbol{w}_0]_j.$$

It follows that

$$\mathrm{trace}(\partial_{\boldsymbol{w}}T(t\boldsymbol{w}_0,\beta))=\frac{1}{2}t^2\sum_{i,j,m}\frac{\partial^3\phi_m(\mathbf{0},\beta)}{\partial x_i\partial x_j\partial x_m}[\boldsymbol{w}_0]_i[\boldsymbol{w}_0]_j$$

which is positive for sufficiently small $t$ if

$$\sum_{i,j,m}\frac{\partial^3\phi_m(\mathbf{0},0)}{\partial x_i\partial x_j\partial x_m}[\boldsymbol{w}_0]_i[\boldsymbol{w}_0]_j>0.$$

□

Derivation of the Liapunov-Schmidt Reduction

In the last section, we developed the theoretical tools necessary to analyze bifurcation of equilibria, of a $G$-equivariant system (5.1)

$$\dot{\boldsymbol{w}} = \phi(\boldsymbol{w}, \beta),$$

when two simplifying assumptions are made. These simplifying assumptions were made so that the assumptions of the Equivariant Branching Lemma (Theorem 47) are met. The first assumption is that $(\boldsymbol{w} = \mathbf{0}, \beta = 0)$ is an equilibrium of (5.1). The second assumption is that $\partial_{\boldsymbol{w}}\phi(\mathbf{0}, 0) = \mathbf{0}$. In other words, we assumed that bifurcation occurs at $(\mathbf{0}, 0)$, and that at the bifurcation, the Jacobian of $\phi$ vanishes.

This section examines in detail how to transform an arbitrary $G$-equivariant system such as (3.15),

$$\dot{\boldsymbol{x}} = \psi(\boldsymbol{x}, \beta),$$

where

$$\psi : B_2 \times \Re \to B_0,$$

as in (3.16), to an equivalent system where the above two assumptions hold.

First, if a bifurcation of equilibria to (3.15) occurs at $(\boldsymbol{x}^*, \beta^*)$, then the translation $\psi(\boldsymbol{x} + \boldsymbol{x}^*, \beta + \beta^*)$ has a bifurcation at $(\mathbf{0}, 0)$ as required by Theorem 47. We continue by assuming that any necessary translation has been performed so that $\psi = \mathbf{0}$ has a bifurcation of solutions at $(\mathbf{0}, 0)$.

Secondly, the Equivariant Branching Lemma requires that

$$\Psi := \partial_{\boldsymbol{x}}\psi(\mathbf{0}, 0) = \mathbf{0},$$

that the Jacobian of $\psi$ must vanish at the bifurcation. Since this is not the case for an arbitrary system, it is necessary to consider the Liapunov-Schmidt reduction of (3.15), $\phi$, which is the restriction of $\psi$ onto $\ker \Psi$ about $(\mathbf{0}, 0)$. More precisely, $\psi$ is restricted to $\ker \Psi$, and $\phi$ is the projection of that restriction onto $\ker \Psi$. To make this formal, decompose $B_2$ and $B_0$ from (3.16) as

$$B_2 = \ker \Psi \oplus \mathcal{M} \text{ and } B_0 = \mathcal{N} \oplus \text{range}\Psi \tag{5.28}$$

where $\mathcal{M}$ and $\mathcal{N}$ are vector space complements of $\ker \Psi$ and range$\Psi$ respectively.

The following derivation is from p.27-28 and p.292-293 of [33] . See also p.10 of [34]. Let $E$ be the projector onto range$\Psi$ with $\ker E = \mathcal{N}$. Thus $I - E$ projects onto $\mathcal{N}$ with $\ker(I - E) = \text{range}\Psi$. Observe that $\psi = \mathbf{0}$ if and only if the components of $\psi$ in range$\Psi$ and in $\mathcal{N}$ are zero:

$$\psi(\boldsymbol{x}, \beta) = \mathbf{0} \Leftrightarrow E\psi(\boldsymbol{x}, \beta) = \mathbf{0} \text{ and } (I - E)\psi(\boldsymbol{x}, \beta) = \mathbf{0}. \tag{5.29}$$

Consider the decomposition $\boldsymbol{x} = \boldsymbol{w} + U$, where $\boldsymbol{w} \in \ker \Psi$ and $U \in \mathcal{M}$, so that the problem $E\psi(\boldsymbol{x}, \beta) = \mathbf{0}$ can be rewritten as

$$E\psi(\boldsymbol{w}, U, \beta) = E\psi(\boldsymbol{w} + U, \beta) = \mathbf{0}.$$

We define the matrix $L$ as

$$L := E\Psi|_{\mathcal{M}}, \tag{5.30}$$

the Jacobian $\partial_{\boldsymbol{x}}\psi(\mathbf{0}, 0)$ projected onto range$\Psi$, and restricted to $\mathcal{M}$. Thus, $L$ is invertible, and the Implicit Function Theorem shows that $E\psi(\boldsymbol{w} + U, \beta) = \mathbf{0}$ can be solved for $U = U(\boldsymbol{w}, \beta)$ near $(\mathbf{0}, 0)$,

$$E\psi(\boldsymbol{x}, \beta) = E\psi(\boldsymbol{w} + U(\boldsymbol{w}, \beta), \beta) = \mathbf{0}. \tag{5.31}$$

Substituting this expression into (5.29), we see that $\psi(\boldsymbol{x}, \beta) = \mathbf{0}$ if and only if

$$(I - E)\psi(\boldsymbol{w} + U(\boldsymbol{w}, \beta), \beta) = \mathbf{0}.$$

This function is the Liapunov-Schmidt reduction $\phi(\boldsymbol{w}, \beta)$:

$$\begin{aligned} \phi &: \ker \Psi \times \Re \to \mathcal{N} \\ \phi(\boldsymbol{w}, \beta) &= (I - E)\psi(\boldsymbol{w} + U(\boldsymbol{w}, \beta), \beta). \end{aligned} \tag{5.32}$$

Using the chain rule, the Jacobian of (5.32) is the matrix

$$\partial_{\boldsymbol{w}}\phi(\boldsymbol{w}, \beta) = (I - E) \cdot \partial_{\boldsymbol{x}}\psi(\boldsymbol{x}, \beta) \cdot (I + \partial_{\boldsymbol{w}}U) \tag{5.33}$$

Since $\ker(I - E) = \text{range}\Psi$, then

$$\partial_{\boldsymbol{w}}\phi(\mathbf{0}, 0) = \mathbf{0} \tag{5.34}$$

and so the Jacobian of $\phi$ vanishes as required. Furthermore, (5.29) and (5.31) show that $\phi = \mathbf{0}$ if and only if $\psi = \mathbf{0}$. Thus, the roots of (5.32) are the equilibria of (3.15). By (5.34), the group and bifurcation theory from the last section can be applied to $\phi = \mathbf{0}$.

Consider the dynamical system formulated with respect to the Liapunov-Schmidt reduction of $\psi$:

$$\dot{\boldsymbol{w}} = \phi(\boldsymbol{w}, \beta).$$

Ascertaining the bifurcation structure of the equilibria of this system, solutions to $\phi(\boldsymbol{w}, \beta) = \mathbf{0}$, means determining the bifurcating branches $(t\boldsymbol{w}, \beta(t))$ from $(\mathbf{0}, 0)$ for $\boldsymbol{w} \in \ker \Psi$. The associated bifurcating branch of $\psi = \mathbf{0}$ is straightforward to get:

$$\begin{gathered} (t\boldsymbol{w}, \beta(t)) \text{ is a bifurcating branch of } \phi = 0 \\ \text{if and only if} \\ \begin{pmatrix} \boldsymbol{x}^* \\ \beta^* \end{pmatrix} + \begin{pmatrix} t\boldsymbol{w} \\ \beta(t) \end{pmatrix} \text{ is a bifurcating branch of } \psi = 0. \end{gathered} \tag{5.35}$$

It is convenient to use an equivalent representation of the Liapunov-Schmidt reduction (5.32). Let

$$\{\boldsymbol{w}_i\}_{i=1}^m \text{ be a basis for } \ker \Psi$$

and let $W$ be the $(NK+K) \times m$ matrix whose column space is $\ker \Psi$. So

$$W = \begin{pmatrix} | & | & & | \\ \boldsymbol{w}_1 & \boldsymbol{w}_2 & \ldots & \boldsymbol{w}_m \\ | & | & & | \end{pmatrix}.$$

Thus, for every $\boldsymbol{w} \in \ker \Psi$, there is a $\boldsymbol{z} \in \Re^m$ such that $W\boldsymbol{z} = \boldsymbol{w}$. Now define $r$ by

$$\begin{aligned} r &: \Re^m \times \Re \to \Re^m \\ r(\boldsymbol{z}, \beta) &= W^T \phi(W\boldsymbol{z}, \beta) \\ &= W^T (I - E) \psi(W\boldsymbol{z} + U(W\boldsymbol{z}, \beta), \beta) \end{aligned} \tag{5.36}$$

where the last equality is from (5.32). We say that $r$ is equivalent to $\phi$ since

$$r(\boldsymbol{z}, \beta) = \mathbf{0} \Leftrightarrow \phi(\boldsymbol{w}, \beta) = \mathbf{0} \Leftrightarrow \psi(\boldsymbol{x}, \beta) = \mathbf{0},$$

which follows from (5.29), (5.31) and (5.32). The Jacobian of $r$, which is similar to (5.33), is the $m \times m$ matrix

$$\partial_{\boldsymbol{z}} r(\boldsymbol{z}, \beta) = W^T (I - E) \cdot \partial_{\boldsymbol{x}} \psi(\boldsymbol{x}, \beta) \cdot (W + \partial_{\boldsymbol{w}} UW). \tag{5.37}$$

So we have introduced the necessary ingredients to define a dynamical system defined by $r$

$$\dot{\boldsymbol{z}} = r(\boldsymbol{z}, \beta).$$

Ascertaining the bifurcation structure of the equilibria of this system, solutions to $r(\boldsymbol{z}, \beta) = \mathbf{0}$, means determining the bifurcating branches $(t\boldsymbol{z}, \beta(t))$ from $(\mathbf{0}, 0)$ for $\boldsymbol{z} \in \Re^m$. The bifurcating branch of $\psi = \mathbf{0}$ is found via the following relationship:

$$\begin{gathered} (t\boldsymbol{z}, \beta(t)) \text{ is a bifurcating branch of } r = 0 \\ \text{if and only if} \\ \begin{pmatrix} \boldsymbol{z}^* \\ \beta^* \end{pmatrix} + \begin{pmatrix} tW\boldsymbol{z} \\ \beta(t) \end{pmatrix} \text{ is a bifurcating branch of } \psi = 0. \end{gathered} \tag{5.38}$$

We now compute the derivative of $r$ with respect to $\beta$, which we will need in chapter 8 when examining saddle-node bifurcations. Beginning with the definition (5.36), we see that

$$\begin{aligned} \partial_\beta r(\boldsymbol{z}, \beta) &= W^T (I - E) \frac{\partial}{\partial \beta} \psi(\boldsymbol{x}, \beta) \\ &= W^T (I - E) \left( \partial_\beta \psi(\boldsymbol{x}, \beta) + \partial_{\boldsymbol{x}} \psi(\boldsymbol{x}, \beta) \frac{\partial}{\partial \beta} \left( W\boldsymbol{z} + U(W\boldsymbol{z}, \beta) \right) \right) \\ &= W^T (I - E) \left( \partial_\beta \psi(\boldsymbol{x}, \beta) + \partial_{\boldsymbol{x}} \psi(\boldsymbol{x}, \beta) \partial_\beta U \right). \end{aligned}$$

Since $(I-E)\partial_{\boldsymbol{x}}\psi(\mathbf{0},0)=\mathbf{0}$, then

$$\partial_\beta r(\mathbf{0},0)=W^T(I-E)\partial_\beta\psi(\mathbf{0},0). \tag{5.39}$$

Next, we compute the three dimensional array of second derivatives of $r$ and the 4 dimensional array of third derivatives of $r$. These prove necessary when we compute $\beta'(0)$ and $\beta''(0)$ in chapter 6 using Lemma 53 and Lemma 63 respectively. To determine the three dimensional array of second derivatives of $r$, we write (5.37) in component form as

$$\frac{\partial r_i}{\partial z_j}=<\boldsymbol{w}_i,(I-E)\partial_{\boldsymbol{x}}\psi(\boldsymbol{x},\beta)\left(\boldsymbol{w}_j+\frac{\partial U}{\partial z_j}\right)>.$$

Thus, we get that

$$\frac{\partial^2 r_i}{\partial z_j\partial z_k}=<\boldsymbol{w}_i,(I-E)\left(\partial_{\boldsymbol{x}}\psi(\boldsymbol{x},\beta)\frac{\partial^2 U}{\partial z_j\partial z_k}+\partial^2_{\boldsymbol{x}}\psi(\boldsymbol{x},\beta)[\boldsymbol{w}_j+\frac{\partial U}{\partial z_j},\boldsymbol{w}_k+\frac{\partial U}{\partial z_k}]\right)> \tag{5.40}$$

It can be shown that ([33] p.31)

$$\partial_{\boldsymbol{w}}U(\mathbf{0},0)=\mathbf{0}, \tag{5.41}$$

from which it follows that $\frac{\partial U}{\partial z_j}(\mathbf{0},0)=\partial_{\boldsymbol{w}}U(\mathbf{0},0)\frac{\partial \boldsymbol{w}}{\partial z_j}(\mathbf{0})=\mathbf{0}$. Furthermore, since $(I-E)\partial_{\boldsymbol{x}}\psi(\mathbf{0},0)=\mathbf{0}$, then

$$\frac{\partial^2 r_i}{\partial z_j\partial z_k}(\mathbf{0},0)=<\boldsymbol{w}_i,(I-E)\partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_j,\boldsymbol{w}_k]>. \tag{5.42}$$

Applying the chain rule to (5.40), we get the 4 dimensional array of third derivatives

$$\begin{aligned}\frac{\partial^3 r_i}{\partial z_j\partial z_k\partial z_l}=\langle\boldsymbol{w}_i \quad , \quad & (I-E)(\partial^2_{\boldsymbol{x}}\psi(\boldsymbol{x},\beta)[\boldsymbol{w}_l+\frac{\partial U}{\partial z_l},\frac{\partial^2 U}{\partial z_j\partial z_k}]+\partial_{\boldsymbol{x}}\psi(\boldsymbol{x},\beta)\frac{\partial^3 U}{\partial z_j\partial z_k\partial z_l}\\ + \quad & \partial^3_{\boldsymbol{x}}\psi[\boldsymbol{w}_j+\frac{\partial U}{\partial z_j},\boldsymbol{w}_k+\frac{\partial U}{\partial z_k},\boldsymbol{w}_l+\frac{\partial U}{\partial z_l}]\\ + \quad & \partial^2_{\boldsymbol{x}}\psi[\boldsymbol{w}_j+\frac{\partial U}{\partial z_j},\frac{\partial^2 U}{\partial z_k\partial z_l}]+\partial^2_{\boldsymbol{x}}\psi[\boldsymbol{w}_k+\frac{\partial U}{\partial z_k},\frac{\partial^2 U}{\partial z_j\partial z_l}])\rangle\end{aligned} \tag{5.43}$$

Using the fact that $\partial_{\boldsymbol{z}}U(\mathbf{0},0)=\mathbf{0}$ and $(I-E)\partial_{\boldsymbol{x}}\psi=\mathbf{0}$, it follows that

$$\begin{aligned}\frac{\partial^3 r_i}{\partial z_j\partial z_k\partial z_l}(\mathbf{0},0)=\langle\boldsymbol{w}_i \quad , \quad & (I-E)(\partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_l,\frac{\partial^2 U}{\partial z_j\partial z_k}(\mathbf{0},0)]\\ + \quad & \partial^3_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_j,\boldsymbol{w}_k,\boldsymbol{w}_l]\\ + \quad & \partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_j,\frac{\partial^2 U}{\partial z_k\partial z_l}(\mathbf{0},0)]\\ + \quad & \partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_k,\frac{\partial^2 U}{\partial z_j\partial z_l}(\mathbf{0},0)])\rangle.\end{aligned} \tag{5.44}$$

To explicitly compute $\frac{\partial^3 r_i}{\partial z_j \partial z_k \partial z_l}(\mathbf{0},0)$, we first derive $\frac{\partial^2 U}{\partial z_j \partial z_k}(\mathbf{0},0)$. To do this, define

$$\theta(\boldsymbol{z},\beta) := E\psi(W\boldsymbol{z} + U(W\boldsymbol{z},\beta),\beta).$$

Observe that $\psi = 0$ implies that $\theta = 0$. Differentiating $\theta = 0$ yields

$$\frac{\partial \theta}{\partial z_j} = E\partial_{\boldsymbol{x}}\psi(\boldsymbol{w}_j + \frac{\partial U}{\partial z_j}) = 0$$

and

$$\frac{\partial^2 \theta}{\partial z_j \partial z_k} = E\left(\partial^2_{\boldsymbol{x}}\psi[\boldsymbol{w}_j + \frac{\partial U}{\partial z_j}, \boldsymbol{w}_k + \frac{\partial U}{\partial z_k}] + \partial_{\boldsymbol{x}}\psi \frac{\partial^2 U}{\partial z_j \partial z_k}\right) = 0.$$

Since $\partial_{\boldsymbol{z}} U(\mathbf{0},0) = \mathbf{0}$, we get

$$\frac{\partial \theta}{\partial z_j \partial z_k}(\mathbf{0},0) = E\left(\partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_j,\boldsymbol{w}_k] + \partial_{\boldsymbol{x}}\psi(\mathbf{0},0)\frac{\partial^2 U}{\partial z_j \partial z_k}\right) = 0,$$

and $E\partial_{\boldsymbol{x}}\psi(\mathbf{0},0) = L$ (from (5.30)) shows that

$$\frac{\partial^2 U}{\partial z_j \partial z_k}(\mathbf{0},0) = -L^{-1}E\partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_j,\boldsymbol{w}_k]. \tag{5.45}$$

Finally, substituting (5.45) into (5.44) shows that

$$\begin{aligned}\frac{\partial^3 r_i}{\partial z_j \partial z_k \partial z_l}(\mathbf{0},0) = \langle \boldsymbol{w}_i \quad , \quad & (I-E)(\partial^3_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_j,\boldsymbol{w}_k,\boldsymbol{w}_l] \\ - \quad & \partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_j, L^{-1}E\partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_k,\boldsymbol{w}_l]] \\ - \quad & \partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_k, L^{-1}E\partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_j,\boldsymbol{w}_l]] \\ - \quad & \partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_l, L^{-1}E\partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{w}_j,\boldsymbol{w}_k]])\rangle. \end{aligned} \tag{5.46}$$

In chapter 6, it proves useful to use Lemma 63 to compute $\beta''(0)$,

$$\beta''(0) = \langle \boldsymbol{x}_0, \partial^3_{\boldsymbol{zzz}} r(\mathbf{0},0)[\boldsymbol{z}_0,\boldsymbol{z}_0,\boldsymbol{z}_0]\rangle,$$

where $r$ is the Liapunov Schmidt reduction of some function $\psi$, $\boldsymbol{z}_0$ is defined as $W\boldsymbol{z}_0 = \boldsymbol{u}$, where $\boldsymbol{z}_0$ is a solution branch of $r$, and $\boldsymbol{u}$ is the corresponding solution branch of $\psi$. The next Lemma writes $< \boldsymbol{z}_0, \partial^3_{\boldsymbol{zzz}} r(\mathbf{0},0)[\boldsymbol{z}_0,\boldsymbol{z}_0,\boldsymbol{z}_0] >$ in terms of $\psi$ and $\boldsymbol{u}$.

LEMMA 66. *Let $W\boldsymbol{z}_0 = \boldsymbol{u}$, where the columns of $W$ are $\{\boldsymbol{w}_i\}$, a basis for* $\ker \partial_{\boldsymbol{x}}\psi(\mathbf{0},0)$. *Then $\langle \boldsymbol{z}_0, \partial^3_{\boldsymbol{zzz}} r(\mathbf{0},0)[\boldsymbol{z}_0,\boldsymbol{z}_0,\boldsymbol{z}_0]\rangle$ is equal to*

$$\langle \boldsymbol{u}, \partial^3_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{u},\boldsymbol{u},\boldsymbol{u}] - 3\partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{u}, L^{-1}E\partial^2_{\boldsymbol{x}}\psi(\mathbf{0},0)[\boldsymbol{u},\boldsymbol{u}]]\rangle$$

*Proof.* The Lemma follows from (5.46). □

Equivariance of the Reduction

By assumption, the vector valued function $\psi$ from (3.15),

$$\dot{\boldsymbol{x}} = \psi(\boldsymbol{x}, \beta),$$

is $G$-equivariant. The discussion in the previous section raises a few questions, the first of which is

*For what group is the Liapunov-Schmidt reduced function $\phi$ equivariant?*

This is answered by Lemma 67.1: If $\mathcal{M}$ and $\mathcal{N}$ from (5.28) are $G$-invariant then $\phi$ is $G$-equivariant. Another question is:

*For what group is the Liapunov Schmidt reduction $r$ equivariant?*

By Lemma 67.2 , the Lie group that acts equivariantly on $r$ is constructed from $G$ in the following way. Let $\{\boldsymbol{w}_i\}_{i=1}^m$ be a basis for $\ker \Psi$. For each $g \in G$ Proposition 46.2 assures that $g\boldsymbol{w}_j = \sum_i a_{ij}\boldsymbol{w}_i$ for $a_{ij} \in \Re$. Define the $m \times m$ matrix $A(g)$ by setting

$$[A(g)]_{ij} := a_{ij}. \tag{5.47}$$

The group for which $r$ is equivariant is

$$\mathcal{A} := \{A(g) | g \in G\}. \tag{5.48}$$

The previous discussion is summarized in the following Lemma.

Lemma 67.

1. *([33] p.306) If $\mathcal{M}$ and $\mathcal{N}$, as defined in (5.28), are $G$-invariant subspaces of $B_2$ and $B_0$ respectively, then the Liapunov-Schmidt reduction of $\psi$ is $G$-equivariant.*
2. *([33] p.307) Let $r$ be defined as in (5.36) and $\mathcal{A}$ defined as in (5.48). Then $r$ is $\mathcal{A}$-equivariant.*

The function $r$ is not used explicitly as we proceed. However, the group $\mathcal{A}$ for which $r$ is equivariant is pivotal to the development of the theory that follows. The reason for this is the following relationship between $G$ and $\mathcal{A}$.

Proposition 68. *Let $\mathcal{A}$ be defined as in (5.48) and let $W$ be the matrix whose columns $\{\boldsymbol{w}_i\}_{i=1}^m$ are a basis for $\ker \Psi$. Then $A(g) \in \mathcal{A}$ fixes $\boldsymbol{x} \in \Re^m$ if and only if $g \in G$ fixes $y = W\boldsymbol{x} \in \ker \Psi$.*

*Proof.*

$$
\begin{aligned}
& A(g)\boldsymbol{x} = \boldsymbol{x} \\
\Leftrightarrow \quad & \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} \sum_j a_{1j} x_j \\ \vdots \\ \sum_j a_{mj} x_j \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \\
\Leftrightarrow \quad & \sum_i (\sum_j a_{ij} x_j) \boldsymbol{w}_i = \sum_j x_j \boldsymbol{w}_j \\
\Leftrightarrow \quad & \sum_j x_j \sum_i a_{ij} \boldsymbol{w}_i = \sum_j x_j \boldsymbol{w}_j \\
\Leftrightarrow \quad & \sum_j x_j g \boldsymbol{w}_j = \sum_j x_j \boldsymbol{w}_j \\
\Leftrightarrow \quad & g \sum_j x_j \boldsymbol{w}_j = \sum_j x_j \boldsymbol{w}_j \\
\Leftrightarrow \quad & gW\boldsymbol{x} = W\boldsymbol{x}.
\end{aligned}
$$

□

# CHAPTER 6

## SYMMETRY BREAKING BIFURCATION

Armed with the tools which we developed in the last chapter, we are now ready to determine the bifurcation structure of local solutions to (1.9)

$$\max_{q\in\Delta}(G(q)+\beta D(q))$$

when Assumption 2 is satisfied. We determine this bifurcation structure by applying the theory of the last chapter to the dynamical system (3.18)

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda}\mathcal{L}(q,\lambda,\beta).$$

We consider the equilibria of (3.18) that are fixed by $S_M$. Bifurcations of these equilibria are symmetry breaking bifurcations since the Equivariant Branching Lemma and the Smoller-Wasserman Theorem ascertain the existence of bifurcating branches which have symmetry corresponding to the maximal isotropy subgroups of $S_M$, $M$ of which are the subgroups $S_{M-1}$.

At the conclusion of the chapter, we will have shown that symmetry breaking bifurcations from $S_M$ to $S_{M-1}$ are always pitchfork-like. We will provide conditions which ascertain whether the bifurcating branches are subcritical or supercritical. All subcritical bifurcations are unstable. We also provide a condition which determines whether supercritical branches are stable or unstable. Furthermore, we determine when unstable bifurcating branches contain no solutions to (1.9).

The bifurcation structure of equilibria of the above dynamical system is the bifurcation structure for stationary points of the optimization problem (3.1)

$$\max_{q\in\Delta_\mathcal{E}}(G(q)+\beta D(q))$$

which in turn gives us the bifurcation structure of local solutions to (1.9)

$$\max_{q\in\Delta}(G(q)+\beta D(q)).$$

We point out that in the case when $G(q)$ from (1.9) and (3.1) is strictly concave, as in the case for the Information Distortion method (2.34), then a singularity of the Hessian of (3.18) always gives a bifurcation (Corollary 108), and so one can always apply the bifurcation structure results, which we present in this chapter, to problems of this type (Corollary 117).

The chapter proceeds as follows. We first determine the specific form of the group for which this system is equivariant (Theorem 70), which is isomorphic to $S_N$. We

then determine an explicit basis for the kernel of the Hessian of $\mathcal{L}$ at the bifurcation (Theorems 85 and 87), which enables us to determine the Liapunov-Schmidt reduction of the system ((6.36) and (6.36)). Next, we determine some of the maximal isotropy subgroups of $S_N$ (Lemma 100), and, using these, the existence of bifurcating branches is proved (Theorem 110). Finally, we examine the structure and stability of the branches.

## Notation

Let $(q^*, \lambda^*, \beta^*)$ denote a bifurcation point of (3.18). In the case where $q^* = q_{\frac{1}{N}}$, the uniform solution defined in (2.7), we will use $(q_{\frac{1}{N}}, \lambda^*, \beta^*)$ to denote the corresponding bifurcation point. The following notation will be used throughout the rest of this chapter:

$\Delta F(q_{\frac{1}{N}}) := \Delta F(q_{\frac{1}{N}}, \beta^*)$

$\Delta \mathcal{L}(q_{\frac{1}{N}}) := \Delta_{q,\lambda} \mathcal{L}(q_{\frac{1}{N}}, \lambda^*, \beta^*)$

$\Delta F(q^*) := \Delta F(q^*, \beta^*)$

$\Delta \mathcal{L}(q^*) := \Delta_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta^*)$

$\gamma_{\nu\eta}$ := the element of some Lie group $\Gamma$ which permutes class $\nu \in \mathcal{Y}_N$ with class $\eta \in \mathcal{Y}_N$.

## $M$-uniform Solutions

We now lay the groundwork to prove the existence of bifurcating branches of equilibria of (3.18) from bifurcation of a special set of equilibria, which we define next.

Definition 69. *A stationary point $q^*$ of (3.1) (or, equivalently, an equilibrium $(q^*, \lambda^*)$ of (3.18)) is* $M$-uniform *if there exists an $M$, $1 \leq M \leq N$, and a $K \times 1$ vector $P$ such that $q^{\nu_i} = P$ for $M$ and only $M$ classes, $\{\nu_i\}_{i=1}^{M}$, of $Y_N$. These $M$ classes of $Y_N$ are* unresolved *classes. The classes of $Y_N$ that are not unresolved are* resolved *classes.*

Hence, this section studies bifurcations of $M$-uniform stationary points $q^*$ of (3.1). In this way, we will study symmetry breaking bifurcations of solutions to (1.9). Note that the solution $q_{\frac{1}{N}}$ is $N$-uniform. Much of the discussion that follows addresses this special case.

A particular solution of (3.1), $q^*$, may be both $M_1$-uniform and $M_2$-uniform for some positive numbers $M_1$ and $M_2$ such that $M_1 + M_2 \leq N$. In other words, $q^{\nu_i} = P$

for $\{\nu_i\}_{i=1}^{M_1}$ and $q^{\eta_i} = R$ for $\{\eta_i\}_{i=1}^{M_2}$. For example, for $N = 6$, there exists a solution which bifurcates from $q_{\frac{1}{N}}$ which is 2-uniform and 4-uniform. There also exists a solution which is "twice" 3-uniform. Furthermore, for arbitrary $N$, every $q \in \Delta$ is "at least" 1-uniform. In these instances, the classification of the classes of $Y_N$ as either resolved or unresolved depends upon how one views $q$. If we consider $q$ as $M_1$-uniform, then we call the classes $\{\nu_i\}_{i=1}^{M_1}$ unresolved, and the rest of the $N - M_1$ classes, including the $M_2$ classes $\{\eta_i\}_{i=1}^{M_2}$, are considered resolved. However, if one views $q$ as being $M_2$-uniform, then we call the classes $\{\eta_i\}_{i=1}^{M_2}$ unresolved, and the rest of the $N - M_2$ classes, including the $M_1$ classes $\{\nu_i\}_{i=1}^{M_1}$, are resolved. We allow this flexibility since, as we will see, viewing a stationary point $q^*$ as both $M_1$ and $M_2$ uniform, for $M_1, M_2 > 1$, enables us to consider two different types of symmetry breaking bifurcation from the solution branch which contains $(q^*, \lambda^*, \beta)$.

Suppose that $q^*$ is $M$-uniform. Based on Definition 69, we now partition $\mathcal{Y}_N$ into two disjoint sets. Let

$$\mathcal{U} \text{ be the set of } M \text{ unresolved classes} \tag{6.1}$$

and let

$$\mathcal{R} \text{ be the set of } N - M \text{ resolved classes.} \tag{6.2}$$

Thus $\mathcal{U} \cap \mathcal{R} = \emptyset$ and $\mathcal{U} \cup \mathcal{R} = \{1, ..., N\} = \mathcal{Y}_N$.

Let $B_\nu$ be the block of $\Delta F(q^*)$ corresponding to class $\nu$. For clarity, we denote

$$B = B_\nu \text{ for } \nu \in \mathcal{U} \tag{6.3}$$

and

$$R_\nu = B_\nu \text{ for } \nu \in \mathcal{R}. \tag{6.4}$$

## The Group of Symmetries

The action of "relabelling of the classes of $Y_N$" addressed by Assumption 15.1 is effected by the action of the finite group $S_N$ on the classes of $Y_N$. We now introduce a finite matrix group, which we will call $\Gamma$, which effects the action of "relabelling of the classes of $Y_N$", on the dynamical system (3.18). This introduction comes in two stages. First we introduce the matrix group $\mathcal{P}$, which is isomorphic to $S_N$, which acts on the elements $q \in \Delta$, and on the function $\nabla F$. Then, we can formally define $\Gamma$, also isomorphic to $S_N$, which acts on the elements

$$\begin{pmatrix} q \\ \lambda \end{pmatrix} \in \Re^{NK+K},$$

and on the function $\nabla_{q,\lambda}\mathcal{L}$. It will be convenient to work with the subgroups $S_M \leq S_N$, for $1 < M < N$. Thus, we also present subgroups of $\mathcal{P}$ and of $\Gamma$ which are isomorphic to $S_M$.

We begin by ascertaining which Lie group representation will be used when $S_N$ actions on $q \in \Delta$ and on the function $\nabla F$ are considered. Let

$$\mathcal{P} < O(NK),$$

where $O(NK)$ is a the group of orthogonal matrices in $\Re^{NK}$

$$O(NK) := \{Q \in \Re^{NK \times NK} | QQ^T = I_{NK}\}.$$

$\mathcal{P}$ acts on $q \in \Re^{NK}$ by permuting all the components of $q$ associated with class $\delta$ to class $\eta$. Formally, for $\rho \in \mathcal{P}, \hat{q} = \rho q$ if and only if for each $\delta$, $1 \leq \delta \leq N$, there is an $\eta, 1 \leq \eta \leq N$ such that $\hat{q}_{\delta k} = q_{\eta k}$ for every $k$. In words, $\mathcal{P}$ is the group of *block permutation* matrices. For example, for $N = 3$, $|\mathcal{P}| = 6$, then the elements $\rho_{13}, \rho_{123} \in \mathcal{P}$ are

$$\rho_{13} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & I_K \\ \mathbf{0} & I_K & \mathbf{0} \\ I_K & \mathbf{0} & \mathbf{0} \end{pmatrix}, \rho_{123} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & I_K \\ I_K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_K & \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is $K \times K$. Hence, Assumption 15.1, which states that $G$ and $D$ must be invariant to the relabelling of the classes of $Y_N$, is equivalent to saying that $G$ and $D$ are $\mathcal{P}$-invariant.

A word of caution is in order. The theory of chapter 5 can not be applied directly to $q \in \Delta$ or to $\nabla F$ for two reasons. First, $\Delta$ is not a Banach space ($\Delta$ is not closed under vector addition, and it does not contain the vector $\mathbf{0}$). Secondly, the theory can describe bifurcations of equilibria to

$$\dot{q} = \nabla F(q, \beta),$$

but these equilibria correspond to solutions of the unconstrained problem

$$\max_{q \in \Re^{NK}} (G(q) + \beta D(q)),$$

which are not even stationary points of our problem (3.1).

Now we define the group that acts on the system (3.18) (i.e. on $(q, \lambda) \in \Re^{NK} \times \Re^K$, and on $\nabla_{q,\lambda}\mathcal{L}$). Let $\Gamma \leq O(NK + K)$ such that

$$\Gamma := \left\{ \begin{pmatrix} \rho & \mathbf{0}^T \\ \mathbf{0} & I_K \end{pmatrix} | \text{ for } \rho \in \mathcal{P} \right\}. \tag{6.5}$$

Observe that $\gamma \in \Gamma$ acts on $\nabla_{q,\lambda}\mathcal{L}$ by

$$\gamma \nabla_{q,\lambda}\mathcal{L}(q, \lambda) = \begin{pmatrix} \rho & \mathbf{0}^T \\ \mathbf{0} & I_K \end{pmatrix} \begin{pmatrix} \nabla_q \mathcal{L} \\ \nabla_\lambda \mathcal{L} \end{pmatrix} = \begin{pmatrix} \rho \nabla_q \mathcal{L} \\ \nabla_\lambda \mathcal{L} \end{pmatrix} \tag{6.6}$$

and on $\begin{pmatrix} q \\ \lambda \end{pmatrix}$ by

$$\gamma \begin{pmatrix} q \\ \lambda \end{pmatrix} = \begin{pmatrix} \rho q \\ \lambda \end{pmatrix}. \tag{6.7}$$

Thus, $\gamma \in \Gamma$ acts on $q \in \Re^{NK}$ as defined by $\rho \in \mathcal{P}$ but leaves the Lagrange multipliers $\lambda = (\lambda_1 \ \ \lambda_2, ..., \lambda_K)^T$ fixed.

We have the following theorem.

THEOREM 70. *$\mathcal{L}(q, \lambda, \beta)$ is $\Gamma$-invariant, $\nabla_{q,\lambda}\mathcal{L}(q, \lambda, \beta)$ is $\Gamma$-equivariant, and $\nabla F$ is $\mathcal{P}$-equivariant.*

*Proof.* By definition of the group $\Gamma$, we see that $\mathcal{L}$ is $\Gamma$-invariant since $F$ is $\mathcal{P}$-invariant. Differentiating both sides of the equation $\mathcal{L}(q, \lambda, \beta) = \mathcal{L}(\gamma \begin{pmatrix} q \\ \lambda \end{pmatrix}, \beta)$ for any $\gamma \in \Gamma$ shows that $\nabla_{q,\lambda}\mathcal{L} = \gamma^T \nabla_{q,\lambda}\mathcal{L}(\gamma \begin{pmatrix} q \\ \lambda \end{pmatrix}, \beta)$. Since $\gamma^{-1} = \gamma^T$, this shows that $\nabla_{q,\lambda}\mathcal{L}$ is $\Gamma$-equivariant. A similar argument shows that $\nabla F$ is $\mathcal{P}$-equivariant. □

For the Information Distortion problem (2.34), Theorem 73 below shows explicitly that $\nabla_{q,\lambda}\mathcal{L}$ is $\Gamma$-equivariant, where $\Gamma$ is defined in (6.5).

The maximal isotropy subgroup structure of $\Gamma$ gives the existence of bifurcating branches from $q_{\frac{1}{N}}$ because $\Gamma$ fixes no nontrivial vector other than multiples of $q_{\frac{1}{N}}$ in $\ker \Delta\mathcal{L}(q_{\frac{1}{N}})$ (see Theorem 47, Theorem 49 and Proposition 104). To get the existence of bifurcating branches from an $M$-uniform solution $q^* \neq q_{\frac{1}{N}}$, we determine a subgroup of $\Gamma$ which fixes no nontrivial vector in $\ker \Delta\mathcal{L}(q^*)$ other than multiples of $q^*$ under generic assumptions (see Proposition 105). With this in mind, we now define such a subgroup of $\Gamma$ (Theorem 71) which is isomorphic to the subgroups $S_M < S_N$ for $1 < M < N$.

The set $Y_N$ is the set of $N$ objects on which $S_N$ acts. Viewed as a subgroup of $S_N$, $S_M$ is the group of permutations on only $M$ of the elements of $Y_N$. The other $N - M$ elements of $Y_N$ are left fixed by the action of $S_M$. Thus, to determine a particular instance of a subgroup $S_M \leq S_N$, one must first determine which $M$ elements of $Y_N$ are permuted freely by $S_M$. Given an $M$-uniform solution, we are interested in the subgroup $S_M \leq S_N$ which permutes the $M$ unresolved classes of $\mathcal{U} \subseteq Y_N$, and leaves fixed the resolved classes $Y_N \setminus \mathcal{U} = \mathcal{R}$ of $Y_N$. Define

$$\Gamma_{\mathcal{U}} := \left\{ \begin{pmatrix} \rho & \mathbf{0}^T \\ \mathbf{0} & I_K \end{pmatrix} | \rho \in \mathcal{P}_{\mathcal{U}} \right\}, \tag{6.8}$$

where $\rho$ is $NK \times NK$, $I_K$ is a $K \times K$ identity, and $\mathbf{0}$ is $K \times NK$. The elements of the subgroup

$$\mathcal{P}_{\mathcal{U}} < \mathcal{P} \tag{6.9}$$

(from (6.5)) fix the classes of $\mathcal{R}$, and freely permute the $M$ classes of $\mathcal{U}$. Thus, $\mathcal{P}_{\mathcal{U}}$ and $\Gamma_{\mathcal{U}}$ are Lie groups isomorphic to $S_M$. If $\mathcal{U} = Y_N$, then we are back to the case where $q^* = q_{\frac{1}{N}}$ and $\Gamma_{\mathcal{U}} = \Gamma$.

THEOREM 71. *Let $|\mathcal{U}| = M$. Then $q \in \mathrm{Fix}(\mathcal{P}_{\mathcal{U}})$ if and only if $q$ is $M$-uniform.*

*Proof.* Recall that $\mathcal{P}_{\mathcal{U}}$ from (6.9), which is isomorphic to $S_M$, permutes the classes of $\mathcal{U}$ and fixes the classes of $\mathcal{R}$. Let $\rho_{\nu\eta} \in \mathcal{P}_{\mathcal{U}}$. If $q \in \mathrm{Fix}(\mathcal{P}_{\mathcal{U}})$, then $\rho_{\nu\eta} q = q$ which implies that $q^\nu = q^\eta$ for every $\nu$ and $\eta$ in $\mathcal{U}$, which shows that $q$ is $M$-uniform.

Now suppose that $q$ is $M$-uniform, which means that $q^\nu = q^\eta$ for every $\nu, \eta \in \mathcal{U}$. Then $\rho_{\nu\eta} q = \hat{q}$, where

$$\hat{q}^c = \begin{cases} q^\nu \text{ if } c = \eta \\ q^\eta \text{ if } c = \nu \\ q^c \text{ otherwise} \end{cases} .$$

This shows that $\rho_{\nu\eta} q = q$. The theorem now follows from Proposition 76.1 since $\mathcal{P}_{\mathcal{U}}$ is generated by the transpositions $\{\rho_{\nu\eta}\}$ over all $\nu, \eta \in \mathcal{U}$. □

The fact that $q \in \Delta$ can be both $M_1$-uniform and $M_2$-uniform for $M_1 \neq M_2$ shows that $q \in \mathrm{Fix}(\mathcal{P}_{\mathcal{U}_1}) \cap \mathrm{Fix}(\mathcal{P}_{\mathcal{U}_2})$ for two distinct subsets of $\mathcal{U}_1, \mathcal{U}_2 \subseteq Y_N$.

One of the basic assumptions on which this thesis relies is that $\Delta F(q^*)$ is block diagonal (Assumption . Another basic but crucial observation about $M$-uniform solution is the following theorem.

THEOREM 72. *If $q \in \mathrm{Fix}(\mathcal{P}_{\mathcal{U}})$ where $|\mathcal{U}| = M$, then $\Delta F$ has $M$ identical blocks.*

*Proof.* Let $\hat{q} \in \mathrm{Fix}(\mathcal{P}_{\mathcal{U}})$. Let $\rho$ be the transposition in $P_{\mathcal{U}}$ which permutes the classes $\nu$ and $\eta$ in $\mathcal{U}$, which exists since $\mathcal{P}_{\mathcal{U}} \cong S_M$ . By Theorem 70, we have that $\nabla F(\rho q) = \rho \nabla F(q)$, and now differentiation and evaluating at $q = \hat{q}$ yields $\Delta F(\hat{q})\rho = \rho \Delta F(\hat{q})$. Thus, using (3.9), we see that $B = B_\eta = B_\nu$. Since $\nu$ and $\eta$ are arbitrary classes of $\mathcal{U}$, then it must be that $B = B_\nu$ for every $\nu \in \mathcal{U}$. □

The converse to the theorem does not hold. To see this, consider the Information Distortion problem (2.34), so that $G = H(Y_N|Y)$ and $D = D_{eff}$. Observe that the $(m,n)^{th}$ component of the $\nu^{th}$ block of $\Delta D_{eff}(q)$ is

$$[\Delta D^\nu(q)]_{mn} = \sum_i \frac{p(x_i, y_m)p(x_i, y_n)}{\sum_k q_{\nu k} p(x_i, y_k)} - \frac{p(y_m)p(y_n)}{\sum_k q_{\nu k} p(y_k)}.$$

For $N = 2$ and some $a$ such that $0 < a < 1$, let $\hat{q}$ be identically $\frac{1}{2}$ except for

$$\hat{q}(\nu = 1|y = 1) = q(\nu = 2|y = 2) = a$$

and

$$\hat{q}(\nu = 2|y = 1) = q(\nu = 1|y = 2) = 1 - a.$$

If $p(X, y_1) = p(X, y_2) = \mathbf{0}$, then $p(y_1) = p(y_2) = 0$ and so $\Delta D^1(\hat{q}) = \Delta D^2(\hat{q})$. This also shows that the corresponding components of $\Delta H(Y_N|Y)$ are zero (see (2.20)). Thus, $\Delta F(\hat{q}, \beta)$ has identical blocks, but $\hat{q} \notin \text{Fix}(\mathcal{P})$.

Now for the result that deals specifically with (3.18) when $F$ is defined as in Information Distortion Problem (2.34), which was promised at the beginning of this section.

THEOREM 73. *When $F$ is defined as in (2.34), $\nabla_{q,\lambda}\mathcal{L}$ is $\Gamma$-equivariant.*

*Proof.* Since the transpositions generate $S_N$ (Proposition 76.1), then it just needs to be shown that for each transposition $\gamma_{\delta\eta} \in \Gamma$ which permutes $q_{\delta k}$ with $q_{\eta k}$ for all $k$, $1 \leq k \leq K$

$$\gamma_{\delta\eta}\nabla_{q,\lambda}\mathcal{L}(q, \lambda, \beta) = \nabla_{q,\lambda}\mathcal{L}(\gamma_{\delta\eta}\begin{pmatrix} q \\ \lambda \end{pmatrix}, \beta).$$

By (6.6) and (6.7), this equation becomes

$$\begin{pmatrix} \rho\nabla_q\mathcal{L}(q, \lambda, \beta) \\ \nabla_\lambda\mathcal{L}(q, \lambda, \beta) \end{pmatrix} = \begin{pmatrix} \nabla_q\mathcal{L}(\rho q, \lambda, \beta) \\ \nabla_\lambda\mathcal{L}(\rho q, \lambda, \beta) \end{pmatrix},$$

where $\rho \in \mathcal{P}$ permutes class $\delta$ with class $\eta$. By (3.4), this requirement becomes

$$\rho\nabla F(q, \beta) = \nabla F(\rho q, \beta) \tag{6.10}$$

$$\nabla_\lambda\mathcal{L}(q, \lambda, \beta) = \nabla_\lambda\mathcal{L}(\rho q, \lambda, \beta). \tag{6.11}$$

We show (6.10) by showing that each term of $\nabla F$ is in fact $\Gamma$-equivariant, which is really just a practice in subscripts. First, we consider

$$\begin{aligned} [\rho\nabla H]_{\nu k} &= \begin{cases} -p(y_k)(\log q_{\nu k} + 1) & \text{if } \nu \notin \{\delta, \eta\} \\ -p(y_k)(\log q_{\eta k} + 1) & \text{if } \nu = \delta \\ -p(y_k)(\log q_{\delta k} + 1) & \text{if } \nu = \eta \end{cases} \\ &= [\nabla H(\rho q)]_{\nu k}. \end{aligned}$$

Thus, $\rho\nabla H(q) = \nabla H(\rho q)$. Lastly we consider

$$\begin{aligned} [\rho\nabla D_{eff}]_{\nu k} &= \begin{cases} \sum_i p(x_i, y_k) \log \frac{\sum_k q_{\nu k} p(x_i, y_k)}{p(x_i)\sum_k q_{\nu k} p(y_k)} & \text{if } \nu \notin \{\delta, \eta\} \\ \sum_i p(x_i, y_k) \log \frac{\sum_k q_{\eta k} p(x_i, y_k)}{p(x_i)\sum_k q_{\eta k} p(y_k)} & \text{if } \nu = \delta \\ \sum_i p(x_i, y_k) \log \frac{\sum_k q_{\delta k} p(x_i, y_k)}{p(x_i)\sum_k q_{\delta k} p(y_k)} & \text{if } \nu = \eta \end{cases} \\ &= [\nabla D_{eff}(\rho q)]_{\nu k}. \end{aligned}$$

Hence, $\rho\nabla D_{eff}(q) = \nabla D_{eff}(\rho q)$.

To get (6.11), we use (3.6), which implies that

$$\begin{aligned} [\nabla_\lambda\mathcal{L}(q, \lambda, \beta)]_k = \sum_\nu q_{\nu k} - 1 &= q_{1k} + q_{2k} +, ..., + q_{Nk} - 1 \\ &= [\nabla_\lambda\mathcal{L}(\rho q, \lambda, \beta)]_k, \end{aligned}$$

where the last equality follows since permuting $q_{\delta k}$ with $q_{\eta k}$ leaves the sum $\sum_\nu q_{\nu k}$ unchanged. □

## The Group $S_M$

In this section we examine the abstract group

$$\mathcal{S}, \text{ the cycle representation of } S_M \tag{6.12}$$

as defined in [8] and [27] for arbitrary $M$, which will prove useful in the sequel. We use the notation $\gamma_{(a_1 a_2 ... a_m)}$ to denote an element of $\Gamma$ which is isomorphic to the $M$-cycle $(a_1 a_2 ... a_m) \in \mathcal{S}$.

We will be working extensively with the subgroups of $S_M$, and define a well studied normal subgroup of $S_M$ next.

DEFINITION 74. *The* alternating group *on $M$ symbols, $A_M$, is the subgroup of all elements of $S_M$ which can be decomposed into an even number of transpositions.*

REMARK 75. *We will use* four *different group representations of $S_M$ in the discussion that follows. The first two representations, $\mathcal{P}_\mathcal{U}$ and $\Gamma_\mathcal{U}$, have just been described in (6.9) and (6.8) respectively. The latter two, $\mathcal{S}$ and $\mathcal{A}$, are defined at (6.12) and (6.45) in the text respectively.*

PROPOSITION 76. *[27] Using the cycle representation $\mathcal{S}$ from (6.12):*

1. *(p.107) $S_M$ is generated by transpositions:*

$$S_M =< T > \text{ where } T = \{(ij) | 1 \leq i < j \leq M\}.$$

2. *(p.28-31,116) Any element of $S_M$ can be written uniquely as a product of disjoint cycles.*

3. *(p.31) An element in $S_M$ is of order $M$ if and only if it is an $M$-cycle.*

4. *(p.110) An element $\sigma \in S_M$ is an $m$-cycle where $m$ is odd if and only if $\sigma \in A_M$.*

5. *(p.127) If $\tau, \sigma \in S_M$ with*

$$\sigma = \Pi_i (a_{i1} ... a_{im_i})$$

*then*

$$\tau \sigma \tau^{-1} = \Pi_i (\tau(a_{i1}) \tau(a_{i2}) ... \tau(a_{im_i}))$$

6. *(p.127) $\sigma$ and $\tau \in S_M$ are conjugate $\Leftrightarrow$ they have the same cycle decomposition. In other words, for any $\sigma, \tau \in S_M$ of a given cycle type, there exists $\zeta \in S_M$ such that $\zeta \tau \zeta^{-1} = \sigma$.*

The next result ascertains some of the maximal subgroups of $S_M$. Liebeck et al. [46] show that the next Theorem gives but one of six different types of the maximal subgroups of $S_M$ for arbitrary $M$.

THEOREM 77. *For any finite $M > 1$, among the maximal subgroups of $S_M$, there are $M$ subgroups which are isomorphic to $S_{M-1}$.*

*Proof.* Using the cycle representation $\mathcal{S}$ of $S_M$ from (6.12), Lemma 76.1 gives that $S_M = \langle T \rangle$ where $T = \{(ij) | 1 \leq i < j \leq M\}$. Consider the set $T_k \subset T$

$$T_k := \{(ij) \in T | i, j \neq k\}. \tag{6.13}$$

It is clear that $\langle T_k \rangle \cong S_{M-1}$. Suppose $\langle T_k \rangle < H \leq S_M$. The theorem is proved if $H = S_M$. Note that $H$ must have some element $\sigma$ which acts on $k$ non-trivially (otherwise, $\langle T_k \rangle = H$). Write $\sigma$ (uniquely) as a product of disjoint cycles (Proposition 76.2)

$$\sigma = \sigma_1 \sigma_2 ... \sigma_m$$

where $m \geq 1$. Then $k$ is contained in some cycle, say $\sigma_l$, for some $l$ where $1 \leq l \leq m$. So $k$ is not in any other cycle $\sigma_n$, for $n \neq l$. This implies $\sigma_n \in \langle T_k \langle < H$ and so $\sigma_n^{-1} \in H$. Therefore, we can multiply $\sigma_1 \sigma_2 ... \sigma_m$ on the left by $\sigma_{l-1}^{-1} \sigma_{l-2}^{-1} ... \sigma_1^{-1}$ and on the right by $\sigma_m^{-1} \sigma_{m-1}^{-1} ... \sigma_l^{-1}$ to show that $\sigma_l \in H$. Now we repeat this strategy: if $\sigma_l = (a_1 a_2 ... k ... a_{p-1} a_p)$ then $\sigma_l$ can be rewritten as

$$\sigma_l = (a_1 a_p)(a_1 a_{p-1}) ... (a_1 k) ... (a_1 a_3)(a_1 a_2)$$

where $a_r \neq k$ for each $r$, $1 \leq r \leq p$. Hence, $(a_1 a_r) \in \langle T_k \rangle < H$ and so we see that $(a_1 k) \in H$ after multiplying on the left and the right of $\sigma_l$ by the appropriate inverses. Now, $(a_1 j) \in \langle T_k \rangle$ for every $j \neq k$ from which it follows that $(a_1 j)(a_1 k)(a_1 j) = (jk) \in H$. Hence, $H$ contains $T$ which implies that $H = S_M$ and so $\langle T_k \rangle$ is a maximal subgroup of $S_M$ for each $k \in \{1, 2, ..., M\}$. □

## The Initial Solution $q_0$

We now examine the solution $q_0$ of (3.1)

$$q_0 = \text{argmax}_{q \in \Delta} F(q, 0) = \text{argmax}_{q \in \Delta} G(q).$$

We show that under some conditions, $q_0$ persists as a solution for $\beta \in [0, \tilde{\beta})$, and then we show that this result holds for the Information Distortion problem (2.34). We conclude the section by providing the location of singularities along the solution branch which contains the initial solution $q_0$.

LEMMA 78. *If $q_0$ is a stationary point of (3.1) for all $\beta \in [0, \hat{\beta})$ for some $\hat{\beta} > 0$, and if $\Delta G(q_0)$ is negative definite on $\ker J$, then $q_0$ is a solution of (3.1) for all $\beta \in [0, \tilde{\beta})$ for some $0 < \tilde{\beta} < \hat{\beta}$.*

*Proof.* Since $q_0$ is a solution of (3.1) at $\beta = 0$, then there exists some vector $\lambda_0$ such that $\nabla_{q,\lambda}\mathcal{L}(q_0, \lambda_0, 0) = \mathbf{0}$. If we let $Z$ be defined as the $NK \times (\dim\ker J)$ matrix whose columns span $\ker J$, then, by the assumption on $\Delta G(q_0)$ and Remark 21.2, the eigenvalues of $Z^T\Delta G(q_0, 0)Z = Z^T\Delta F(q_0, 0)Z = Z^T\Delta_q\mathcal{L}(q_0, \lambda_0, 0)Z$ are negative and bounded away from zero. Since $\Delta F$ changes continuously in $\beta$ (Assumption 15.2), then $Z^T\Delta F(q_0, \beta)Z$ has negative eigenvalues for every $0 < \beta < \tilde{\beta}$ for some $\tilde{\beta} < \hat{\beta}$. Applying Theorem 20 completes the proof. □

Theorem 71 shows that for any problem (3.1), $q_0$ is fixed by the action of the full group $\mathcal{P}$ if and only if $q = q_{\frac{1}{N}}$, where $q_{\frac{1}{N}}$ is the uniform quantizer defined in (2.7). For $F$ as defined for the Information Distortion and the Information Bottleneck cost functions, (2.34) and (2.35), $q_0 = q_{\frac{1}{N}}$. In both cases, $q_{\frac{1}{N}}$ is a solution to (3.1) for all $\beta$ in some $[0, \tilde{\beta})$. We prove this claim for (2.34) in the following lemma.

LEMMA 79. *$(q_{\frac{1}{N}}, \beta)$ is a solution of (2.34) for all $\beta \in [0, \tilde{\beta})$ for some $\tilde{\beta} > 0$.*

*Proof.* Consider

$$\max_{q\in\Delta_{\mathcal{E}}} H(q). \tag{6.14}$$

Now the Lagrangian (3.3) becomes

$$\mathcal{L}(q, \lambda) = H(q) + \sum_k \lambda_k(\sum_\nu q_{\nu k} - 1).$$

By Theorem 16 and (2.19), solutions $\tilde{q}$ of (6.14) are determined by considering solutions of

$$\nabla_q\mathcal{L}_{\nu k} = -p(y_k)\left(\log_2 q_{\nu k} + \frac{1}{\ln 2}\right) + \lambda_k = 0 \tag{6.15}$$

$$\nabla_\lambda\mathcal{L}_k = \sum_\nu q_{\nu k} - 1 = 0. \tag{6.16}$$

From (6.15), $\log_2 q_{\nu k} = \frac{\lambda_k}{p(y_k)} - \frac{1}{\ln 2}$ from which it follows $q_{\nu k} = 2^{\frac{\lambda_k}{p(y_k)} - \frac{1}{\ln 2}}$. From (6.16),

$$1 = \sum_\nu q_{\nu k} = \sum_\nu 2^{\frac{\lambda_k}{p(y_k)} - \frac{1}{\ln 2}} = N2^{\frac{\lambda_k}{p(y_k)} - \frac{1}{\ln 2}}$$

which implies $\lambda_k = p(y_k)\left(\frac{1}{\ln 2} - \log_2 N\right)$. Substituting this last expression for $\lambda_k$ back into (6.15) proves that $q_{\nu k} = \frac{1}{N}$ for every $\nu$ and $k$ satisfies the KKT conditions. Since $\Delta H(q)$ is negative definite for every $q$ (see (2.20)), then $q_{\frac{1}{N}}$ is the global solution of (6.14) by Theorem 20.

Since $\nabla H(q_{\frac{1}{N}}) + \beta\nabla D_{eff}(q_{\frac{1}{N}}) = \mathbf{0}$ for every Thus$\beta$, then $q_{\frac{1}{N}}$ is a stationary point of (2.34) for every $\beta$. The Lemma now follows from Lemma 78 since $\Delta H(q)$ is negative definite for every $q \in \Delta$. □

Compare the result of the last Lemma to the unconstrained problem

$$\max_{q \in \Re^{NK}} H(q), \tag{6.17}$$

where $H$ is the entropy function from (2.34). Since $\Delta H$ is negative definite, then the unique point that satisfies $\nabla H = \mathbf{0}$ (see (2.19)) is the global maximum:

$$\begin{aligned} \nabla H = \mathbf{0} &\Leftrightarrow -p(y_k)\left(\log_2 q_{\nu k} + \frac{1}{\ln 2}\right) = 0 \\ &\Leftrightarrow \log_2 q_{\nu k} = -\frac{1}{\ln 2} \\ &\Leftrightarrow \ln q_{\nu k} = -1 \\ &\Leftrightarrow q_{\nu k} = \frac{1}{e}. \end{aligned}$$

Hence, for arbitrary $N$, the constrained maximum $q_{\frac{1}{N}}$ of (6.14) is not even a stationary point of (6.17).

By Theorem 72, if $q_0 = q_{\frac{1}{N}} \in \mathrm{Fix}(\mathcal{P})$, then all of the blocks of $\Delta F(q_0, \beta)$ are identical. Thus, the blocks $\{B_i\}_{i=1}^N$ (from (3.9)) of $\Delta F(q_0, \beta) = \Delta F(q_{\frac{1}{N}}, \beta)$ can be written as

$$B_i = B. \tag{6.18}$$

Consider the branch of equilibria $(q_{\frac{1}{N}}, \lambda^*, \beta)$ to (3.18) for $0 \leq \beta \leq \hat{\beta}$. If the hypotheses of Lemma 78 are met, and if $\Delta G(q_{\frac{1}{N}})$ is nonsingular, then one can ascertain the values of $\beta$ at which bifurcation occurs along this branch by solving an eigenvalue problem. In particular, this result holds for the Information Distortion problem (2.34).

THEOREM 80. *Suppose that $q_{\frac{1}{N}}$ is a stationary point of (3.1) for all $\beta \in [0, \hat{\beta})$ for some $\hat{\beta} > 0$, and that $\Delta G(q_{\frac{1}{N}})$ is negative definite on $\ker J$. Further suppose that $\Delta G(q_{\frac{1}{N}})$ is nonsingular. Then the bifurcation from the solution $(q_{\frac{1}{N}}, \beta)$ can only occur at the reciprocal of the eigenvalues of $-\Delta G^{-1}(q_{\frac{1}{N}})\Delta D(q_{\frac{1}{N}})$.*

*Proof.* By Theorem 24, $\Delta \mathcal{L}(q_{\frac{1}{N}})$ is singular at bifurcation. By Theorems 70 and 72, $\Delta F(q_{\frac{1}{N}})$ has identical blocks. Thus, by Corollary 35, $\Delta F(q_{\frac{1}{N}})$ is singular,

$$\det(\Delta F(q_{\frac{1}{N}}, \beta) = \det(\Delta G(q_{\frac{1}{N}}) + \beta \Delta D(q_{\frac{1}{N}})) = 0,$$

at bifurcation. By assumption, $\Delta G(q_{\frac{1}{N}})$ is nonsingular, and hence invertible so that

$$\frac{1}{-\det(\Delta G(q_{\frac{1}{N}}))} \det\left(\Delta G(q_{\frac{1}{N}}) + \beta \Delta D(q_{\frac{1}{N}})\right) = 0$$

from which it follows that

$$\det\left(-\Delta G(q_{\frac{1}{N}})^{-1})\Delta D(q_{\frac{1}{N}}) - \frac{1}{\beta}I\right) = 0$$

which is the eigenvalue problem for the matrix $-\Delta G^{-1}(q_{\frac{1}{N}})\Delta D(q_{\frac{1}{N}})$. □

Kernel of the Hessian at Symmetry Breaking Bifurcation

Bifurcation of equilibria of (3.18)

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda}\mathcal{L}(q,\lambda,\beta)$$

at a point $(q^*,\lambda^*,\beta^*)$ causes the Jacobian of the system, $\Delta\mathcal{L}(q^*)$, to be singular (Theorem 24). As we have seen, the bifurcating directions are contained in $\ker\Delta\mathcal{L}(q^*)$, the kernel of the Hessian of the Lagrangian (3.3) (see (5.35) and (5.38)).

The purpose of this section is to determine a basis for $\ker\Delta\mathcal{L}(q^*)$ at symmetry breaking bifurcation of an $M$-uniform solution $(q^*,\lambda^*,\beta^*)$, given that the following assumptions are met.

Assumption 81.

1. *$q^*$ is $M$-uniform, for $1 < M \leq N$.*
2. *For $B$, the block(s) of the Hessian defined in (6.3),*

$$\ker B \text{ has dimension } 1 \text{ with } K\times 1 \text{ basis vector } \boldsymbol{v} \tag{6.19}$$

3. *For $\{R_\nu\}$, the block(s) of the Hessian defined in (6.4), we have*

$$R_\nu \text{ is nonsingular for every } \nu\in\mathcal{R}. \tag{6.20}$$

4. *The matrix $B\sum_\nu R_\nu^{-1} + MI_K$ is nonsingular.*

Observe that Theorem 72 guarantees that the blocks of the Hessian have the structure presupposed by Assumptions 81.2 and 81.3. When $q^*$ is $N$-uniform, then all of the blocks of the Hessian are identical as in (6.18).

In chapter 8, we examine the type of bifurcation to be expected when Assumption 81 does not hold.

Remark 82. *For the Information Bottleneck problem (2.35),*

$$\max_{q\in\Delta} F(q,\beta) = \max_{q\in\Delta}(I(Y;Y_N) + \beta I(X,Y_N)),$$

*Assumption 81.3 is never satisfied. This is due to the fact that $q$ is always in the kernel of $\Delta F(q,\beta)$ for every $\beta$ (Theorem 43). This implies that the kernel of the $\nu^{th}$ block of $\Delta F(q^*)$ contains the $K\times 1$ vector $[q^*]_\nu$ at bifurcation $(q^*,\beta^*)$ in addition to the vector $\boldsymbol{v}$ from Assumption 81.2. We comment on this scenario in the section at the end of this chapter.*

We begin by determining a basis for $\ker \Delta F(q^*)$. Define the $NK\times 1$ vectors

$$\{\boldsymbol{v}_i\}_{i=1}^{M}$$

by

$$[\boldsymbol{v}_i]_\nu := \begin{cases} \boldsymbol{v} \text{ if } \nu \text{ is the } i^{th} \text{ unresolved class of } \mathcal{U} \\ \mathbf{0} \text{ otherwise} \end{cases} \tag{6.21}$$

where $\mathbf{0}$ is $K\times 1$, which are clearly linearly independent. From Assumptions 81.2 and 81.3, we get that $\dim\ker \Delta F(q^*) = M$. This shows the following:

LEMMA 83. *$\{\boldsymbol{v}_i\}_{i=1}^{M}$ as defined in (6.21) is a basis for* $\ker \Delta F(q^*)$.

Thus, if $q^* = q_{\frac{1}{N}}$ then $\ker \Delta F(q_{\frac{1}{N}})$ has dimension $N$ with $NK\times 1$ basis vectors

$$\boldsymbol{v}_1 = \begin{pmatrix} \boldsymbol{v} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \boldsymbol{v}_2 = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{v} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \dots, \boldsymbol{v}_N = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \boldsymbol{v} \end{pmatrix}. \tag{6.22}$$

Now, let

$$\boldsymbol{w}_i = \begin{pmatrix} \boldsymbol{v}_i \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{v}_M \\ \mathbf{0} \end{pmatrix} \tag{6.23}$$

for $i = 1, \dots, M-1$ where $\mathbf{0}$ is $K\times 1$. For example, if $M = N-1$ and $\mathcal{R} = \{2\}$, then $\{\boldsymbol{w}_i\}_{i=1}^{M-1} =$

$$\underbrace{\left\{\begin{pmatrix} \boldsymbol{v} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ -\boldsymbol{v} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \boldsymbol{v} \\ \vdots \\ \mathbf{0} \\ -\boldsymbol{v} \\ \mathbf{0} \end{pmatrix}, \dots, \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \boldsymbol{v} \\ -\boldsymbol{v} \\ \mathbf{0} \end{pmatrix}\right\}}_{N-2 \text{ vectors}}.$$

LEMMA 84. *Given that Assumption 81 holds,* $\{\boldsymbol{w}_i\}_{i=1}^{M-1}$ *from (6.23) are linearly independent vectors of* $\ker \Delta\mathcal{L}(q^*)$.

*Proof.* To show $\{\boldsymbol{w}_i\} \in \ker \Delta\mathcal{L}(q^*)$, compute

$$\begin{aligned}
\Delta\mathcal{L}(q^*)\boldsymbol{w}_i &= \begin{pmatrix} \Delta F(q^*) & J^T \\ J & \mathbf{0} \end{pmatrix} \left( \begin{pmatrix} \boldsymbol{v}_i \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{v}_M \\ \mathbf{0} \end{pmatrix} \right) \\
&= \begin{pmatrix} \Delta F(q^*)\boldsymbol{v}_i \\ J\boldsymbol{v}_i \end{pmatrix} - \begin{pmatrix} \Delta F(q^*)\boldsymbol{v}_M \\ J\boldsymbol{v}_M \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{0} \\ J\boldsymbol{v}_i \end{pmatrix} - \begin{pmatrix} \mathbf{0} \\ J\boldsymbol{v}_M \end{pmatrix} \text{ (by Lemma 83)} \\
&= \begin{pmatrix} \mathbf{0} \\ [I_K \quad I_K \quad ... \quad I_K]\boldsymbol{v}_i \end{pmatrix} - \begin{pmatrix} \mathbf{0} \\ [I_K \quad I_K \quad ... \quad I_K]\boldsymbol{v}_M \end{pmatrix} \text{ (by (3.7))} \\
&= \begin{pmatrix} \mathbf{0} \\ \boldsymbol{v} \end{pmatrix} - \begin{pmatrix} \mathbf{0} \\ \boldsymbol{v} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}.
\end{aligned}$$

To get linear independence, suppose there exists $c_i \in \Re$ for $i = 1, ..., M-1$ such that

$$\sum_{i=1}^{M-1} c_i \boldsymbol{w}_i = \mathbf{0}.$$

Then

$$\sum_{i=1}^{M-1} (c_i \boldsymbol{v}_i - c_i \boldsymbol{v}_M) = \sum_{i=1}^{M-1} c_i \boldsymbol{v}_i - \sum_{i=1}^{M-1} c_i \boldsymbol{v}_M = \mathbf{0}. \tag{6.24}$$

Set

$$d_i = c_i \text{ for } i = 1, ..., M-1 \text{ and } d_M = -\sum_{i=1}^{M-1} c_i. \tag{6.25}$$

Then (6.24) and (6.25) imply that

$$\sum_{i=1}^{M} d_i \boldsymbol{v}_i = \mathbf{0}.$$

By Lemma 83, $d_i = 0$ for every $i$, from which it follows that $c_i = 0$ for every $i$. □

Now we are ready to prove the main results of this section.

THEOREM 85. $\{\boldsymbol{w}_i\}_{i=1}^{N-1}$ *is a basis for* $\ker \Delta\mathcal{L}(q_{\frac{1}{N}})$.

*Proof.* By Lemma 84, $\ker \Delta\mathcal{L}(q_{\frac{1}{N}}) \supseteq \text{span}\{\boldsymbol{w}_i\}$. To get the other containment, let $\boldsymbol{k} \in \ker \Delta\mathcal{L}(q_{\frac{1}{N}})$ and decompose it as in (4.1) and (4.6). Since the blocks of $\Delta F(q_{\frac{1}{N}})$ are identical (see (6.18)), $\boldsymbol{k} \in \ker \Delta\mathcal{L}(q_{\frac{1}{N}})$ if and only if

$$\begin{pmatrix} B\boldsymbol{x}_1 \\ B\boldsymbol{x}_2 \\ \vdots \\ B\boldsymbol{x}_N \end{pmatrix} = - \begin{pmatrix} \boldsymbol{k}_J \\ \boldsymbol{k}_J \\ \vdots \\ \boldsymbol{k}_J \end{pmatrix}$$

(see (4.7)). Equation (4.8) implies that $B\sum_\nu \boldsymbol{x}_\nu = -\sum_\nu \boldsymbol{k}_J = \boldsymbol{0}$ from which we get $\boldsymbol{k}_J = \boldsymbol{0}$. Hence $\boldsymbol{k} = \begin{pmatrix} \boldsymbol{k}_F \\ \boldsymbol{0} \end{pmatrix}$ and (4.2) assures that $\boldsymbol{k}_F \in (\ker \Delta F(q_{\frac{1}{N}})) \cap (\ker J)$. Therefore $\boldsymbol{k}_F = \sum_i c_i \boldsymbol{v}_i$ (Lemma 83) and $J\boldsymbol{k}_F = \boldsymbol{0}$. The last equation can be written as

$$J\boldsymbol{k}_F = J \begin{pmatrix} c_1\boldsymbol{v} \\ c_2\boldsymbol{v} \\ \vdots \\ c_N\boldsymbol{v} \end{pmatrix} = \boldsymbol{0}$$

from which $\sum_i c_i \boldsymbol{v} = \boldsymbol{v} \sum_i c_i = \boldsymbol{0}$. Therefore $\sum_i c_i = 0$ and so

$$c_N = -\sum_{i=1}^{N-1} c_i. \tag{6.26}$$

Thus

$$\begin{aligned} \boldsymbol{k}_F &= \sum_{i=1}^{N-1} c_i \boldsymbol{v}_i + c_N \boldsymbol{v}_N \\ &= \sum_{i=1}^{N-1} c_i \boldsymbol{v}_i - \sum_{i=1}^{N-1} c_i \boldsymbol{v}_N \\ &= \sum_{i=1}^{N-1} c_i (\boldsymbol{v}_i - \boldsymbol{v}_N). \end{aligned} \tag{6.27}$$

Since $\boldsymbol{k}$ is arbitrary, then the vectors $\{\boldsymbol{w}_i\} = \{\begin{pmatrix} \boldsymbol{v}_i - \boldsymbol{v}_N \\ \boldsymbol{0} \end{pmatrix}\}$ span $\ker \Delta\mathcal{L}(q_{\frac{1}{N}})$. By Lemma 84, $\{\boldsymbol{w}_i\}$ are linearly independent and so they are a basis for $\ker \Delta\mathcal{L}(q_{\frac{1}{N}})$. □

Remark 86. *Corollary 35 shows that*

$$\Delta F(q_{\frac{1}{N}}) \text{ is singular } \Leftrightarrow \Delta\mathcal{L}(q_{\frac{1}{N}}) \text{ is singular.}$$

*Theorem 85 gives a stronger result for $N > 2$. It shows that every $\boldsymbol{k} \in \ker \Delta\mathcal{L}(q_{\frac{1}{N}})$ can be written as $\boldsymbol{k} = \begin{pmatrix} \boldsymbol{k}_F \\ \boldsymbol{0} \end{pmatrix}$ where $\boldsymbol{k}_F \in \ker \Delta F(q_{\frac{1}{N}}) \cap \ker J$ so that $\boldsymbol{k} = \sum_{i=1}^{N-1} c_i \begin{pmatrix} \boldsymbol{v}_i - \boldsymbol{v}_N \\ \boldsymbol{0} \end{pmatrix}$.*

*Conversely, if* $\{\boldsymbol{v}_i\}_{i=1}^N$ *is the basis for* $\ker \Delta F(q_{\frac{1}{N}})$ *from (6.22), then a vector in* $\ker \Delta\mathcal{L}(q_{\frac{1}{N}})$ *is a linear combination of the vectors* $\{\begin{pmatrix} \boldsymbol{v}_i - \boldsymbol{v}_j \\ \mathbf{0} \end{pmatrix}\}$ *for any* $i \neq j$.

THEOREM 87. *Given that Assumption 81 holds, then* $\{\boldsymbol{w}_i\}_{i=1}^{M-1}$ *from (6.23) are a basis for* $\ker \Delta\mathcal{L}(q^*)$.

*Proof.* By Lemma 84, $\ker \Delta\mathcal{L}(q^*) \supseteq \operatorname{span}\{\boldsymbol{w}_i\}$. To get the other containment, let $\boldsymbol{k} \in \ker \Delta\mathcal{L}(q^*)$ and decompose it as in (4.1) and (4.6). Then by (4.7) we have

$$\begin{pmatrix} B_1\boldsymbol{x}_1 \\ B_2\boldsymbol{x}_2 \\ \vdots \\ B_N\boldsymbol{x}_N \end{pmatrix} = - \begin{pmatrix} \boldsymbol{k}_J \\ \boldsymbol{k}_J \\ \vdots \\ \boldsymbol{k}_J \end{pmatrix}. \tag{6.28}$$

Using the notation from (6.3) and (6.4), (6.28) implies

$$\begin{aligned} B\boldsymbol{x}_\eta &= -\boldsymbol{k}_J \text{ for } \eta \in \mathcal{U} \\ R_\nu \boldsymbol{x}_\nu &= -\boldsymbol{k}_J \text{ for } \nu \in \mathcal{R} \end{aligned} \tag{6.29}$$

from which it follows that

$$\boldsymbol{x}_\nu = R_\nu^{-1} B\boldsymbol{x}_\eta$$

for any $\eta \in \mathcal{U}$. By (4.4), $J\boldsymbol{k}_F = \mathbf{0}$ which implies $\sum_{i=1}^N \boldsymbol{x}_i = \mathbf{0}$ and so

$$\begin{aligned} & \textstyle\sum_{\nu\in\mathcal{R}} \boldsymbol{x}_\nu + \sum_{\eta\in\mathcal{U}} \boldsymbol{x}_\eta = \mathbf{0} \\ \implies & \textstyle\sum_{\nu\in\mathcal{R}} R_\nu^{-1} B\boldsymbol{x}_{\hat{\eta}} + \sum_{\eta\in\mathcal{U}} \boldsymbol{x}_\eta = \mathbf{0} \end{aligned} \tag{6.30}$$

where $\hat{\eta}$ is some fixed class in $\mathcal{U}$. By (6.29), for every $\eta \in \mathcal{U}$, $\boldsymbol{x}_\eta$ can be written as

$$\boldsymbol{x}_\eta = \boldsymbol{x}_p + d_\eta \boldsymbol{v} \tag{6.31}$$

where either $\boldsymbol{x}_p = \mathbf{0}$ or $\boldsymbol{x}_p \in \Re^K \setminus \ker B$, $d_\eta \in \Re$ and $\boldsymbol{v}$ is the basis vector of $\ker B$ from (6.19). From (6.31) it follows that

$$\begin{aligned} B\sum_{\nu\in\mathcal{R}} R_\nu^{-1} \quad & B(\boldsymbol{x}_p + d_{\hat{\eta}}\boldsymbol{v}) + B\sum_{\eta\in\mathcal{U}} (\boldsymbol{x}_p + d_\eta \boldsymbol{v}) = \mathbf{0} \\ \Leftrightarrow \quad & B\sum_{\nu\in\mathcal{R}} R_\nu^{-1} B\boldsymbol{x}_p + \sum_{\eta\in\mathcal{U}} B\boldsymbol{x}_p = \mathbf{0} \\ \Leftrightarrow \quad & (B\sum_{\nu\in\mathcal{R}} R_\nu^{-1} + MI_K) B\boldsymbol{x}_p = \mathbf{0} \\ \Leftrightarrow \quad & B\boldsymbol{x}_p = \mathbf{0} \end{aligned}$$

since we are assuming that $B\sum_{\nu\in\mathcal{R}} R_\nu^{-1} + MI_K$ is nonsingular (Assumption 81.4). In fact, $\boldsymbol{x}_p = \mathbf{0}$ since a nontrivial $\boldsymbol{x}_p \notin \ker B$. Therefore, $\boldsymbol{x}_\eta = d_\eta \boldsymbol{v}$ for every $\eta \in \mathcal{U}$. Now (6.29) shows that $\boldsymbol{k}_J = \mathbf{0}$ and so

$$\boldsymbol{x}_\nu = \mathbf{0} \text{ for } \nu \in \mathcal{R}. \tag{6.32}$$

Hence $\boldsymbol{k} = \begin{pmatrix} \boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix}$ where $[\boldsymbol{k}_F]_\nu = \begin{cases} d_\nu \boldsymbol{v} \text{ if } \nu \in \mathcal{U} \\ \mathbf{0} \text{ if } \nu \in \mathcal{R} \end{cases}$ from which it follows that $\boldsymbol{k}_F \in \ker \Delta F(q^*)$. Therefore, (4.4) assures that $\boldsymbol{k}_F \in (\ker \Delta F(q_{\frac{1}{N}})) \cap (\ker J)$ and so Lemma 83 gives

$$\boldsymbol{k}_F = \sum_{i=1}^{M} c_i \boldsymbol{v}_i \text{ and } J\boldsymbol{k}_F = \mathbf{0}$$

and now (6.26) implies

$$c_M = -\sum_{i=1}^{M-1} c_i.$$

Thus

$$\boldsymbol{k}_F = \sum_{i=1}^{M-1} c_i(\boldsymbol{v}_i - \boldsymbol{v}_M)$$

as in (6.27). Therefore, the vectors $\{\boldsymbol{w}_i\} = \{\begin{pmatrix} \boldsymbol{v}_i - \boldsymbol{v}_M \\ \mathbf{0} \end{pmatrix}\}$ span $\ker \Delta\mathcal{L}(q^*)$. By Lemma 84, $\{\boldsymbol{w}_i\}$ are linearly independent and so they are a basis for $\ker \Delta\mathcal{L}(q^*)$. □

Remark 88. *Theorem 36 shows that if the unresolved blocks of $\Delta F(q^*)$ are singular, then $\Delta\mathcal{L}(q^*)$ is singular. In particular, Theorem 87 shows that if Assumption 81 holds (so that $B$ is singular and $B\sum_\nu R_\nu^{-1} + MI_K$ is nonsingular for $M > 1$), then every $\boldsymbol{k} \in \ker \Delta\mathcal{L}(q^*)$ can be written as $\boldsymbol{k} = \begin{pmatrix} \boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix}$ where $\boldsymbol{k}_F \in \ker \Delta F(q^*) \cap \ker J$ so that*

$$\boldsymbol{k} = \sum_{i=1}^{M-1} c_i \begin{pmatrix} \boldsymbol{v}_i - \boldsymbol{v}_M \\ \mathbf{0} \end{pmatrix}.$$

*Conversely, if $\Delta\mathcal{L}(q^*)$ is singular, $R_\nu$ is nonsingular, and if $B\sum_\nu R_\nu^{-1} + MI_K$ is nonsingular, then Theorem 87 shows that $\ker \Delta F(q^*) \cap \ker J \neq \emptyset$, so then $\Delta F(q^*)$ (and $B$) must be singular. We examine the case when $B\sum_\nu R_\nu^{-1} + MI_K$ is singular (which does not necessarily cause a singularity in $\Delta F(q^*)$) in chapter 8.*

In light of the previous Remark, we have the following Lemma, which will prove useful in chapter 8.

COROLLARY 89. *Suppose that $q^*$ is $M$-uniform. If the unresolved blocks of $\Delta F(q^*)$ are singular and if the resolved blocks are nonsingular, then $\Delta\mathcal{L}(q^*)$ is singular. Conversely, if $\Delta\mathcal{L}(q^*)$ is singular, and if Assumptions 81.3 and 81.4 hold for $M > 1$, then $\Delta F(q^*)$ is singular. In both cases,* $\dim\ker\Delta F = M$ *and* $\dim\ker\Delta_{q,\lambda}\mathcal{L} = M - 1$.

We have produced a basis of the kernel of $\Delta\mathcal{L}(q^*)$ for arbitrary optimization problems of the form (1.9)

$$\max_{q\in\Delta}(G(q) + \beta D(q))$$

as long as Assumption 81) is met. For the Information Distortion problem (2.34), we have that $G = H(Y_N|Y)$, the conditional entropy (2.17), which is a strictly concave function (see (2.20)). For any problem (1.9) where $G$ is a strictly concave function, we have the following Lemma.

LEMMA 90. *Let $(q^*, \beta^*)$ be some singular point of $\Delta F(q^*)$ such that $G$ is strictly concave (and no further assumptions on $D$). Let $\boldsymbol{u}$ be any nontrivial vector in* $\ker\Delta F(q^*)$. *Then $\boldsymbol{u}^T\Delta D(q^*)\boldsymbol{u} > 0$.*

*Proof.* $\Delta F(q^*)\boldsymbol{u} = \mathbf{0}$ implies $\boldsymbol{u}^T\Delta F(q^*)\boldsymbol{u} = 0$ which in turn gives

$$\boldsymbol{u}^T\Delta G(q^*)\boldsymbol{u} + \beta^*\boldsymbol{u}^T\Delta D(q^*)\boldsymbol{u} = 0. \tag{6.33}$$

Since $G$ is strictly concave, $\Delta G(q)$ is negative definite for any $q$ which implies that $\boldsymbol{u}^T\Delta G(q^*)\boldsymbol{u} < 0$. For (6.33) to hold, we must have $\boldsymbol{u}^T\Delta D(q^*)\boldsymbol{u} > 0$. □

## Liapunov-Schmidt Reduction

In order to apply the theory of chapter 5 to (3.18) at a given bifurcation point $(q^*, \lambda^*, \beta^*)$, we must translate the bifurcation to $(\mathbf{0}, \mathbf{0}, 0)$ and require that the Jacobian vanishes at bifurcation. To accomplish the former, consider the system

$$\mathcal{F}(q, \lambda, \beta) = \nabla_{q,\lambda}\mathcal{L}(q + q^*, \lambda + \lambda^*, \beta + \beta^*), \tag{6.34}$$

so that

$$\partial_{q,\lambda}\mathcal{F}(\mathbf{0}, \mathbf{0}, 0) = \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*).$$

To assure that the Jacobian vanishes as required, we consider the Liapunov-Schmidt reduction of $\mathcal{F}$ at bifurcation $(\mathbf{0}, \mathbf{0}, 0)$. That is, we restrict $\nabla_{q,\lambda}\mathcal{L}$ to $\ker\Delta\mathcal{L}(q^*)$ about $(q^*, \lambda^*, \beta^*)$. Since we will be using the explicit basis $\{\boldsymbol{w}_i\}_{i=1}^{M-1}$ from (6.23), we require that at the point $(q^*, \lambda^*, \beta^*)$, Assumption 81 holds. First, we determine the relevant spaces in the reduction. The Jacobian of the right hand side of (3.18), $\Delta\mathcal{L}(q^*)$, is symmetric. Furthermore, the spaces $B_2$ and $B_0$ defined in (5.28) are each the finite dimensional Euclidean space $\Re^{NK+K}$. Hence, we can take the vector space complements $\mathcal{M}$ and $\mathcal{N}$ from (5.28) as

$$\mathcal{M} = (\ker\Delta\mathcal{L}(q^*))^\perp = \text{range}\Delta\mathcal{L}(q^*)^T = \text{range}\Delta\mathcal{L}(q^*)$$

and

$$\mathcal{N} = (\text{range}\Delta\mathcal{L}(q^*))^\perp = \ker \Delta\mathcal{L}(q^*)^T = \ker \Delta\mathcal{L}(q^*).$$

Therefore, the Liapunov-Schmidt reduced equation of $\mathcal{F}$ is

$$\begin{aligned} \phi &: \ker \Delta\mathcal{L}(q^*) \times \Re \to \ker \Delta\mathcal{L}(q^*) \\ \phi(\boldsymbol{w}, \beta) &= (I - E)\mathcal{F}(\boldsymbol{w} + U(\boldsymbol{w}, \beta), \beta). \end{aligned} \tag{6.35}$$

As outlined in (5.32), $I - E$ is the projection onto $\ker \Delta\mathcal{L}(q^*)$ with $\ker(I - E) = \text{range}\Delta\mathcal{L}(q^*)$, $\boldsymbol{w} \in \ker \Delta\mathcal{L}(q^*)$ and $U(\boldsymbol{w}, \beta) \in \text{range}\Delta\mathcal{L}(q^*)$. In particular, we define the orthogonal projection onto $\text{range}\Delta\mathcal{L}(q^*)$ as

$$\begin{aligned} E &= \Delta\mathcal{L}(q^*)(\Delta\mathcal{L}(q^*)^T\Delta\mathcal{L}(q^*))^-\Delta\mathcal{L}(q^*)^T \\ &= \Delta\mathcal{L}(q^*)(\Delta\mathcal{L}(q^*)^2)^-\Delta\mathcal{L}(q^*). \end{aligned}$$

We now investigate an equivalent representation of the Liapunov-Schmidt reduction (6.35) on $\Re^{M-1}$, a representation of $\ker \Delta\mathcal{L}(q^*)$, as in (5.36). Let $W = (\boldsymbol{w}_1\ \boldsymbol{w}_2 ... \boldsymbol{w}_{N-1})$, the $(NK + K) \times (M - 1)$ matrix whose column space is $\ker \Delta\mathcal{L}(q^*)$, where $\boldsymbol{w}_i$ are defined in (6.23) (Theorem 85). Thus, for every $\boldsymbol{w} \in \ker \Delta\mathcal{L}(q^*)$, there exists $\boldsymbol{x} \in \Re^{M-1}$ such that $W\boldsymbol{x} = \boldsymbol{w}$. Now define

$$\begin{aligned} r &: \Re^{M-1} \times \Re \to \Re^{M-1} \\ r(\boldsymbol{x}, \beta) &= W^T\phi(W\boldsymbol{x}, \beta) \qquad (6.36) \\ &= W^T(I - E)\mathcal{F} \text{ (by (6.35))} \\ &= W^T\mathcal{F} - W^TE\mathcal{F} \\ &= W^T\mathcal{F} \qquad (6.37) \end{aligned}$$

where the last equality is justified by the fact that $E\mathcal{F} \in \text{range}\Delta\mathcal{L}(q^*)$ and that the column space of $W$ is $\ker \Delta\mathcal{L}(q^*)$, which are orthogonal. The function $r$ is equivalent to $\phi$ in the sense that $r = \mathbf{0}$ if and only if $\phi = \mathbf{0}$.

From (6.36) we see that the $(M - 1) \times (M - 1)$ Jacobian of $r$ is

$$\partial_{\boldsymbol{x}} r(\boldsymbol{x}, \beta) = W^T\partial_{\boldsymbol{w}}\phi(\boldsymbol{w}, \beta)W \tag{6.38}$$

where $\boldsymbol{w} = W\boldsymbol{x}$. Using (6.37), we see that $\partial_{\boldsymbol{x}} r(\boldsymbol{x}, \beta)$ as in (5.37) can be written as

$$\partial_{\boldsymbol{x}} r(\boldsymbol{x}, \beta) = W^T\Delta_{q,\lambda}\mathcal{L}(q + q^*, \lambda + \lambda^*, \beta + \beta^*)(W + \partial_{\boldsymbol{w}} U(W\boldsymbol{x}, \beta)W) \tag{6.39}$$

where $\begin{pmatrix} q \\ \lambda \end{pmatrix} = W\boldsymbol{x} + U(W\boldsymbol{x}, \beta)$.

The three dimensional array of second derivatives of $r$ from (5.40) becomes

$$\frac{\partial^2 r_i}{\partial x_j \partial x_k} = < \boldsymbol{w}_i, (I - E)\left(\Delta_{q,\lambda}\mathcal{L}\frac{\partial^2 U}{\partial x_j \partial x_k} + \partial_Q^3\mathcal{L}[\boldsymbol{w}_j + \frac{\partial U}{\partial x_j}, \boldsymbol{w}_k + \frac{\partial U}{\partial x_k}]\right) >, \tag{6.40}$$

where $Q = \begin{pmatrix} q \\ \lambda \end{pmatrix}$ and $\Delta_{q,\lambda}\mathcal{L}$ and $\partial_Q^3\mathcal{L}$ are both evaluated at $(q+q^*, \lambda+\lambda^*, \beta+\beta^*)$. From (5.42), we see that at bifurcation when $(q, \lambda, \beta) = (\mathbf{0}, \mathbf{0}, 0)$, that

$$\begin{aligned} \frac{\partial^2 r_i}{\partial x_j \partial x_k}(\mathbf{0}, 0) &= < \boldsymbol{w}_i, (I-E)\partial_Q^3\mathcal{L}(q^*, \lambda^*, \beta^*)[\boldsymbol{w}_j, \boldsymbol{w}_k] > \quad (6.41) \\ &= < \boldsymbol{w}_i, \partial_Q^3\mathcal{L}(q^*, \lambda^*, \beta^*)[\boldsymbol{w}_j, \boldsymbol{w}_k] > \quad (6.42) \end{aligned}$$

where the last equality follows from the fact that $< \boldsymbol{w}_i, (I-E)V >=< \boldsymbol{w}_i, V >$ for any vector $V$ since $\boldsymbol{w}_i \perp EV$. Now let $\hat{\boldsymbol{w}}_i = \boldsymbol{v}_i - \boldsymbol{v}_M$ for each $i$ between 1 and $M-1$. Then $\boldsymbol{w}_i = \begin{pmatrix} \hat{\boldsymbol{w}}_i \\ \mathbf{0} \end{pmatrix}$. Thus, (6.42) simplifies to show that

$$\begin{aligned} \frac{\partial^2 r_i}{\partial x_j \partial x_k}(\mathbf{0}, 0) &= < \boldsymbol{w}_i, \partial_Q^3\mathcal{L}(q^*, \lambda^*, \beta^*)[\boldsymbol{w}_j, \boldsymbol{w}_k] > \\ &= < \hat{\boldsymbol{w}}_i, \partial_q^3 F(q^*, \beta^*)[\hat{\boldsymbol{w}}_j, \hat{\boldsymbol{w}}_k] > \\ &= \sum_{\nu,\delta,\eta \in Y_N} \sum_{l,m,n \in Y} \frac{\partial^3 F(q^*, \beta^*)}{\partial q_{\nu l} \partial q_{\delta m} \partial q_{\eta n}} [\hat{\boldsymbol{w}}_i]_{\nu l} [\hat{\boldsymbol{w}}_j]_{\delta m} [\hat{\boldsymbol{w}}_k]_{\eta n}. \end{aligned}$$

Recall that $\frac{\partial^2 F}{\partial q_{\nu k} \partial q_{\delta m}} = 0$ if $\nu \neq \delta$ (see (3.9)), and so $\frac{\partial^3 F}{\partial q_{\nu k} \partial q_{\delta m} \partial q_{\eta l}} = 0$ unless $\nu = \delta = \eta$. Thus, the last equation can be further simplified as

$$\frac{\partial^2 r_i}{\partial x_j \partial x_k}(\mathbf{0}, 0) = \sum_{\nu \in Y_N} \sum_{l,m,n \in Y} \frac{\partial^3 F(q^*, \beta^*)}{\partial q_{\nu l} \partial q_{\nu m} \partial q_{\nu n}} [\hat{\boldsymbol{w}}_i]_{\nu l} [\hat{\boldsymbol{w}}_j]_{\nu m} [\hat{\boldsymbol{w}}_k]_{\nu n}.$$

Now, substituting $\hat{\boldsymbol{w}}_i = \boldsymbol{v}_i - \boldsymbol{v}_M$ and using the definition of $\boldsymbol{v}_i$ from (6.21) we get that

$$\frac{\partial^2 r_i}{\partial x_j \partial x_k}(\mathbf{0}, 0) = \sum_{\nu \in \mathcal{U}} \sum_{l,m,n \in Y} \frac{\partial^3 F(q^*, \beta^*)}{\partial q_{\nu l} \partial q_{\nu m} \partial q_{\nu n}} \left( \delta_{ijk\nu} [\boldsymbol{v}]_l [\boldsymbol{v}]_m [\boldsymbol{v}]_n - \delta_{\nu M} [\boldsymbol{v}]_l [\boldsymbol{v}]_m [\boldsymbol{v}]_n \right). \quad (6.43)$$

Finally, we use the fact that for any $\nu, \eta \in \mathcal{U}$, $\frac{\partial^2 F}{\partial q_{\nu m} \partial q_{\nu n}} = \frac{\partial^2 F}{\partial q_{\eta m} \partial q_{\eta n}}$ which implies that $\frac{\partial^3 F}{\partial q_{\nu l} \partial q_{\nu m} \partial q_{\nu n}} = \frac{\partial^3 F}{\partial q_{\eta l} \partial q_{\eta m} \partial q_{\eta n}}$. Thus

$$\frac{\partial^2 r_i}{\partial x_j \partial x_k}(\mathbf{0}, 0) = \sum_{l,m,n \in Y} \frac{\partial^3 F(q^*, \beta^*)}{\partial q_{\nu l} \partial q_{\nu m} \partial q_{\nu n}} \left( \delta_{ijk} [\boldsymbol{v}]_l [\boldsymbol{v}]_m [\boldsymbol{v}]_n - [\boldsymbol{v}]_l [\boldsymbol{v}]_m [\boldsymbol{v}]_n \right). \quad (6.44)$$

An immediate consequence of (6.44) is that $\frac{\partial^2 r_i}{\partial x_i \partial x_i}(\mathbf{0}, 0) = 0$ for each $i$. Furthermore, (6.44) shows that $\frac{\partial^2 r_i}{\partial x_j \partial x_k}(\mathbf{0}, 0) = \frac{\partial^2 r_{i'}}{\partial x_{j'} \partial x_{k'}}(\mathbf{0}, 0)$ for any $(i, j, k)$ and $(i', j', k')$ such that at least one of $i$, $j$ and $k$ are distinct and at least one of $i'$, $j'$ and $k'$ are distinct.

Equivariance of the Reduction

Theorem 73 answered the question

*For what group is $\nabla_{q,\lambda}\mathcal{L}$ equivariant?*

The next question

*For what group is the Liapunov-Schmidt reduced function of $\nabla_{q,\lambda}\mathcal{L}, \phi$, equivariant?*

is answered by Proposition 46.2 and Proposition 67.1: Since $\mathcal{M} = \text{range}\Delta\mathcal{L}(q^*)$ and $\mathcal{N} = \ker \Delta\mathcal{L}(q^*)$ from (5.28) are $\Gamma$-invariant, and $\nabla_{q,\lambda}\mathcal{L}$ is $\Gamma$-equivariant (Theorem 73), then $\phi$ is $\Gamma$-equivariant. Since $\Gamma_{\mathcal{U}} < \Gamma$, then $\phi$ is also $\Gamma_{\mathcal{U}}$-equivariant .

The next question that arises is:

*For what group is $r$ equivariant?*

By Lemma 67.2, the Lie group that acts equivariantly on $r$ is constructed as in (5.47) and (5.48): for each $\gamma \in \Gamma_{\mathcal{U}}$, and for $\{\boldsymbol{w}_i\}_{i=1}^{M-1}$ as in (6.23), $\gamma\boldsymbol{w}_j = \sum_i a_{ij}\boldsymbol{w}_i$ for $a_{ij} \in \Re$. Define the $(M-1)\times(M-1)$ matrix $A(\gamma)$ by setting

$$[A(\gamma)]_{ij} := a_{ij}.$$

Then

$$\mathcal{A}_M := \mathcal{A} = \{A(\gamma)|\gamma \in \Gamma_{\mathcal{U}}\}. \tag{6.45}$$

The previous discussion is summarized in the following Lemma.

Lemma 91.

1. *Let $\phi$ be defined as in (6.35) and let $\Gamma_{\mathcal{U}}$ be defined as in (6.5). Then $\phi$ is $\Gamma_{\mathcal{U}}$-equivariant.*
2. *Let $r$ be defined as in (6.36) and let $\mathcal{A}$ defined as in (6.45). Then $r$ is $\mathcal{A}$-equivariant.*

The group $\mathcal{A}$ for which $r$ is equivariant is pivotal to the development of the theory that follows. Therefore, we analyze $\mathcal{A}$ in more detail and, before giving an explicit algorithm for generating any $A(\gamma) \in \mathcal{A}$ from $\gamma \in \Gamma_{\mathcal{U}}$, we first show an example.

Example 92. *We derive the explicit groups $\mathcal{A}_M$ from (6.45) for $M = N = 2$ and 3.*

*When $N = 2$, $\Gamma \cong S_2$ is a group of 2 $(NK+K)\times(NK+K)$ matrices,*

$$\Gamma := \{I_{NK+K}, \gamma_{12}\}$$

*(see (6.5)), and* $\mathcal{A}_2$ *is the group of scalars isomorphic to* $S_2$*. To determine the two scalar elements of* $\mathcal{A}_2$*, we observe that the basis of* $\ker \Delta F(q_{\frac{1}{N}})$ *is* $\{\boldsymbol{v}_i\}_{i=1}^2$ *where* $\boldsymbol{v}_i$ *are defined in (6.21), and so the single basis vector for* $\ker \Delta \mathcal{L}(q_{\frac{1}{N}})$ *is*

$$\boldsymbol{w}_1 = \begin{pmatrix} \boldsymbol{v}_1 - \boldsymbol{v}_2 \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v} \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix}$$

*(Theorem 85). Thus*

$$\gamma_{12}\boldsymbol{w}_1 = \gamma_{12} \begin{pmatrix} \boldsymbol{v} \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{v} \\ \boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix},$$

*which shows that* $\gamma_{12}\boldsymbol{w}_1 = -\boldsymbol{w}_1$*. By definition then,* $A(\gamma_{12}) = -1$*. Together with the group identity, this shows that* $\mathcal{A}_2 = \{1, -1\}$*.*

*For* $M = N = 3$*,* $\Gamma$ *is a group of* $6$ $(NK + K) \times (NK + K)$ *matrices,*

$$\Gamma := \{I_{NK+K}, \gamma_{12}, \gamma_{13}, \gamma_{23}, \gamma_{123}, \gamma_{132}\},$$

*and* $\mathcal{A}_3$ *is the group of* $2 \times 2$ *matrices isomorphic to* $S_3$*. The basis for* $\ker \Delta F(q_{\frac{1}{N}})$ *is* $\{\boldsymbol{v}_i\}_{i=1}^3$*, which implies that (Theorem 85) the two basis vectors of* $\ker \Delta \mathcal{L}(q_{\frac{1}{N}})$ *are*

$$\boldsymbol{w}_1 = \begin{pmatrix} \boldsymbol{v}_1 - \boldsymbol{v}_3 \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{0} \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix}, \boldsymbol{w}_2 = \begin{pmatrix} \boldsymbol{v}_2 - \boldsymbol{v}_3 \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{v} \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix}.$$

*To determine* $A(\gamma_{12})$*, we compute*

$$\gamma_{12}\boldsymbol{w}_1 = \gamma_{12} \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{0} \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{v} \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{w}_2,$$

$$\gamma_{12}\boldsymbol{w}_2 = \gamma_{12} \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{v} \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{0} \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{w}_1$$

*which shows that* $A(\gamma_{12}) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. *To compute the element* $A(\gamma_{123}) \in \mathcal{A}_3$, *we compute*

$$\begin{aligned} \gamma_{123}\boldsymbol{w}_1 &= \gamma_{12}\begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{0} \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{v} \\ \boldsymbol{v} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix} = -\boldsymbol{w}_1 + \boldsymbol{w}_2, \\ \gamma_{123}\boldsymbol{w}_2 &= \gamma_{12}\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{v} \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{v} \\ \boldsymbol{0} \\ \boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix} = -\boldsymbol{w}_1. \end{aligned}$$

*Thus,* $A(\gamma_{123}) = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}$. *One can continue in this fashion to show that the elements of* $\mathcal{A}_3$,

$$A(I_{NK+K}), A(\gamma_{12}), A(\gamma_{13}), A(\gamma_{23}), A(\gamma_{123}), A(\gamma_{132}),$$

*are*

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix}, \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}$$

*respectively.*

Armed with the intuition provided by the preceding example, we now give the following algorithm for generating any $A(\gamma) \in \mathcal{A}$ from $\gamma \in \Gamma_{\mathcal{U}}$.

ALGORITHM 93. *Let* $\Gamma_{\mathcal{U}}$ *be defined as in (6.8). Let* $\{\boldsymbol{w}_i\}$ *be defined as in (6.23), the basis of* $\ker \Delta\mathcal{L}(q^*)$ *for some* $M$*-uniform solution* $q^*$ *where Assumption 81 holds. Let* $A := A(\gamma)$ *(defined in (6.45)) for some* $\gamma \in \Gamma_{\mathcal{U}}$. *Suppose that* $\gamma$ *maps class* $j$ *to class* $k$ *and class* $M$ *to class* $m$. *Then for* $1 \leq k \leq M-1$,

$$[A(\gamma)]_{k^{th}\ row} = \begin{cases} j^{th}\ row\ of\ I_{M-1}\ if\ m = M\ \ or\ if\ k \neq m \neq M \\ -1 \ldots -1\ if\ k = m \neq M \end{cases}$$

*Proof.* $\boldsymbol{a}_j$, the $j^{th}$ column of $A$, is constructed by considering

$$\gamma\boldsymbol{w}_j = \gamma\begin{pmatrix} \boldsymbol{v}_j \\ \boldsymbol{0} \end{pmatrix} - \gamma\begin{pmatrix} \boldsymbol{v}_M \\ \boldsymbol{0} \end{pmatrix}.$$

There are a few cases to consider:

1. If $m = M$ then

$$\begin{aligned} \gamma \boldsymbol{w}_j &= \begin{pmatrix} \boldsymbol{v}_k \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{v}_M \\ \mathbf{0} \end{pmatrix} \\ &= \boldsymbol{w}_k \end{aligned}$$

   Therefore, if $M$ is fixed and $j \mapsto k$ for any $k$,

$$\boldsymbol{a}_j = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ 0 \end{pmatrix} \tag{6.46}$$

   where the 1 is in the $k^{th}$ row.

2. If $m \neq M$ and $k = M$ then

$$\begin{aligned} \gamma \boldsymbol{w}_j &= \begin{pmatrix} \boldsymbol{v}_M \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{v}_m \\ \mathbf{0} \end{pmatrix} \\ &= -\boldsymbol{w}_m. \end{aligned}$$

   Therefore, if $M$ is not fixed and $j \mapsto M \mapsto m$,

$$\boldsymbol{a}_j = \begin{pmatrix} 0 \\ \vdots \\ -1 \\ 0 \\ 0 \end{pmatrix} \tag{6.47}$$

   where -1 is in the $m^{th}$ row.

3. Lastly, if $m \neq M$ and if $k \neq M$, then

$$\begin{aligned} \gamma \boldsymbol{w}_j &= \begin{pmatrix} \boldsymbol{v}_k \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{v}_m \\ \mathbf{0} \end{pmatrix} \\ &= \boldsymbol{w}_k - \boldsymbol{w}_m \end{aligned}$$

   Therefore, if $M$ is not fixed, $j \mapsto k \neq M$ and $M \mapsto m$,

$$\boldsymbol{a}_j = \begin{pmatrix} 0 \\ \vdots \\ -1 \\ \vdots \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \tag{6.48}$$

where $-1$ is in the $m^{th}$ row and 1 is in the $k^{th}$ row.

Now, $A$ is constructed by acting on $\boldsymbol{w}_j$ for every $j$, $1 \leq j \leq M-1$. Thus, if $M$ is fixed, then $A$ is a permutation matrix, where the $k^{th}$ row is the $j^{th}$ row of $I_{M-1}$ (by (6.46)). If $M$ is not fixed, then by (6.47) and (6.48), the $m^{th}$ row of $A$ is $-1$. Furthermore, by (6.48), the $k^{th}$ row (for $k \neq M$ and $k \neq m$) is the $j^{th}$ row of $I_{M-1}$. □

REMARK 94. *For any $\gamma \in \Gamma$ such that class $M$ is fixed (i.e. $m = M$), $A(\gamma)$ is a permutation matrix.*

THEOREM 95. *Let $\mathcal{A}$ be defined as in (6.45) such that Assumption 81 holds. The action of $\mathcal{A}$ is absolutely irreducible on $\Re^{M-1}$.*

*Proof.* Assumption 81 is necessary since the explicit form of $\mathcal{A}$ depends on the basis of $\ker \Delta\mathcal{L}(q^*)$ from Theorem 87. We use induction to show that if $X$ is an $(M-1) \times (M-1)$ matrix that commutes with every $A \in \mathcal{A}_M$, then $X = c(\beta) I_{M-1}$.

For $M = 2$, $\mathcal{A}_2$ is the group $\{1, -1\}$. For $M = 3$, we have

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}.$$

By algorithm 93

$$\mathcal{A}_3 \supset \left\{A(\gamma_{(12)}), A(\gamma_{(13)})\right\} = \left\{ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix} \right\}.$$

If $X$ commutes with all of the elements of $\mathcal{A}_3$, then $X \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} X$ and so $\begin{pmatrix} x_{12} & x_{11} \\ x_{22} & x_{21} \end{pmatrix} = \begin{pmatrix} x_{21} & x_{22} \\ x_{11} & x_{12} \end{pmatrix}$. Hence $x_{12} = x_{21} = b$ and $x_{11} = x_{22} = c$ for some $b, c \in \Re$. Thus

$$X = \begin{pmatrix} c & b \\ b & c \end{pmatrix}.$$

Furthermore, $X \begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix} X$, which shows that

$$\begin{pmatrix} -c & b-c \\ -b & c-b \end{pmatrix} = \begin{pmatrix} -c-b & -c-b \\ b & c \end{pmatrix}.$$

Thus $b = 0$ and so

$$X = c \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Now assume the inductive hypothesis, that $cI_{M-1}$ is the only matrix that commutes with all of the elements of $\mathcal{A}_M$ for some $c \in \Re$. Consider $\mathcal{A}_{M+1}$, the group of $M \times M$ matrices when there are $M+1$ classes. Let $X$ be an $M \times M$ matrix such that

$$XA = AX \quad \forall A \in \mathcal{A}_{M+1} \tag{6.49}$$

Write $X$ as

$$X = \begin{pmatrix} X_0 & \boldsymbol{x} \\ \boldsymbol{y}^T & d_X \end{pmatrix}$$

where $X_0$ is $(M-1) \times (M-1)$, $\boldsymbol{x}$ and $\mathbf{y}$ are $(M-1) \times 1$ and $d_X \in \Re$. Write every $A \in \mathcal{A}_{M+1}$ as

$$A = \begin{pmatrix} A_0 & \boldsymbol{a} \\ \boldsymbol{b}^T & d_A \end{pmatrix}$$

where $A_0$ is $(M-1) \times (M-1)$, $\boldsymbol{a}$ and $\mathbf{b}$ are $(M-1) \times 1$ and $d_A \in \Re$. Equation (6.49) becomes

$$X_0 A_0 + \boldsymbol{x}\boldsymbol{b}^T = A_0 X_0 + \boldsymbol{a}\boldsymbol{y}^T. \tag{6.50}$$

For the element $A := A(\gamma_{M(M+1)})$, $A_0 = I_{M-1}$, $\boldsymbol{a} = \mathbf{0}$, $\boldsymbol{b} = -\mathbf{1}$, and $d_A = -1$ (Algorithm 93). Equation (6.50) becomes

$$X_0 + \boldsymbol{x}\boldsymbol{b}^T = X_0$$

so that

$$\boldsymbol{x}\boldsymbol{b}^T = (-\boldsymbol{x} \quad -\boldsymbol{x} \quad \ldots -\boldsymbol{x}) = \mathbf{0}$$

which shows that $\boldsymbol{x} = \mathbf{0}$. To show that $\boldsymbol{y} = \mathbf{0}$, we consider the transposition $A := A(\gamma_{1M})$. By Algorithm 93, $A_0 = \begin{pmatrix} \mathbf{0} \\ I_M \end{pmatrix}$, $\boldsymbol{a} = \boldsymbol{e}_1$, $\boldsymbol{b} = \boldsymbol{e}_1$, and $d_A = 0$. Substituting these and $\boldsymbol{x} = \mathbf{0}$ into equation (6.50),

$$X_0 \begin{pmatrix} \mathbf{0} \\ I_M \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ I_M \end{pmatrix} X_0 + \begin{pmatrix} \boldsymbol{y}^T \\ \mathbf{0}^T \\ \vdots \\ \mathbf{0}^T \end{pmatrix}$$

shows that $[\boldsymbol{y}]_1 = 0$. Evaluating (6.50) for $A := A(\gamma_{iM})$ for every $1 \leq i < M$ shows that $\boldsymbol{y} = \mathbf{0}$

To complete the proof, we need to show that $d_X = c$, which is accomplished by considering

$$XA(\gamma_{(M-1)M}) = A(\gamma_{(M-1)M})X$$

which becomes

$$\begin{pmatrix} X_0 & \mathbf{0} \\ \mathbf{0}^T & d_X \end{pmatrix} \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} X_0 & \mathbf{0} \\ \mathbf{0}^T & d_X \end{pmatrix}.$$

Since $X_0 = cI_{M-1}$, then this equation can be rewritten as

$$\begin{pmatrix} X_0 & \mathbf{0} \\ \mathbf{0}^T & d_X \end{pmatrix} \begin{pmatrix} \begin{pmatrix} I_{M-2} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix} & \mathbf{0} \\ \mathbf{0}^T \quad 1 & 0 \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} I_{M-2} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{pmatrix} & \mathbf{0} \\ \mathbf{0}^T \quad 1 & 0 \end{pmatrix} \begin{pmatrix} X_0 & \mathbf{0} \\ \mathbf{0}^T & d_X \end{pmatrix}$$

where $\mathbf{0}$ is an $(M-2)\times 1$ vector of zeros. Multiplying these block matrices out shows that

$$\begin{pmatrix} cI_{M-2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & 0 & c \\ \mathbf{0} & d_X & 0 \end{pmatrix} = \begin{pmatrix} cI_{M-2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & 0 & d_X \\ \mathbf{0}^T & c & \boldsymbol{x}_{M-1} \end{pmatrix}.$$

It follows that $d_X = c$ □

LEMMA 96. *Let $\mathcal{A}_M$ be defined as in (6.45) such that Assumption 81 holds. Then $\mathcal{A}_M \cong S_M$.*

*Proof.* Consider the map

$$\begin{aligned} \varphi : \Gamma_{\mathcal{U}} &\rightarrow \mathcal{A}_M \\ \gamma &\mapsto A(\gamma) \end{aligned}$$

where the group $\Gamma_{\mathcal{U}}$, which is isomorphic to $S_M$, is defined in (6.8). The proof is complete if $\varphi$ is shown to be a group homomorphism with $\ker\varphi = \{I_{NK+K}\}$ [27]. To show the former, for $\gamma_1, \gamma_2 \in \Gamma_{\mathcal{U}}$, let $[A(\gamma_1)]_{ij} = a_{ij}, [A(\gamma_2)]_{ij} = b_{ij}$ and $[A(\gamma_1\gamma_2)]_{ij} = c_{ij}$. Observe that $A(\gamma_1\gamma_2)$ is constructed by considering

$$\begin{aligned} \gamma_1\gamma_2\boldsymbol{w}_j &= \gamma_1(\gamma_2\boldsymbol{w}_j) \\ &= \gamma_1(\sum_i b_{ij}\boldsymbol{w}_i) \\ &= \sum_i b_{ij}\gamma_1\boldsymbol{w}_i \\ &= \sum_i b_{ij}(\sum_k a_{ki}\boldsymbol{w}_k) \\ &= \sum_k(\sum_i a_{ki}b_{ij})\boldsymbol{w}_k. \end{aligned}$$

Thus, $c_{kj} = \sum_i a_{ki}b_{ij}$ which implies that $A(\gamma_1\gamma_2) = A(\gamma_1)A(\gamma_2)$ and so $\varphi(\gamma_1\gamma_2) = \varphi(\gamma_1)\varphi(\gamma_2)$. Hence, $\varphi$ is a group homomorphism.

To show that $\ker\varphi = \{I_{NK+K}\}$, suppose $\varphi(\hat{\gamma}) = I_{M-1} \in \mathcal{A}_M$ for some $\hat{\gamma} \in \Gamma_{\mathcal{U}}$. Then for every $j$,

$$\begin{aligned}\hat{\gamma}\boldsymbol{w}_j &= \sum_i a_{ij}\boldsymbol{w}_i \\ &= \sum_i \delta_{ij}\boldsymbol{w}_i \\ &= \boldsymbol{w}_j.\end{aligned}$$

By (6.5), $\hat{\gamma} = \begin{pmatrix} \hat{\rho} & \mathbf{0} \\ \mathbf{0} & I_K \end{pmatrix}$ for some $\hat{\rho} \in \mathcal{P}$. By (6.23), $\boldsymbol{w}_j = \begin{pmatrix} \boldsymbol{v}_j - \boldsymbol{v}_M \\ \mathbf{0} \end{pmatrix}$. Hence, for every $j$,

$$\begin{aligned}&\hat{\gamma}\boldsymbol{w}_j = \boldsymbol{w}_j \\ \Longrightarrow\quad &\begin{pmatrix} \hat{\rho} & \mathbf{0} \\ \mathbf{0} & I_K \end{pmatrix}\begin{pmatrix} \boldsymbol{v}_j - \boldsymbol{v}_M \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_j - \boldsymbol{v}_M \\ \mathbf{0} \end{pmatrix} \\ \Longrightarrow\quad &\hat{\rho}(\boldsymbol{v}_j - \boldsymbol{v}_M) = \boldsymbol{v}_j - \boldsymbol{v}_M. \qquad (6.51)\end{aligned}$$

$\hat{\rho}$ is a $NK \times NK$ permutation matrix in $\mathcal{P}_{\mathcal{U}}$ so that $\hat{\rho} = \begin{pmatrix} E_1^{\nu_1} \\ \vdots \\ E_N^{\nu_N} \end{pmatrix}$ where $E_i^{\nu_i}$ is a $K \times NK$ matrix of 0's with identity $I_K$ in the $K \times K$ block component corresponding to class $\nu_i$. Hence, for every $j$, (6.51) becomes

$$\begin{pmatrix} E_1^{\nu_1} \\ \vdots \\ E_N^{\nu_N} \end{pmatrix}(\boldsymbol{v}_j - \boldsymbol{v}_M) = \boldsymbol{v}_j - \boldsymbol{v}_M$$

which is true if and only if, for every $j$,

$$\begin{aligned}E_j^{\nu_j}(\boldsymbol{v}_j - \boldsymbol{v}_M) &= \boldsymbol{v} \qquad (6.52)\\ E_M^{\nu_M}(\boldsymbol{v}_j - \boldsymbol{v}_M) &= -\boldsymbol{v} \\ E_i^{\nu_i}(\boldsymbol{v}_j - \boldsymbol{v}_M) &= \mathbf{0} \ \ \forall i \notin \{j, M\}\end{aligned}$$

where $\boldsymbol{v}$ is defined in (6.19). Observe that

$$E_k^{\nu_k}(\boldsymbol{v}_l - \boldsymbol{v}_M) = \begin{cases} \boldsymbol{v} \text{ if } \nu_k = l \\ -\boldsymbol{v} \text{ if } \nu_k = M \\ \mathbf{0} \text{ otherwise} \end{cases}. \qquad (6.53)$$

By (6.52), $k = l$ for every $k$ and $l$. By (6.52) and (6.53), $\nu_k = l$ for every $k$ and $l$. Thus, $\nu_k = k$ for every $k$ so that

$$\begin{aligned}E_1^{\nu_1} &= (I_K \ \ \mathbf{0}...\mathbf{0}) \\ E_2^{\nu_2} &= (\mathbf{0} \ \ I_K...\mathbf{0}) \\ &\vdots \\ E_M^{\nu_M} &= (\mathbf{0} \ \ \mathbf{0}...I_K).\end{aligned}$$

Hence, $\hat{\rho} = I_{NK}$ which implies $\hat{\gamma} = I_{NK+K}$ from which it follows that $\ker\varphi = \{I_{NK+K}\}$. By the First Isomorphism Theorem ([27] p.97), $\varphi$ is a group isomorphism and so we have that $\mathcal{A}_M \cong \Gamma_{\mathcal{U}} \cong S_M$. □

Isotropy Subgroups

To show the existence of bifurcating branches from bifurcation of equilibria of (3.18),

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda}\mathcal{L}(q,\lambda,\beta),$$

the Equivariant Branching Lemma and the Smoller-Wasserman Theorem require two things. First, we must we work with the Liapunov Schmidt reduction $r(\boldsymbol{x},\beta)$ (6.36) of $\nabla_{q,\lambda}\mathcal{L}$,

$$r : \Re^{M-1} \times \Re \to \Re^{M-1}.$$

Secondly, we must determine the maximal isotropy subgroups of $\mathcal{A}_M$, (6.45), the group for which the reduction $r(\boldsymbol{x},\beta)$ is equivariant (Lemma 91.2), as well as the elements contained in the fixed point spaces for each of the maximal isotropy subgroups. For arbitrary $M$, the lattice of maximal subgroups of $S_M$, let alone the full lattice of subgroups, is unknown [16, 46]. This section ascertains some of the maximal isotropy subgroups of $\mathcal{A}_M$, in particular the subgroups which are isomorphic to $S_{M-1}$ (Lemma 100), which enables us to show the existence of bifurcating solutions from an $M$-uniform solution $q^*$ of (3.18) for any $M > 1$ when Assumption 81 holds.

First, we show a class of subgroups of $\mathcal{A}_M$ that do not fix any vector in $\Re^{M-1}$, motivated by the following example.

Example 97. *Recall the explicit construction of the group $\mathcal{A}_3$ in Example 92. Observe that an element $A(\gamma)$ of $\mathcal{A}_M$ fixes a vector in $\Re^{M-1}$ if and only if $A(\gamma)$ has the eigenvalue* 1. *This shows for the elements of $\mathcal{A}_3$, $A(\gamma_{123}) = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}$ and $A(\gamma_{132}) = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}$, $\mathrm{Fix}\langle A(\gamma_{123})\rangle$ and $\mathrm{Fix}\langle A(\gamma_{132})\rangle$ are empty.*

The observation made in the previous example is true for the general case, which we prove next.

Lemma 98. *Let $\Gamma_{\mathcal{U}}$ be defined as in (6.8). If $\gamma$ is an element of order $M$ in $\Gamma_{\mathcal{U}}$, then $\dim(\mathrm{Fix}\langle\gamma\rangle \cap \ker\Delta\mathcal{L}(q^*)) = 0$. Equivalently, if $\mathcal{A}_M$ is defined as in (6.45) and if $A$ is an element of order $M$ in $\mathcal{A}_M$, then $\dim(\mathrm{Fix}\langle A\rangle) = 0$.*

*Proof.* Let $\gamma$ be some element of order $M$ in $\Gamma_{\mathcal{U}}$. Then $A := A(\gamma)$ is an element of order $M$ in $\mathcal{A}_M$ (Lemma 96). First, note that $|A| = M \Leftrightarrow A$ is an $M$-cycle

(Proposition 76.3), by which it is meant that $A$ is isomorphic to an $M$-cycle in $\mathcal{S}$ (see (6.12)). Suppose there exists $\boldsymbol{x} \in \Re^{M-1}$ such that $A\boldsymbol{x} = \boldsymbol{x}$. Next, let

$$C = BAB^{-1} \text{ for some } B \in \mathcal{A}_M, \tag{6.54}$$

which is possible if and only if $C$ is (isomorphic to an element in $\mathcal{S}$) of the same cycle type as $A$ (Proposition 76.6). Hence $C$ is an $M$-cycle. Furthermore, all $M$-cycles can be generated as in (6.54)(Proposition 76.6). Lastly, note that $CB\boldsymbol{x} = BAB^{-1}(B\boldsymbol{x}) = BA\boldsymbol{x} = B\boldsymbol{x}$ if and only if $C$ fixes $B\boldsymbol{x}$. Thus,

$$\begin{gathered} \text{there is an } M\text{-cycle in } \mathcal{A}_M \text{ which fixes some nontrivial } \boldsymbol{x} \in \Re^{M-1} \\ \text{if and only if} \\ \text{every } M\text{-cycle in } \mathcal{A}_M \text{ fixes some nontrivial vector in } \Re^{M-1}. \end{gathered} \tag{6.55}$$

The proof is completed by showing that there is an $M$-cycle in $\mathcal{A}_M$ which does not fix any nontrivial vector in $\Re^{M-1}$.

Consider the $M$-cycle $\gamma \cong (123...(M-1)M) \in S_M$. By Algorithm 93,

$$A(\gamma) = \begin{pmatrix} -1 & -1 & ... & -1 & -1 \\ 1 & 0 & ... & 0 & 0 \\ 0 & 1 & ... & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & ... & 1 & 0 \end{pmatrix}$$

is the $(M-1) \times (M-1)$ isomorphic matrix representation in $\mathcal{A}_M$ of $\gamma$. Observe that the characteristic polynomial for $A(\gamma)$ is $\lambda^{M-1} + \lambda^{M-2} + ... + \lambda + 1$ which does not have 1 as a root. This implies that there does not exist a nontrivial $\boldsymbol{x} \in \Re^{M-1}$ such that $A(\gamma)\boldsymbol{x} = \boldsymbol{x}$. By (6.55), no $M$-cycle of $\mathcal{A}_M$ fixes a nontrivial element of $\Re^{M-1}$. By Proposition 68, no $M$-cycle of $\Gamma_{\mathcal{U}}$ fixes a nontrivial element of $\ker \Delta\mathcal{L}(q^*)$. □

THEOREM 99. *For the alternating group $A_M$ (see Definition 74),* $\dim \mathrm{Fix}(A_M) = 0$.

*Proof.* Suppose that $M$ is odd. Then $A_M$ contains elements of order $M$ (by Definition 74 and Proposition 76.4) which implies $\dim \mathrm{Fix}(A_M) = 0$ by Lemma 98. Now suppose that $M$ is even. Then $A_M$ contains elements of cycle length $M-1$ (Proposition 76.4). Consider the $(M-1)$-cycles $\gamma_{(1...(M-1))}$ and $\gamma_{(2...M)} \in A_M$. By Algorithm 93, $A(\gamma_{(1...(M-1))}) \in \mathcal{A}$ is a permutation matrix from which it follows that $\langle A(\gamma_{(1...(M-1))})\rangle$ fixes the $(M-1) \times 1$ vector $\mathbf{1}$. By Proposition 46.3, $A(\gamma_{(1M)})\mathbf{1}$ is fixed by the group $A(\gamma_{(1M)})\langle A(\gamma_{(1...(M-1))})\rangle A(\gamma_{(1M)})^{-1}$, which is equal to the group $\langle A(\gamma_{(2...M)})\rangle$ by Proposition 76.5. To compute $A(\gamma_{(1M)})\mathbf{1}$, we use the explicit form of

$A(\gamma_{(1M)})$ given by Algorithm 93,

$$A(\gamma_{(1M)})\mathbf{1} = \begin{pmatrix} -1 & -1 & -1 & ... & -1 \\ 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & 0 & ... & 1 \end{pmatrix} \mathbf{1} = \begin{pmatrix} 1-M \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

The Trace Formula (Proposition 46.4) shows that

$$\dim(\mathrm{Fix}\langle A(\gamma_{(1\ldots(M-1))})\rangle) = 1,$$

since $A(\gamma_{(1\ldots(M-1))})$ is a permutation matrix, and the only element of $\langle A(\gamma_{(1\ldots(M-1))})\rangle$ which contributes to $\sum_{A\in\langle A(\gamma_{(1\ldots(M-1))})\rangle}\mathrm{trace}(A)$ is $A(\gamma_{(1\ldots(M-1))})^{M-1} = I_{M-1}$. Thus,

$$\dim(\mathrm{Fix}\langle A(\gamma_{(2\ldots M)})\rangle) = 1,$$

from which it follows that $\dim(\mathrm{Fix}(A_M)) \leq 1$. Hence, any vector $\boldsymbol{u}$ fixed by $A_M$ must be in $\big(\mathrm{Fix}\langle A(\gamma_{(1\ldots(M-1))})\rangle\big) \cap \big(\mathrm{Fix}\langle A(\gamma_{(2\ldots M)})\rangle\big)$. Thus, $\boldsymbol{u} = a\mathbf{1} = b\begin{pmatrix} 1-M \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ for some $a, b \in \Re$ which implies that $a = b = 0$. □

$M$-uniform solutions are in the fixed point space of $\Gamma_{\mathcal{U}}$ (Theorem 71), which is isomorphic to $S_M$. To apply the theory of chapter 5 to $M$-uniform solutions of the gradient flow (3.18) at a bifurcation point $(q^*, \lambda^*, \beta^*)$, one must ascertain the maximal isotropy subgroups of $\Gamma_{\mathcal{U}}$. We now find some of these subgroups. In particular, we show that the $M$ subgroups of $\Gamma_{\mathcal{U}}$, that are isomorphic to $S_{M-1} < S_M$, are maximal isotropy subgroups of $\Gamma_{\mathcal{U}}$. The representation of these subgroups in $\Gamma_{\mathcal{U}}$ is $\langle T_k\rangle$ (see (6.13)). In fact, these maximal isotropy subgroups of $\Gamma_{\mathcal{U}}$ have fixed point spaces of dimension 1. We also obtain an explicit basis of the fixed point space for each subgroup. This derivation is done in two parts, Lemma 100 and Lemma 103.

LEMMA 100. *Let $\Gamma_{\mathcal{U}}$ be defined as in (6.8). Let $T_k$ be the set of transpositions in $\Gamma_{\mathcal{U}}$ such that the $k^{th}$ unresolved class in $\mathcal{U}$ is fixed. (as in (6.13)). Let $\hat{\boldsymbol{u}}_k$ be a $NK \times 1$ vector such that*

$$[\hat{\boldsymbol{u}}_k]_\nu = \begin{cases} (M-1)\boldsymbol{v} & \text{if } \nu \text{ is the } k^{th} \text{ unresolved class of } \mathcal{U} \\ -\boldsymbol{v} & \text{if } \nu \neq k \text{ is any other unresolved class of } \mathcal{U} \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{6.56}$$

*and let*

$$\boldsymbol{u}_k = \begin{pmatrix} \hat{\boldsymbol{u}}_k \\ \mathbf{0} \end{pmatrix} \tag{6.57}$$

*where $\mathbf{0}$ is $K \times 1$. Then $\langle T_k\rangle$ is the isotropy subgroup of $\boldsymbol{u}_k$. Equivalently, if $T_k$ is the set of transpositions in $\mathcal{A}_M$, then $\langle T_k\rangle$ is the isotropy subgroup of $A(\gamma_{kM})\mathbf{1}$.*

*Proof.* First, we show that $\langle T_M \rangle$ fixes $\boldsymbol{u}_M$. By Algorithm 93 and Lemma 96, the matrices in $\langle T_M \rangle < \Gamma_{\mathcal{U}}$ map to each and every one of the $(M-1)\times(M-1)$ permutation matrices, $\{A_i\}_{i=1}^{(M-1)!}$, in $\mathcal{A}_M$ (see (6.45)). It is clear that $\{A_i\}$, being permutation matrices, fix $\mathbf{1}$, an $(M-1)\times 1$ vector of ones. By Proposition 68, the corresponding vector which is fixed in $\ker \Delta\mathcal{L}(q^*)$ by $\langle T_M \rangle < \Gamma_{\mathcal{U}}$ is $W\mathbf{1} = \sum_{i=1}^{M-1} \boldsymbol{w}_i = -\boldsymbol{u}_M$. Here, $W$ is the $NK \times (M-1)$ matrix

$$W = \begin{pmatrix} | & | & | & | \\ \boldsymbol{w}_1 & \boldsymbol{w}_2 & \dots & \boldsymbol{w}_{M-1} \\ | & | & | & | \end{pmatrix},$$

and $\{\boldsymbol{w}_i\}$ are defined in (6.23). The group $\Gamma_{\mathcal{U}}$ does not fix $\boldsymbol{u}_M$ since $\gamma \boldsymbol{u}_M \neq \boldsymbol{u}_M$ for any $\gamma \in \Gamma_{\mathcal{U}}$ which does not fix class $M$. Therefore, since there does not exist a proper subgroup of $\Gamma_{\mathcal{U}}$ that is strictly larger than $\langle T_M \rangle$ (Theorem 77), then $\langle T_M \rangle$ must be the isotropy subgroup for $\boldsymbol{u}_M$.

Let $\gamma_{kM}$ be the transposition in $\Gamma_{\mathcal{U}}$ that permutes class $k$ with class $M$. Now apply Proposition 46.3 which assures that $\gamma_{kM}\boldsymbol{u}_M = \boldsymbol{u}_k$ has isotropy subgroup

$$\gamma_{kM}\langle T_M \rangle \gamma_{kM}^{-1}.$$

By Proposition 76.5, the conjugation $\gamma_{kM}\langle T_M \rangle \gamma_{kM}^{-1}$ simply replaces each permutation to and from the $k^{th}$ class in each element of $\langle T_M \rangle$ with permutations to and from the $M^{th}$ class. That is

$$\gamma_{kM}\langle T_M \rangle \gamma_{kM}^{-1} = \langle T_k \rangle$$

□

Remark 101. *When $M = N$, $\boldsymbol{u}_k$ as defined in (6.57) is*

$$\boldsymbol{u}_k = \begin{pmatrix} -\boldsymbol{v} \\ \vdots \\ -\boldsymbol{v} \\ (N-1)\boldsymbol{v} \\ -\boldsymbol{v} \\ \vdots \\ -\boldsymbol{v} \\ \mathbf{0} \end{pmatrix} \tag{6.58}$$

*where $(N-1)\boldsymbol{v}$ is in the $k^{th}$ row.*

Example 102. *Recall the explicit form of the group $\mathcal{A}_3$ in Example 92. The elements $A(\gamma_{12}), A(\gamma_{13}), A(\gamma_{23})$, given by the $2\times 2$ matrices*

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix} \tag{6.59}$$

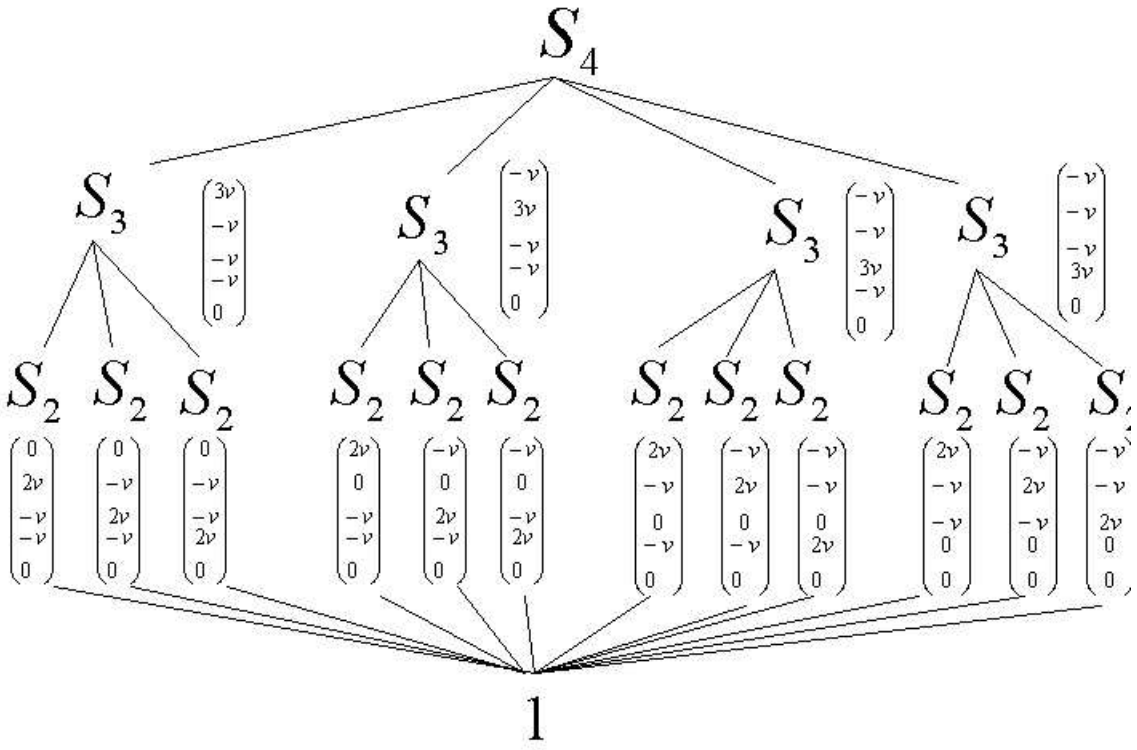



Figure 13. The lattice of the maximal isotropy subgroups $S_M < S_N$ for $N = 4$ from Lemma 100 and the corresponding basis vectors of the fixed point spaces of the corresponding groups from Lemma 100.

*in* $\mathcal{A}_3$ *respectively, are the sole generators of the subgroups* $\langle T_k \rangle = \mathcal{A}_2 < \mathcal{A}_3$ *which are isomorphic to* $S_2$. *That is,* $\langle T_3 \rangle = \langle A(\gamma_{12}) \rangle$, $\langle T_2 \rangle = \langle A(\gamma_{13}) \rangle$, *and* $\langle T_1 \rangle = \langle A(\gamma_{23}) \rangle$, *each group of which has order* 2. *The eigenvectors of each of the matrices of (6.59) with the eigenvalue* 1 *are*

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ 1 \end{pmatrix} \tag{6.60}$$

*respectively, which shows that*

$$\dim \mathrm{Fix}\langle A(\gamma_{12}) \rangle = \dim \mathrm{Fix}\langle A(\gamma_{13}) \rangle = \dim \mathrm{Fix}\langle A(\gamma_{23}) \rangle = 1.$$

*When* $M = N = 3$, *the vectors that correspond to (6.60) which are fixed in* $\ker \Delta\mathcal{L}(q_{\frac{1}{N}}) = \mathrm{span}(\{\boldsymbol{w}_1, \boldsymbol{w}_2\})$ *by the corresponding subgroups* $\langle \gamma_{12} \rangle$, $\langle \gamma_{13} \rangle$, *and* $\langle \gamma_{23} \rangle$ *of* $\Gamma$ *are*

$$\boldsymbol{w}_1 + \boldsymbol{w}_2 = \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{v} \\ -2\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix}, \boldsymbol{w}_1 - 2\boldsymbol{w}_2 = \begin{pmatrix} \boldsymbol{v} \\ -2\boldsymbol{v} \\ \boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix}, -2\boldsymbol{w}_1 + \boldsymbol{w}_2 = \begin{pmatrix} -2\boldsymbol{v} \\ \boldsymbol{v} \\ \boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix}$$

*respectively (see Figure 14).*

LEMMA 103. *Let $\Gamma_{\mathcal{U}}$ be defined as in (6.8). Let $T_k$ be the set of transpositions in $\Gamma_{\mathcal{U}}$ such that the $k^{th}$ unresolved class in $\mathcal{U}$ is fixed (as in (6.13)). Then*

$$\dim \mathrm{Fix}(\langle T_k \rangle \cap \ker \Delta\mathcal{L}(q^*)) = 1.$$

*Equivalently, if $T_k$ is the set of transpositions in $\mathcal{A}_M$, then* $\dim\langle T_k \rangle = 1$.

*Proof.* Consider $\langle T_M \rangle < \mathcal{A}_M$. By Algorithm 93, $M$ is fixed, and so $\langle T_M \rangle$ is a Lie group of $(M-1)\times(M-1)$ permutation matrices. By the Trace Formula (Proposition 46.4), $\dim \mathrm{Fix}(\langle T_M \rangle) = \frac{1}{|\langle T_M \rangle|}\sum_{A\in\langle T_M\rangle} \mathrm{trace}(A)$. Note that the $i^{th}$ row of an element of $\langle T_M \rangle$ contributes to $\sum_{A\in\langle T_M\rangle} \mathrm{trace}(A)$ only when there is a 1 in the $i^{th}$ component of that row. When there is a 1 in the $i^{th}$ component of the $i^{th}$ row, there are $(M-2)!$ possible combinations of the other $(M-2)$ rows. Thus, the first row of the elements of $\langle T_M \rangle$ is counted $(M-2)!$ times in $\sum_{A\in\langle T_M\rangle} \mathrm{trace}(A)$, the second row of the elements of $\langle T_M \rangle$ is counted $(M-2)!$ times, ... , and the $(M-1)^{st}$ row of the elements of $\langle T_M \rangle$ is counted $(M-2)!$ times. It follows that $\sum_{A\in\langle T_M\rangle} \mathrm{trace}(A) = (M-1)(M-2)!$ and so $\dim \mathrm{Fix}\langle T_M \rangle = \frac{(M-1)!}{|\langle T_M \rangle|} = 1$. Thus $\mathrm{Fix}\langle T_M \rangle$ has basis $\{\boldsymbol{x}\}$ for some $\boldsymbol{x} \in \Re^{M-1}$. Now suppose that there exists $k$ such that $\langle T_k \rangle$ is an isotropy subgroup of $\mathcal{A}_M$ for two vectors $x_1$ and $x_2$ in $\Re^{M-1}$. By Proposition 76.6, there is a $C \in \mathcal{A}_M$ such that $\langle T_M \rangle = C\langle T_k \rangle C^{-1}$. By Proposition 46.3, $\langle T_M \rangle$ is the isotropy subgroup of $Cx_1$ and $Cx_2$ which implies that $Cx_1 = a\boldsymbol{x}$ and $Cx_2 = b\boldsymbol{x}$ for some nonzero $a, b \in \Re$. Thus, $x_1 = \frac{b}{a}x_2$ from which it follows that $\dim(\mathrm{Fix}\langle T_k \rangle) = 1$ for every $k$. The Lemma now follows from Proposition 68. □

Two lines of reasoning have been developed to show that $\langle T_k \rangle$ is a maximal isotropy subgroup of $\Gamma_{\mathcal{U}}$ (or of $\mathcal{A}_M$). The first uses Theorem 77 and Lemma 100. The second relies on the two previous Lemmas, Lemma 100 and Lemma 103, since an isotropy group with a fixed point space of dimension 1 is necessarily maximal.

Theorem 71 shows that $\mathrm{Fix}(\Gamma_{\mathcal{U}})$ is the vector space of points in $\Re^{NK+K}$ generated by the vectors $(q, \lambda)$ where $q$ is $M$-uniform. The final ingredient that is required to apply the theory of chapter 5 to a bifurcation point $(q^*, \lambda^*, \beta^*)$ when $q^*$ is an $M$-uniform solution is to show that

$$\mathrm{Fix}(\Gamma_{\mathcal{U}}) \cap \ker \Delta\mathcal{L}(q^*) = \{\mathbf{0}\},$$

which is equivalent to showing that

$$\mathrm{Fix}(\mathcal{A}_M) = \{\mathbf{0}\}.$$

This section is finished with two proofs which show this. When Assumption 81 is satisfied, this result already follows from the fact that $\mathcal{A}_M$ acts absolutely irreducibly on $\Re^{M-1}$ (Theorem 95) and Propositions 46.5 and 46.6. The next theorem deals with the solution $q_{\frac{1}{N}}$. It is presented separately because Assumptions 81.3 and 81.4 are not required.

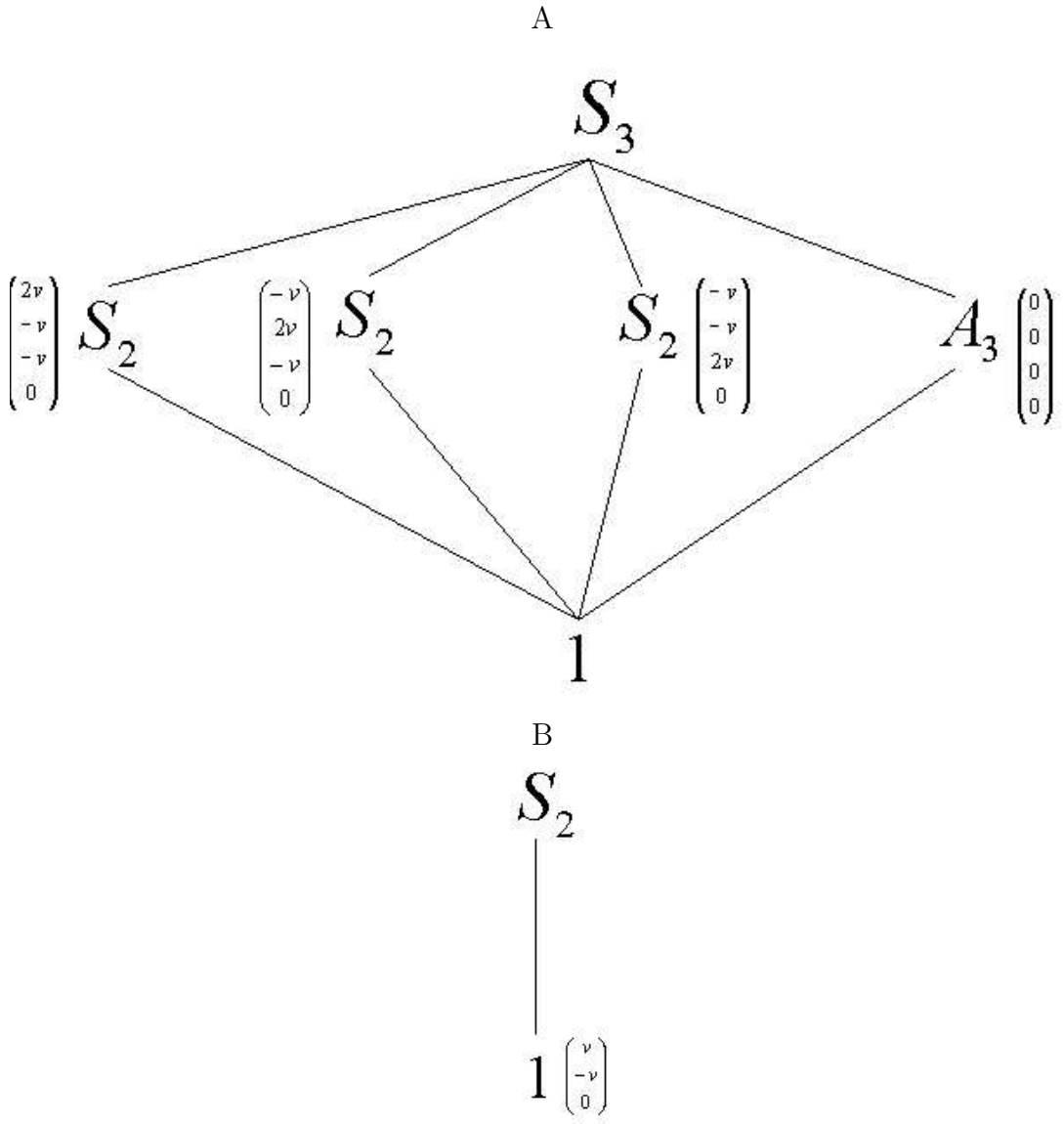



Figure 14. Panel (A) shows the full lattice of subgroups $S_2 < S_3$ for $N = 4$ and the corresponding basis vectors, from Theorem 99 and Lemma 100, of the fixed point spaces of the corresponding groups. Panel (B) shows the full lattice of subgroups of $S_2$, and the corresponding basis vectors, from Lemma 100, of the fixed point spaces of the corresponding groups.

PROPOSITION 104. *Let $(q_{\frac{1}{N}}, \lambda^*, \beta^*)$ be some bifurcation point of (3.18) such that Assumptions 81.1 and 81.2 hold, and let $\Gamma$ be defined as in (6.5). Then*

$$\mathrm{Fix}(\Gamma) \cap \ker \Delta\mathcal{L}(q_{\frac{1}{N}}) = \{\mathbf{0}\}.$$

*Proof.* Let $\boldsymbol{k}$ be a $(NK+K)\times 1$ vector in $\mathrm{Fix}(\Gamma) \cap \ker \Delta\mathcal{L}(q_{\frac{1}{N}})$. Decompose $\boldsymbol{k}$ as in (4.1) and (4.6). Then $\gamma\boldsymbol{k} = \boldsymbol{k}$ for every $\gamma \in \Gamma$. By Remark 86, $\gamma\boldsymbol{k}$ becomes

$$\gamma\boldsymbol{k} = \begin{pmatrix} \rho & \mathbf{0} \\ \mathbf{0} & I_K \end{pmatrix} \begin{pmatrix} \boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \rho\boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix} \quad \forall \rho \in \mathcal{P} \text{ (see (6.5))}.$$

Hence $\rho\boldsymbol{k}_F = \boldsymbol{k}_F$ and so $\boldsymbol{x} = \boldsymbol{x}_i = \boldsymbol{x}_j$ for every $i$ and $j$, $1 \le i, j \le N$. From (4.8), $\sum_\nu \boldsymbol{x}_\nu = \mathbf{0}$ from which it follows that $\sum_\nu \boldsymbol{x} = \mathbf{0}$ which shows that $\boldsymbol{x} = \mathbf{0}$ and so $\boldsymbol{k} = \mathbf{0}$. □

The result for an arbitrary $M$-uniform solution when Assumption 81 is satisfied is next.

PROPOSITION 105. *Let $\Gamma_{\mathcal{U}}$ be defined as in (6.8). Let $(q^*, \lambda^*, \beta^*)$ be some bifurcation point of (3.18) where $q^*$ is $M$-uniform such that Assumption 81 holds. Then* $\mathrm{Fix}(\Gamma_{\mathcal{U}}) \cap \ker \Delta\mathcal{L}(q^*) = \{\mathbf{0}\}$.

*Proof.* Let $\boldsymbol{k} \in \mathrm{Fix}(\Gamma_{\mathcal{U}}) \cap \ker \Delta\mathcal{L}(q^*)$. Decompose $\boldsymbol{k}$ as in (4.1) and (4.6). Since $\boldsymbol{k} \in \mathrm{Fix}(\Gamma_{\mathcal{U}})$, then $\gamma\boldsymbol{k} = \boldsymbol{k}$ for every $\gamma \in \Gamma_{\mathcal{U}}$. This and Remark 88 (which we can apply since Assumption 81.4 holds) imply that

$$\begin{pmatrix} \rho & \mathbf{0} \\ \mathbf{0} & I_K \end{pmatrix} \begin{pmatrix} \boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix}$$
$$\implies \rho\boldsymbol{k}_F = \boldsymbol{k}_F \quad \forall \rho \in \mathcal{P}_{\mathcal{U}}.$$

Hence, $\boldsymbol{x}_\nu = \boldsymbol{u}$ for some $K \times 1$ vector $\boldsymbol{u}$ for every $\nu \in \mathcal{U}$. Thus (4.8) becomes

$$J\boldsymbol{k}_F = \sum_{\nu\in\mathcal{R}} \boldsymbol{x}_\nu + M\boldsymbol{u} = \mathbf{0}. \tag{6.61}$$

Since Assumption 81.4 holds, then Remark 88 also shows that $\Delta F(q^*)\boldsymbol{k}_F = \mathbf{0}$, which gives

$$\begin{aligned} B\boldsymbol{u} &= \mathbf{0} \\ R_\nu \boldsymbol{x}_\nu &= \mathbf{0} \quad \forall \nu \in \mathcal{R} \end{aligned} \tag{6.62}$$

and so

$$\boldsymbol{x}_\nu = \mathbf{0} \quad \forall \nu \in \mathcal{R}$$

since $R_\nu$ are nonsingular. □

## Bifurcating Branches from $M$-uniform Solutions

We have laid the groundwork so that in this section, we finally may present the main result of this chapter, which is the existence of explicit bifurcating branches from an $M$-uniform $q^*$ at some $\beta^*$ and vector of Lagrange multipliers $\lambda^*$ (110). To accomplish this, the Equivariant Branching Lemma or the Smoller-Wasserman Theorem is applied to the Liapunov Schmidt reduction $r(\boldsymbol{x}, \beta)$ (6.36) of $\nabla_{q,\lambda}\mathcal{L}$ at a bifurcation point $(q,^*, \lambda^*, \beta^*)$, where $\nabla_{q,\lambda}\mathcal{L}$ defines the dynamical system (3.18)

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda}\mathcal{L}(q, \lambda, \beta).$$

To satisfy the requirements of these theorems, in the last section, we found some maximal isotropy subgroups of $\Gamma_{\mathcal{U}}$, and the corresponding elements of $\ker \Delta\mathcal{L}(q^*)$ in the fixed point spaces of these subgroups. Equivalently, we have found some maximal isotropy subgroups of $\mathcal{A}_M$, and the corresponding elements of $\Re^{M-1}$ in the fixed point spaces of these subgroups.

Before getting to the main result, we first prove that degenerate singularities of $\Delta\mathcal{L}(q^*)$ (see Definition 25) do not occur on any branch of equilibria $(q^*, \lambda^*, \beta^*)$ to (3.18) (i.e. not necessarily $M$-uniform) when $D$ is convex on $\ker \Delta F(q^*)$. In particular, this condition holds when $G$ from (3.2) is strictly concave (Corollary 108). For the Information Distortion problem (2.34), $G = H(Y_N|Y)$ is strictly concave. For the Information Bottleneck problem (2.35), $G = I(Y, Y_N)$ is not *strictly* concave, and in chapter 4, we showed that $F$ is highly degenerate. Thus, this theorem does not apply to this case.

THEOREM 106. *Let $q^*$ be any stationary point to (3.1) where $\Delta D(q^*)$ (defined in (3.2)) is positive definite on $\ker \Delta F(q^*)$ and Assumptions 81.2–81.4 hold. Then $(q^*, \lambda^*, \beta^*)$ is a singularity of $\ker \Delta\mathcal{L}(q^*)$ if and only if $(q^*, \lambda^*, \beta^*)$ is a bifurcation point.*

*Proof.* Necessity follows from Theorem 24. To get sufficiency, let $r(\boldsymbol{x}, \beta)$ be the Liapunov Schmidt reduction from (6.37). By Proposition 46.1 and Theorem 95, we have that $\partial_{\boldsymbol{x}} r(\mathbf{0}, \beta) = c(\beta) I_{M-1}$. The theorem is proved by showing that $c'(0) \neq 0$. In fact, we will show that $c'(0) > 0$.

Let $\dim \ker \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*) = d > 0$. By (6.39)

$$\partial_{\boldsymbol{x}} r(\mathbf{0}, \beta) = W^T \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta + \beta^*)(W + \partial_{\boldsymbol{w}} U(\mathbf{0}, \beta) W) = c(\beta) I_{M-1}.$$

Choose some arbitrary $\boldsymbol{z} \in \Re^d$ and let $\boldsymbol{k} = W\boldsymbol{z}$ so that $\boldsymbol{k} \in \ker \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*)$. Multiplying on the left by $\boldsymbol{z}^T$ and on the right by $\boldsymbol{z}$ gives

$$\boldsymbol{k}^T \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta + \beta^*)(I_{NK+K} + \partial_{\boldsymbol{w}} U(\mathbf{0}, \beta))\boldsymbol{k} = c(\beta)\boldsymbol{z}^T\boldsymbol{z}$$

By Remark 88, $\boldsymbol{k} = \begin{pmatrix} \boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix}$ where $\boldsymbol{k}_F \in \ker \Delta F(q^*, \beta^*) \cap \ker J$. Thus

$$\boldsymbol{k}^T \Delta_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta + \beta^*) = (\Delta_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta + \beta^*)\boldsymbol{k})^T = \begin{pmatrix} \Delta F(q^*, \beta + \beta^*)\boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix}^T$$

where $\mathbf{0}$ is $K \times 1$. It follows that

$$c(\beta) = \frac{\begin{pmatrix} \boldsymbol{k}_F^T \Delta F(q^*, \beta + \beta^*) & \mathbf{0}^T \end{pmatrix} (I_{NK+K} + \partial_{\boldsymbol{w}} U(\mathbf{0}, \beta)) \begin{pmatrix} \boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix}}{||\boldsymbol{z}||^2}. \tag{6.63}$$

From (3.2), we rewrite $\Delta F(q^*, \beta + \beta^*) = \Delta G(q^*) + (\beta + \beta^*)\Delta D(q^*) = \Delta G(q^*) + \beta^* \Delta D(q^*) + \beta \Delta D(q^*)$. Then

$$\boldsymbol{k}_F^T \Delta F(q^*, \beta + \beta^*) = (\Delta F(q^*, \beta + \beta^*)\boldsymbol{k}_F)^T \quad = \quad \beta (\Delta D(q^*)\boldsymbol{k}_F)^T.$$

Furthermore,

$$\boldsymbol{z}^T \boldsymbol{z} = \boldsymbol{z}^T W^T W \boldsymbol{z} = \boldsymbol{k}^T \boldsymbol{k} = \boldsymbol{k}_F^T \boldsymbol{k}_F.$$

So (6.63) becomes

$$c(\beta) = \beta \frac{\begin{pmatrix} \boldsymbol{k}_F^T \Delta D(q^*) & \mathbf{0}^T \end{pmatrix} (I_{NK+K} + \partial_{\boldsymbol{w}} U(\mathbf{0}, \beta)) \begin{pmatrix} \boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix}}{||\boldsymbol{k}_F||^2}.$$

Finally, we have that

$$c'(\beta) = \frac{\begin{pmatrix} \boldsymbol{k}_F^T \Delta D(q^*) & \mathbf{0}^T \end{pmatrix} \left( I_{NK+K} + \partial_{\boldsymbol{w}} U(\mathbf{0}, \beta) + \beta \frac{\partial(\partial_{\boldsymbol{w}} U(\mathbf{0}, \beta))}{\partial \beta} \frac{\partial \boldsymbol{w}}{\partial \beta} \right) \begin{pmatrix} \boldsymbol{k}_F \\ \mathbf{0} \end{pmatrix}}{||\boldsymbol{k}_F||^2}$$

and now (5.41) shows that

$$c'(0) = \frac{\boldsymbol{k}_F^T \Delta D(q^*) \boldsymbol{k}_F}{||\boldsymbol{k}_F||^2}. \tag{6.64}$$

Since we are assuming that $\Delta D(q^*)$ is positive definite on $\ker \Delta F(q^*, \beta^*)$, then

$$c'(0) = \frac{\boldsymbol{k}_F^T \Delta D(q^*) \boldsymbol{k}_F}{||\boldsymbol{k}_F||^2} > 0$$

for all $\boldsymbol{k}_F \in \ker \Delta F(q^*, \beta^*)$. □

Remark 107.

1. *Theorem 106 holds if $\Delta D(q^*)$ is negative definite on* $\ker J$.

*2. Equation (6.64) can be written as*

$$c'(0) = \frac{\boldsymbol{v}^T B_D^\nu(q^*)\boldsymbol{v}}{\|\boldsymbol{v}\|^2},$$

*where $\boldsymbol{v}$ is defined in (6.19), $\nu \in \mathcal{U}$, and $B_D^\nu(q^*)$ is the $\nu^{th}$ block of $\Delta D(q^*)$. This shows that $c'(0)$ is well defined.*

COROLLARY 108. *Let $q^*$ be any stationary point to (3.1) when $G$ (defined in (3.2)) is strictly concave and Assumptions 81.2–81.4 hold. Then $(q^*, \lambda^*, \beta^*)$ is a singularity of $\ker \Delta\mathcal{L}(q^*)$ if and only if $(q^*, \lambda^*, \beta^*)$ is a bifurcation point.*

*Proof.* Applying Lemma 90 to (6.64), we see that

$$c'(0) = \frac{\boldsymbol{k}_F^T \Delta D(q^*)\boldsymbol{k}_F}{||\boldsymbol{k}_F||^2} > 0$$

for all $\boldsymbol{k}_F \in \ker \Delta F(q^*, \beta^*)$. □

REMARK 109. *If $\Delta D(q^*)$, where $D$ is defined in (1.9), is positive definite on $\ker \Delta F(q^*)$, as is the case with the Information Distortion problem (2.34), then we only need assume that $(q^*, \lambda^*, \beta^*)$ is a singularity point since Theorem 106 assures that a bifurcation occurs at the singularity.*

We have developed enough theory to produce our main result of this chapter, which is the existence of explicit bifurcating solutions from $q^*$ at some $\beta^*$ and vector of Lagrange multipliers $\lambda^*$.

THEOREM 110. *Let $(q^*, \lambda^*, \beta^*)$ be a bifurcation point of (3.18) such that Assumption 81 holds. Then there exists $M$ bifurcating solutions, $\begin{pmatrix} q^* \\ \lambda^* \\ \beta^* \end{pmatrix} + \begin{pmatrix} t\boldsymbol{u}_k \\ \beta(t) \end{pmatrix}$, where $\boldsymbol{u}_k$ is defined in (6.57) for $1 \le k \le M$, each with isotropy group isomorphic to $S_{M-1}$.*

*Proof.* Let $r(\boldsymbol{x}, \beta)$ be the Liapunov-Schmidt reduction as defined in (6.36) which is $\mathcal{A}_M$-equivariant by Lemma 91.2. By Theorem 95, $\mathcal{A}_M$ acts absolutely irreducibly on $\ker \partial_{\boldsymbol{x}} r(\boldsymbol{x}, \beta)$ so that $\partial_{\boldsymbol{x}} r(\mathbf{0}, \beta) = c(\beta) I_{M-1}$ for some scalar function $c(\beta)$. The derivative $c'(0) \neq 0$ by the assumption that $(q^*, \lambda^*, \beta^*)$ is a bifurcation point. By (5.34) and (6.38), $\partial_{\boldsymbol{x}} r(\mathbf{0}, 0) = \mathbf{0}$. Lemma 100 shows that $\langle T_k \rangle$ is an isotropy subgroup in $\mathcal{A}_M$ and Lemma 103 shows that $\dim \mathrm{Fix}\langle T_k \rangle = 1$. Therefore the hypotheses of the Equivariant Branching Lemma (Theorem 47) are satisfied, whose application, along with (5.38), proves the theorem. □

When $q^* = q_{\frac{1}{N}}$, we can drop Assumptions 81.3 and 81.4, which state that $B\sum_\nu R_\nu^{-1} + MI_K$ is nonsingular.

COROLLARY 111. *From a bifurcation at* $(q_{\frac{1}{N}}, \lambda^*, \beta^*)$ *of (3.18) such that Assumptions 81.1 and 81.2 hold, there exists* $N$ *bifurcating solutions,* $\begin{pmatrix} q_{\frac{1}{N}} \\ \lambda^* \\ \beta^* \end{pmatrix} + \begin{pmatrix} t\boldsymbol{u}_k \\ \beta(t) \end{pmatrix}$, *where* $\boldsymbol{u}_k$ *is defined in (6.58) for* $1 \leq k \leq N$, *each with isotropy group isomorphic to* $S_{N-1}$.

*Proof.* In the proof for Theorem 110, use Theorem 85 instead of Theorem 87 to ascertain the basis of the kernel of $\Delta\mathcal{L}(q_{\frac{1}{N}})$. The corollary then follows without the hypothesis that $B\sum_\nu R_\nu^{-1} + MI_K$ is nonsingular. □

Using the relationship offered by Theorem 71, then we see that Theorem 110 and Corollary 111 show that there exists $M$ bifurcating $(M-1)$-uniform solutions from an $M$-uniform solution branch. This is the following corollary.

COROLLARY 112. *Let* $(q^*, \lambda^*, \beta^*)$ *be a bifurcation of (3.18) such that Assumption 81 holds. Then there exists* $M$ *bifurcating* $(M-1)$ *uniform solutions.*

REMARK 113.

1. *If* $\Delta D(q^*)$, *where* $D$ *is defined in (1.9), is positive definite on* $\ker \Delta F(q^*)$, *as is the case with the Information Distortion problem (2.34), then we only need assume that* $(q^*, \lambda^*, \beta^*)$ *is a singularity point in Theorem 110, and Corollaries 111, since Theorem 106 assures that a bifurcation occurs at the singularity.*

2. *An alternate proof of Theorem 110 using the Smoller-Wasserman Theorem proceeds thusly. We can use the same line of reasoning presented in the proof to Theorem 110, with the exception that we appeal to Theorem 77 and Lemma 100 to show that* $\langle T_k \rangle < \mathcal{A}_M$ *is a maximal isotropy subgroup of* $\boldsymbol{u}_k$.

   *The advantage of using the Smoller-Wasserman Theorem for the proof is that we get the existence of bifurcating branches for each and every maximal isotropy subgroup, not merely the ones where the dimension of the fixed point space of the isotropy group is 1.*

By Corollary 112, when assuming Assumption 81, then bifurcation on an $M$-uniform solution branch guarantees the existence of $M$ bifurcating $(M-1)$-uniform solutions. When $M = 3$, Theorem 110 assures that three 2-uniform solutions bifurcate from each of the 3-uniform solution branches. From bifurcation of these 2-uniform solutions, then Theorem 110 assures that two 1-uniform solutions bifurcate from each of the 2-uniform solution branches. A 1-uniform solution is one that is not fixed by the action of the full group $\Gamma$. In other words, by Theorem 71, for every $\gamma \in \Gamma$,

$$\gamma q \neq q \text{ if and only if } q \text{ is } 1-\text{uniform.}$$

Thus far we have excluded consideration of the possibility of bifurcation from 1-uniform solution branches (Assumption 81.1). We now address this scenario. The

next theorem shows that, under generic assumptions, that 1-uniform solutions do not bifurcate.

THEOREM 114. *Let $(q^*, \lambda, \beta)$ be an equilibria of (3.18) such that $q^*$ is 1-uniform and that Assumptions 81.2 and 81.3 hold. If $B\sum_\nu R_\nu^{-1} + I_K$ is nonsingular, then $\Delta\mathcal{L}(q^*)$ is nonsingular and there are no bifurcating solutions at $(q^*, \lambda, \beta)$.*

*Proof.* The proof to Theorem 87 begins with considering an arbitrary

$$\boldsymbol{k} \in \ker \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda, \beta),$$

and then decomposing $\boldsymbol{k}$ as in (4.1) and (4.6). For a 1-uniform solution, $\dim\ker\Delta\mathcal{L}(q^*) = 1$ by Assumption 81.2. The proof holds for the case where $q^*$ is 1-uniform, up until we get (6.32),

$$\boldsymbol{x}_\nu = \mathbf{0} \text{ for } \nu \in \mathcal{R} \tag{6.65}$$

which holds since we assume that $B\sum_\nu R_\nu^{-1}+I_K$ is nonsingular. Furthermore, $|\mathcal{U}| = 1$ since $q^*$ is 1-uniform and so the equation

$$\sum_{\nu\in\mathcal{R}} \boldsymbol{x}_\nu + \sum_{\eta\in\mathcal{U}} \boldsymbol{x}_\eta = \mathbf{0}$$

from (6.30) becomes

$$\sum_{\nu\in\mathcal{R}} \boldsymbol{x}_\nu + \boldsymbol{x}_\eta = \mathbf{0}. \tag{6.66}$$

By (6.65) and (6.66), $\boldsymbol{x}_\eta = \mathbf{0}$, which implies that $\boldsymbol{k} = \mathbf{0}$. Since $\boldsymbol{k}$ is an arbitrary element of $\ker\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda, \beta)$, then $\ker\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda, \beta) = \{\mathbf{0}\}$ and so $\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda, \beta)$ is nonsingular. Therefore, it follows from Theorem 24 that no bifurcation can occur at $(q^*, \lambda, \beta)$. □

## Bifurcating Branches when $M \leq 4$

In this section, we explain the types of bifurcation that the theory predicts when the number of classes is $N = 2, 3$ and 4.

For the case when $N = 2$, symmetry breaking bifurcation is possible only along the 2-uniform solution branch $(q_{\frac{1}{2}}, \lambda, \beta)$. Thus, symmetry breaking bifurcations will be classical pitchforks [34]: there will be 2 1-uniform bifurcating branches, each with isotropy group $S_1$ (Corollary 111 and Figure 14(B)). In other words, these 1-uniform solutions have no symmetry. It follows that further symmetry breaking bifurcations are not possible on either of these 2 1-uniform branches. Furthermore, Theorem 114 shows that, generically, no other type of bifurcation is possible either.

When $N = 3$, symmetry breaking bifurcation can occur either on the branch $(q_{\frac{1}{3}}, \lambda, \beta)$, or on some 2-uniform branch. Thus, from each symmetry breaking bifurcation which occurs on the branch $(q_{\frac{1}{3}}, \lambda, \beta)$, the only bifurcating branches with symmetry are the 3 2-uniform branches (Corollary 111) as depicted in Figure 14(A). From symmetry breaking bifurcation on each of the 2-uniform branches, 2 1-uniform solutions will bifurcate, each with isotropy group $S_1$ (Theorem 110). Now, further symmetry breaking bifurcations are impossible on any of the 1-uniform branches. Furthermore, Theorem 114 shows that, generically, no other type of bifurcation is possible either.

At symmetry breaking bifurcation when $N = 4$ along the branch $(q_{\frac{1}{4}}, \lambda, \beta)$, Corollary 111 shows that there are 4 3-uniform bifurcating solutions. See Figure 13 for the group lattice, and for a representation of the quantizers $q^*$ which have isotropy groups isomorphic to $S_3$, see panel (1) in Figure 18, and Figures 23(B) and 24. In addition to these branches, Figure 25 shows the existence of 3 other bifurcating branches which are "twice" 2-uniform.

As we have seen for $N = 3$, at a symmetry breaking bifurcation on any of the 3-uniform branches, Theorem 110 shows that there exists 3 2-uniform bifurcating solutions. See Figure 13 for the group lattice, and to see a representation of the quantizers $q^*$ which have isotropy group isomorphic to $S_2$, see panels (2)–(3) of Figure 18 and panels (2)–(5) of Figure 19. At symmetry breaking bifurcation on each of the 2-uniform branches, Theorem 110 shows that there exists 2 1-uniform bifurcating solutions. See panel (5) of Figure 18.

## Bifurcation Structure of $M$-uniform Solutions

This section examines the structure of bifurcating branches from $M$-uniform solutions

$$\left( \begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}_k, \beta^* + \beta(t) \right), \tag{6.67}$$

whose existence is guaranteed by Theorem 110, where $\boldsymbol{u}_k$ is defined in (6.57). We show that bifurcation from an $M$-uniform solution is always pitchfork-like (Theorem 120). We provide a condition, called the *bifurcation discriminator*, which ascertains whether the bifurcating branches are subcritical or supercritical (Theorems 127 and 128). All subcritical bifurcations are unstable (Proposition 55). We also provide a condition which determines whether supercritical branches are stable or unstable (Theorem 128). We conclude by determining when unstable bifurcating branches contain no solutions to (1.9) (Theorem 129).

To apply the tools that we developed earlier in chapter 6, one needs to check that Assumption 50 holds for (3.18)

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda} \mathcal{L}(q, \lambda, \beta).$$

We provide a condition in the next Lemma, which ascertains when Assumption 50 is met.

LEMMA 115. *Suppose that Assumption 81 holds. Then Assumption 50 is satisfied by the Liapunov-Schmidt reduction $r(\boldsymbol{x}, \beta)$ as defined in (6.37) if and only if $\Delta D(q^*)$ is positive definite on* $\ker \Delta F(q^*)$.

*Proof.* Lemma 91.2 shows that $r$ is $\mathcal{A}$-equivariant. Assumption 15.2 on $F$ implies that $r$ is infinitely differentiable. Theorem 95 shows that $\mathcal{A}$ acts absolutely irreducibly on $\Re^{M-1} = \ker \partial_{\boldsymbol{x}} r(\mathbf{0}, 0)$ so that $\partial_{\boldsymbol{x}} r(\mathbf{0}, \beta) = c(\beta) I_{M-1}$. Condition (6.63) shows that $c(0) = 0$ and condition (6.64),

$$c'(0) = \frac{\boldsymbol{k}_F^T \Delta D(q^*) \boldsymbol{k}_F}{||\boldsymbol{k}_F||^2},$$

shows that $c'(0) > 0$ since we are assuming that $\Delta D(q^*)$ is positive definite on $\ker \Delta F(q^*)$. Finally, Lemma 103 shows that the isotropy subgroup $\langle T_k \rangle \leq \Gamma_{\mathcal{U}}$ has a fixed point space of dimension 1. □

REMARK 116. *Condition (6.64) shows that $\Delta D(q^*)$ is negative definite on* $\ker \Delta F(q^*)$ *if and only if $c'(0) < 0$. In this case, symmetry breaking bifurcations are pitchfork-like (Theorem 120), and the bifurcation discriminator defined in (6.81) still dictates whether a bifurcating branch is subcritical or supercritical (Remark 125).*

When $G$ from (3.1) and (3.2),

$$\max_{q \in \Delta} (G(q) + \beta D(q)),$$

is strictly concave, as in the case of the Information Distortion problem (2.34), then the condition in Lemma 115 is satisfied.

COROLLARY 117. *Assumption 50 is satisfied by the Liapunov-Schmidt reduction $r(\boldsymbol{x}, \beta)$ as defined in (6.37) when $G$ is strictly concave.*

*Proof.* Lemma 90 shows that $\Delta D(q^*)$ is positive definite on $\ker \Delta \mathcal{L}(q^*)$. □

REMARK 118. *One can determine whether or not $\Delta D(q^*)$ is positive definite on* $\ker \Delta \mathcal{L}(q^*)$ *by applying the argument from Remark 21.2 to the case where $Z$ is the $NK \times M$ matrix with full column rank whose columns span* $\ker \Delta F(q^*)$. *Thus, $\Delta D(q^*)$ is positive definite on* $\ker \Delta F(q^*)$ *if and only if the matrix $Z^T \Delta D(q^*) Z$ is positive definite on $\Re^M$.*

The next theorem shows that the bifurcating solutions (6.67) are pitchfork-like. Before getting to this result, we first prove a necessary Lemma.

LEMMA 119. *Let $D_i = \partial^2_{\boldsymbol{x}} r_i(\mathbf{0}, 0)$. If $\nu \neq \eta$ are any two integers between 1 and $M-1$, and if neither $k$ nor $h$ are in $\{\eta, \nu\}$, then the following are true:*

1. $\sum_{l=1}^{M-1} [D_l]_{\nu\nu} = -[D_\nu]_{\nu\nu}$.
2. $[D_\eta]_{kk} = [D_\nu]_{kk}$

*Proof.* Since $r(\boldsymbol{x}, \beta)$ is $\mathcal{A}$-equivariant, then for every $A \in \mathcal{A}$,

$$Ar(\boldsymbol{x}, \beta) = r(A\boldsymbol{x}, \beta).$$

From the Taylor expansion in (6.75), it follows that

$$A \begin{pmatrix} r_1(\boldsymbol{x}, 0) \\ r_2(\boldsymbol{x}, 0) \\ \vdots \\ r_{M-1}(\boldsymbol{x}, 0) \end{pmatrix} = \begin{pmatrix} c(0)\boldsymbol{x}_1 + \boldsymbol{x}^T A^T D_1 A\boldsymbol{x} + \mathcal{O}((A\boldsymbol{x})^3) \\ c(0)\boldsymbol{x}_2 + \boldsymbol{x}^T A^T D_2 A\boldsymbol{x} + \mathcal{O}((A\boldsymbol{x})^3) \\ \vdots \\ c(0)\boldsymbol{x}_{M-1} + \boldsymbol{x}^T A^T D_{M-1} A\boldsymbol{x} + \mathcal{O}((A\boldsymbol{x})^3) \end{pmatrix}.$$

Since the quadratic terms on each side must be equal, we have that

$$A \begin{pmatrix} \boldsymbol{x}^T D_1 \boldsymbol{x} \\ \boldsymbol{x}^T D_2 \boldsymbol{x} \\ \vdots \\ \boldsymbol{x}^T D_{M-1} \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}^T A^T D_1 A\boldsymbol{x} \\ \boldsymbol{x}^T A^T D_2 A\boldsymbol{x} \\ \vdots \\ \boldsymbol{x}^T A^T D_{M-1} A\boldsymbol{x} \end{pmatrix}. \tag{6.68}$$

First, we prove part 2. Consider the element $A := A(\gamma_{\nu\eta}) \in \mathcal{A}$, the $(M-1) \times (M-1)$ permutation matrix that permutes class $\nu$ with class $\eta$, where both $\nu$ and $\eta$ are less than $M$, and all the other classes fixed. We equate the $\nu^{th}$ component on each side of (6.68) where $A = A(\gamma_{\nu\eta})$

$$[A \begin{pmatrix} \boldsymbol{x}^T D_1 \boldsymbol{x} \\ \boldsymbol{x}^T D_2 \boldsymbol{x} \\ \vdots \\ \boldsymbol{x}^T D_{M-1} \boldsymbol{x} \end{pmatrix}]_\nu = [\begin{pmatrix} \boldsymbol{x}^T A^T D_1 A\boldsymbol{x} \\ \boldsymbol{x}^T A^T D_2 A\boldsymbol{x} \\ \vdots \\ \boldsymbol{x}^T A^T D_{M-1} A\boldsymbol{x} \end{pmatrix}]_\nu. \tag{6.69}$$

The left hand side of equation (6.69) is

$$\sum_{i,j} x_i x_j [D_\eta]_{ij} \;\; = \;\; \sum_j x_j^2 [D_\eta]_{jj} + \sum_{i,j \neq i} x_i x_j [D_\eta]_{ij}. \tag{6.70}$$

Now we compute the right hand side of (6.69). Since

$$[A\boldsymbol{x}]_i = \begin{cases} x_\nu & \text{if } i = \eta \\ x_\eta & \text{if } i = \nu \\ x_i & \text{otherwise} \end{cases},$$

the right hand side of (6.69) is

$$\begin{aligned}(A\boldsymbol{x})^T D_\nu A\boldsymbol{x} &= \sum_{i,j}[A\boldsymbol{x}]_i[A\boldsymbol{x}]_j[D_\nu]_{ij} \\ &= [A\boldsymbol{x}]_k^2[D_\nu]_{kk} + 2\sum_{j\neq k}[A\boldsymbol{x}]_j[A\boldsymbol{x}]_k[D_\nu]_{jk} + \sum_{i\neq k, j\neq k}[A\boldsymbol{x}]_i[A\boldsymbol{x}]_j[D_\nu]_{ij} \\ &= x_k^2[D_\nu]_{kk} + 2\sum_{j\neq k}[A\boldsymbol{x}]_j x_k[D_\nu]_{jk} + \sum_{i\neq k, j\neq k}[A\boldsymbol{x}]_i[A\boldsymbol{x}]_j[D_\nu]_{ij}, \quad (6.71)\end{aligned}$$

where the last equality follows if $k \notin \{\nu, \eta\}$. Comparing the coefficients of the $x_k^2$ in (6.70) and (6.71), we get that

$$[D_\eta]_{kk} = [D_\nu]_{kk}$$

as long as $k \notin \{\nu, \eta\}$, proving part 2.

To get part 1, we now consider $A := A(\gamma_{\nu M})$, the element which permutes class $\nu \neq M$ with class $M$ and leaves all other classes fixed. By Algorithm 93

$$A = \begin{pmatrix} I_{\nu-1} & & & \mathbf{0} \\ -1 & -1 & \ldots & -1 \\ \mathbf{0}^T & & & I_{M-1-\nu} \end{pmatrix}$$

where $\mathbf{0}$ is a $(\nu-1)\times(M-1-\nu)$ matrix of zeros. Computing (6.69) for $A = A(\gamma_{\nu M})$ yields

$$[A\begin{pmatrix} \boldsymbol{x}^T D_1 \boldsymbol{x} \\ \boldsymbol{x}^T D_2 \boldsymbol{x} \\ \vdots \\ \boldsymbol{x}^T D_{M-1} \boldsymbol{x} \end{pmatrix}]_k = [\begin{pmatrix} \boldsymbol{x}^T A^T D_1 A\boldsymbol{x} \\ \boldsymbol{x}^T A^T D_2 A\boldsymbol{x} \\ \vdots \\ \boldsymbol{x}^T A^T D_{M-1} A\boldsymbol{x} \end{pmatrix}]_k. \quad (6.72)$$

The left hand side of (6.72) is

$$-\boldsymbol{x}^T \sum_l D_l \boldsymbol{x} = -\sum_{i,j} x_i x_j \sum_l [D_l]_{ij}. \quad (6.73)$$

Now we compute the right hand side of (6.72). First, observe that

$$[A\boldsymbol{x}]_i = \begin{cases} x_i & \text{if } i \neq \nu \\ -\sum_l x_l & \text{if } i = \nu \end{cases}.$$

The right hand side of (6.72) is $\sum_{i,j}[A\boldsymbol{x}]_i[A\boldsymbol{x}]_j[D_\nu]_{ij}$ which is equal to

$$\begin{aligned}&[A\boldsymbol{x}]_\nu^2[D_\nu]_{\nu\nu} + 2\sum_{j\neq\nu}[A\boldsymbol{x}]_j[A\boldsymbol{x}]_\nu[D_\nu]_{j\nu} + \sum_{i\neq\nu, j\neq\nu}[A\boldsymbol{x}]_i[A\boldsymbol{x}]_j[D_\nu]_{ij} \\ = \;& (-\sum_l x_l)^2[D_\nu]_{\nu\nu} - 2\sum_{j\neq\nu} x_j(\sum_l x_l)[D_\nu]_{j\nu} + \sum_{i\neq\nu, j\neq\nu} x_i x_j [D_\nu]_{ij}. \quad (6.74)\end{aligned}$$

Now, we equate the coefficients of the $x_\nu^2$ terms in (6.73) and (6.74), which yields

$$-\sum_l [D_l]_{\nu\nu} = [D_\nu]_{\nu\nu}.$$

□

Now, as promised, we use Lemma 119 to prove the desired result. Observe that we need not assume that $c'(0) \neq 0$, and so we do not make any assumption on $\Delta D(q^*)$.

THEOREM 120. *All of the bifurcating branches guaranteed by Theorem 110 and Corollary 111 are pitchfork-like. That is, for each branch, $\beta'(0) = 0$.*

*Proof.* By Lemma 53, we need to show that $< \boldsymbol{x}_0, \partial_{\boldsymbol{x}}^2 r(\mathbf{0}, 0)[\boldsymbol{x}_0, \boldsymbol{x}_0] >= 0$, for every $\boldsymbol{x}_0$ such that $W\boldsymbol{x}_0 = \boldsymbol{u}_k$ for some $k$. In fact, we show that $\partial_{\boldsymbol{x}}^2 r(\mathbf{0}, 0) = \mathbf{0}$. As in (5.21), for each integer $i$ between 1 and $M-1$, consider the Taylor series of the $i^{th}$ component of $r$, $r_i(\boldsymbol{x}, \beta)$, about $\boldsymbol{x} = \mathbf{0}$ for fixed $\beta$,

$$r_i(\boldsymbol{x}, \beta) = r_i(\mathbf{0}, \beta) + \partial_{\boldsymbol{x}} r_i(\mathbf{0}, \beta)^T \boldsymbol{x} + \boldsymbol{x}^T \partial_{\boldsymbol{x}}^2 r_i(0, \beta)\boldsymbol{x} + \mathcal{O}(\boldsymbol{x}^3).$$

Equation (5.3) shows that $r_i(\mathbf{0}, \beta) = 0$, and by Assumption 50.2, $\partial_{\boldsymbol{x}} r_i(\mathbf{0}, \beta) = c(\beta)\boldsymbol{e}_i$, so that

$$r_i(\boldsymbol{x}, \beta) = c(\beta)x_i + \boldsymbol{x}^T \partial_{\boldsymbol{x}}^2 r_i(0, \beta)\boldsymbol{x} + \mathcal{O}(\boldsymbol{x}^3).$$

Evaluating at $\beta = 0$ and letting $D_i$ be the $(M-1) \times (M-1)$ matrix $\partial_{\boldsymbol{x}}^2 r_i(0, 0)$, we get

$$r_i(\boldsymbol{x}, 0) = c(0)\boldsymbol{x}_i + \boldsymbol{x}^T D_i \boldsymbol{x} + \mathcal{O}(\boldsymbol{x}^3). \tag{6.75}$$

Now we show that the diagonal of $D_i$ is identically $\mathbf{0}$. Equation (6.44) shows that $[D_i]_{ii} = 0$. This and Lemma 119.1 show that

$$\sum_{i \neq \nu} [D_i]_{\nu\nu} = 0 \tag{6.76}$$

for every $1 \leq \nu \leq (M-1)$. Lemma 119.2 shows that $[D_i]_{\nu\nu} = [D_j]_{\nu\nu}$ for every $i$ and $j$ not equal to $\nu$. This and (6.76) shows that the diagonal of $D_i$ is zero,

$$[D_i]_{\nu\nu} = 0,$$

whenever $i \neq \nu$.

To complete the proof, we again appeal to (6.44) which shows that $[D_j]_{kl} = [D_i]_{\nu\nu} = 0$ for every $j$, $k$ and $l$. Thus, $D_i$ is identically zero. □

As in Definition 51, the orientation of the branch $\left( \begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}_k, \beta^* + \beta(t) \right)$ is determined by the sign of $t\beta'(t)$ for sufficiently small $t$. Theorem 120 shows that $\beta'(0) = 0$, so that by Remark 54.4 we need to consider $\beta''(0)$ to determine whether

a branch is subcritical or supercritical. By Lemma 63 and Corollary 64, $\beta''(0)$ depends on $< \boldsymbol{x}_0, \partial^3_{\boldsymbol{x}} r(\mathbf{0},0)[\boldsymbol{x}_0,\boldsymbol{x}_0,\boldsymbol{x}_0] >$, where $\boldsymbol{x}_0$ is the bifurcating direction for solutions to $r = 0$. That is, $W\boldsymbol{x}_0 = \boldsymbol{u}_k$, where $\boldsymbol{u}_k$, defined in (6.57), is the bifurcating direction of equilibria of (3.18). We explicitly compute the multilinear form $< \boldsymbol{x}_0, \partial^3_{\boldsymbol{x}} r(\mathbf{0},0)[\boldsymbol{x}_0,\boldsymbol{x}_0,\boldsymbol{x}_0] >$ in terms of the original problem (3.18) in the next theorem.

THEOREM 121. *If Assumption 81 holds, then* $\langle \boldsymbol{x}_0, \partial^3_{\boldsymbol{x}} r(\mathbf{0},0)[\boldsymbol{x}_0,\boldsymbol{x}_0,\boldsymbol{x}_0]\rangle$ *is equal to*

$$(M^2 - M)((M^2 - 3M + 3)\zeta_2 - 3\zeta_1)$$

*where*

$$\begin{aligned} \zeta_1 &= \langle \boldsymbol{u}_k, \partial^3_Q \mathcal{L}[\boldsymbol{u}_k, L^{-1}E\sum_{s,t} \frac{\partial^2 \nabla_Q \mathcal{L}}{\partial q_{\nu s}\partial q_{\nu t}}[\boldsymbol{v}]_s[\boldsymbol{v}]_t]\rangle, \\ \zeta_2 &= \langle \boldsymbol{v}, f[\boldsymbol{v},\boldsymbol{v},\boldsymbol{v}]\rangle. \end{aligned}$$

*The multilinear form* $< \boldsymbol{v}, f[\boldsymbol{v},\boldsymbol{v},\boldsymbol{v}] >$ *denotes*

$$< \boldsymbol{v}, f[\boldsymbol{v},\boldsymbol{v},\boldsymbol{v}] >= \sum_{r,s,t,u\in Y} \frac{\partial^4 F(q^*,\beta^*)}{\partial q_{\nu r}\partial q_{\nu s}\partial q_{\nu t}\partial q_{\nu u}}[\boldsymbol{v}]_r[\boldsymbol{v}]_s[\boldsymbol{v}]_t[\boldsymbol{v}]_u,$$

$\partial^3_Q \mathcal{L}$ *is evaluated at* $(q^*,\lambda^*,\beta^*)$, $Q = \begin{pmatrix} q \\ \lambda \end{pmatrix}$, $\boldsymbol{v}$ *is defined in (6.19),* $\boldsymbol{u}_k$ *is the bifurcating direction from (6.57), and* $\nu$ *is any class in* $\mathcal{U}$.

*Proof.* Assumption 81 is required because we assume the specific basis from Theorem 87 when decomposing $\boldsymbol{u}_k \in \Delta\mathcal{L}(q^*)$. By definition of the Liapunov Schmidt reduction (6.36), there exists a $\boldsymbol{u}_k \in \Delta\mathcal{L}(q^*)$ such that $W\boldsymbol{x}_0 = \boldsymbol{u}_k$. By Lemma 66, $\langle \boldsymbol{x}_0, \partial^3_{\boldsymbol{x}} r(\mathbf{0},0)[\boldsymbol{x}_0,\boldsymbol{x}_0,\boldsymbol{x}_0]\rangle$ is equal to

$$\langle \boldsymbol{u}_k, \partial^3_Q \mathcal{F}(\mathbf{0},0)[\boldsymbol{u}_k,\boldsymbol{u}_k,\boldsymbol{u}_k] - 3\partial^2_Q \mathcal{F}(\mathbf{0},0)[\boldsymbol{u}_k, L^{-1}E\partial^2_Q \mathcal{F}(\mathbf{0},0)[\boldsymbol{u}_k,\boldsymbol{u}_k]]\rangle.$$

Using the definition of $\mathcal{F}$ in (6.34), this becomes

$$\begin{aligned} &\langle \boldsymbol{u}_k, \partial^4_Q \mathcal{L}(q^*,\lambda^*,\beta^*)[\boldsymbol{u}_k,\boldsymbol{u}_k,\boldsymbol{u}_k]\rangle \\ - \;& 3\langle \boldsymbol{u}_k, \partial^3_Q \mathcal{L}(q^*,\lambda^*,\beta^*)[\boldsymbol{u}_k, L^{-1}E\partial^3_Q \mathcal{L}(q^*,\lambda^*,\beta^*)[\boldsymbol{u}_k,\boldsymbol{u}_k]]\rangle. \end{aligned} \tag{6.77}$$

The first term of (6.77) can be rewritten as

$$\langle \hat{\boldsymbol{u}}_k, \partial^4_{qqqq} F(q^*,\beta^*)[\hat{\boldsymbol{u}}_k,\hat{\boldsymbol{u}}_k,\hat{\boldsymbol{u}}_k]\rangle$$

using (6.57). The component form is

$$\sum_{\nu,\delta,\eta,\omega\in Y_N} \sum_{r,s,t,u\in Y} \frac{\partial^4 F(q^*,\beta^*)}{\partial q_{\nu r}\partial q_{\delta s}\partial q_{\eta t}\partial q_{\omega u}}[\hat{\boldsymbol{u}}_k]_{\nu r}[\hat{\boldsymbol{u}}_k]_{\delta s}[\hat{\boldsymbol{u}}_k]_{\eta t}[\hat{\boldsymbol{u}}_k]_{\omega u}. \tag{6.78}$$

Recall that $\frac{\partial^2 F}{\partial q_{\nu r}\partial q_{\delta s}} = 0$ if $\nu \neq \delta$ (see (3.9)), and so $\frac{\partial^4 F(q^*,\beta^*)}{\partial q_{\nu r}\partial q_{\delta s}\partial q_{\eta t}\partial q_{\omega u}} = 0$ unless $\nu = \delta = \eta = \omega$. This and (6.56) allow us to simplify (6.78) as

$$(M-1)^4 \sum_{r,s,t,u\in Y} \frac{\partial^4 F}{\partial q_{\mu r}\partial q_{\mu s}\partial q_{\mu t}\partial q_{\mu u}}[\boldsymbol{v}]_r[\boldsymbol{v}]_s[\boldsymbol{v}]_t[\boldsymbol{v}]_u + \sum_{\nu\in\mathcal{U}\setminus\{\mu\}}\sum_{r,s,t,u\in Y} \frac{\partial^4 F}{\partial q_{\nu r}\partial q_{\nu s}\partial q_{\nu t}\partial q_{\nu u}}[\boldsymbol{v}]_r[\boldsymbol{v}]_s[\boldsymbol{v}]_t[\boldsymbol{v}]_u \tag{6.79}$$

where $\mu$ is the $k^{th}$ unresolved class of $\mathcal{U}$ and $\partial^4_{qqqq}F$ is evaluated at $(q^*,\beta^*)$. Since $\frac{\partial^2 F}{\partial q_{\nu r}\partial q_{\nu s}} = \frac{\partial^2 F}{\partial q_{\mu r}\partial q_{\mu s}}$, then $\frac{\partial^4 F}{\partial q_{\nu r}\partial q_{\nu s}\partial q_{\nu t}\partial q_{\nu u}} = \frac{\partial^4 F}{\partial q_{\mu r}\partial q_{\mu s}\partial q_{\mu t}\partial q_{\mu u}}$ for any $\nu,\mu \in \mathcal{U}$. Since $|\mathcal{U}| = M$, then (6.79) becomes

$$((M-1)^4 + (M-1)) \sum_{r,s,t,u\in Y} \frac{\partial^4 F(q^*,\beta^*)}{\partial q_{\nu r}\partial q_{\nu s}\partial q_{\nu t}\partial q_{\nu u}}[\boldsymbol{v}]_r[\boldsymbol{v}]_s[\boldsymbol{v}]_t[\boldsymbol{v}]_u.$$

Observe that $(M-1)^4 + (M-1) = (M^2-M)(M^2-3M+2)$.

Now we consider the second term of (6.77)

$$-3\langle \boldsymbol{u}_k, \partial^3_Q\mathcal{L}(\mathbf{0},0)[\boldsymbol{u}_k, L^{-1}E\partial^3_Q\mathcal{L}(\mathbf{0},0)[\boldsymbol{u}_k,\boldsymbol{u}_k]]\rangle.$$

In particular, we examine the $(NK+K)\times 1$ vector

$$L^{-1}E\partial^3_Q\mathcal{L}(\mathbf{0},0)[\boldsymbol{u}_k,\boldsymbol{u}_k] \;=\; L^{-1}E\sum_{\delta,\eta\in Y_N}\sum_{r,s\in Y}\frac{\partial^2\nabla_Q\mathcal{L}}{\partial q_{\delta r}\partial q_{\eta s}}[\boldsymbol{u}_k]_{\delta r}[\boldsymbol{u}_k]_{\eta s}. \tag{6.80}$$

Note that the derivatives with respect to $\lambda$ on the left hand side of (6.80) are ignored since they are zero. Now, using (6.56) as before, we rewrite (6.80) as

$$((M-1)^2 + (M-1))L^{-1}E\sum_{\delta,\eta\in Y_N}\sum_{r,s\in Y}\frac{\partial^2\nabla_Q\mathcal{L}}{\partial q_{\delta r}\partial q_{\eta s}}[\boldsymbol{v}]_r[\boldsymbol{v}]_s.$$

Since $(M-1)^2 + (M-1) = M^2 - M$, we are done. □

Remark 122. *The term*

$$-3\langle \boldsymbol{u}_k, \partial^3_Q\mathcal{L}[\boldsymbol{u}_k, L^{-1}E\sum_{r,s}\frac{\partial^2\nabla_Q\mathcal{L}}{\partial q_{\nu r}\partial q_{\nu s}}[\boldsymbol{v}]_r[\boldsymbol{v}]_s]\rangle$$

*in Theorem 121 can not be written in terms of $F$ due to multiplication by the $(NK+K)\times(NK+K)$ matrix $L^{-1}E$.*

Definition 123. *The* discriminant *of the bifurcating branch,*

$$\left(\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}_k, \beta^* + \beta(t)\right),$$

*is defined as*

$$\begin{aligned}\zeta(q^*,\beta^*,\boldsymbol{u}_k) &= 3\langle \boldsymbol{u}_k, \partial_Q^3\mathcal{L}[\boldsymbol{u}_k, L^{-1}E\sum_{r,s}\frac{\partial^2\nabla_Q\mathcal{L}}{\partial q_{\nu r}\partial q_{\nu s}}[\boldsymbol{v}]_r[\boldsymbol{v}]_s]\rangle \\ &\quad -(M^2-3M+3)\langle \boldsymbol{v}, f[\boldsymbol{v},\boldsymbol{v},\boldsymbol{v}]\rangle,\end{aligned} \tag{6.81}$$

*where the derivatives of* $\mathcal{L}$ *are evaluated at* $(q^*,\lambda^*,\beta^*)$*, and*

$$< \boldsymbol{v}, f[\boldsymbol{v},\boldsymbol{v},\boldsymbol{v}] >= \sum_{r,s,t,u\in Y}\frac{\partial^4 F(q^*,\beta^*)}{\partial q_{\nu r}\partial q_{\nu s}\partial q_{\nu t}\partial q_{\nu u}}[\boldsymbol{v}]_r[\boldsymbol{v}]_s[\boldsymbol{v}]_t[\boldsymbol{v}]_u.$$

We now have the following result.

COROLLARY 124. *If* $\Delta D(q^*)$ *is positive definite on* $\ker\Delta F(q^*)$*, and if Assumption 81 holds, then* $\operatorname{sgn}(\beta''(0)) = \operatorname{sgn}(\zeta(q^*,\beta^*,\boldsymbol{u}_k))$

*Proof.* Corollary 64 and Theorem 121. □

REMARK 125. *If* $\Delta D(q^*)$ *is negative definite on* $\ker\Delta F(q^*)$*, then Lemma 63 shows that*

$$\operatorname{sgn}(\beta''(0)) = -\operatorname{sgn}(\zeta(q^*,\beta^*,\boldsymbol{u}_k)).$$

The following lemma provides a way to compute the discriminant, $\zeta(q^*,\beta^*,\boldsymbol{u}_k)$, for the Information Distortion problem (2.34), where $F = H(q) + \beta D_{eff}(q)$.

LEMMA 126. *For the Information Distortion problem (2.34), the sign of* $\frac{\partial^3\mathcal{L}}{\partial q_{\nu r}\partial q_{\nu s}\partial q_{\nu t}}$ *is equal to*

$$\delta_{rst}\frac{p(y_r)}{q_{\nu r}^2} + \beta\left(\frac{p(y_r)p(y_s)p(y_t)}{(\sum_j p(y_j)q_{\nu j})^2} - \sum_i \frac{p(x_i,y_r)p(x_i,y_s)p(x_i,y_t)}{(\sum_j p(x_i,y_j)q_{\nu j})^2}\right).$$

*The sign of the expression* $\frac{\partial^4 F}{\partial q_{\nu r}\partial q_{\nu s}\partial q_{\nu t}\partial q_{\nu u}}$ *is equal to*

$$2\beta\left(\sum_i \frac{p(x_i,y_r)p(x_i,y_s)p(x_i,y_t)p(x_i,y_u)}{(\sum_j p(x_i,y_j)q_{\nu j})^3} - \frac{p(y_r)p(y_s)p(y_t)p(y_u)}{(\sum_j p(y_j)q_{\nu j})^3}\right) - 2\delta_{rstu}\frac{p(y_r)}{q_{\nu r}^3}.$$

*Proof.* The lemma follows from (2.21),(2.22),(2.24), and (2.25) □

We now present the results for the general case which determine whether a bifurcating branch from symmetry breaking bifurcation is subcritical or supercritical.

THEOREM 127. *Suppose that Assumption 81 holds and that* $\Delta D(q^*)$ *is positive definite on* $\ker\Delta F(q^*)$*. If* $\zeta(q^*,\beta^*,\boldsymbol{u}_k) < 0$*, then the bifurcating branch*

$$\left(\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}_k, \beta^* + \beta(t)\right),$$

*guaranteed by Theorem 110, is subcritical and consists of unstable solutions.*

*Proof.* The theorem follows from Lemma 115, Corollary 124, Remark 54.4, and Proposition 55. □

THEOREM 128. *Suppose that Assumption 81 holds and that $\Delta D(q^*)$ is positive definite on $\ker \Delta F(q^*)$. If $\zeta(q^*, \beta^*, \boldsymbol{u}_k) > 0$, then the bifurcating branch*

$$\left( \begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}_k, \beta^* + \beta(t) \right),$$

*guaranteed by Theorem 110, is supercritical. Furthermore, if*

$$\theta(q^*, \beta^*, \boldsymbol{u}_k) := \sum_m \langle \boldsymbol{w}_m, \theta_1 - 2\theta_2 - \theta_3 \rangle > 0$$

*where*

$$\begin{aligned} \theta_1 &= \partial_Q^4 \mathcal{L}[\boldsymbol{u}_k, \boldsymbol{u}_k, \boldsymbol{w}_m], \\ \theta_2 &= \partial_Q^3 \mathcal{L}[\boldsymbol{u}_k, L^{-1} E \partial_Q^3 \mathcal{L}[\boldsymbol{u}_k, \boldsymbol{w}_m]], \\ \theta_3 &= \partial_Q^3 \mathcal{L}[\boldsymbol{w}_m, L^{-1} E \partial_Q^3 \mathcal{L}[\boldsymbol{u}_k, \boldsymbol{u}_k]], \end{aligned}$$

*where $Q = \begin{pmatrix} q \\ \lambda \end{pmatrix}$, then the branch consists of unstable solutions.*

*Proof.* Lemma 115, Corollary 124 and Remark 54.4 show that the branch is supercritical. By Theorem 120, we can now invoke Proposition 65. The proof is complete once we show that $\sum_{i,j,m} \frac{\partial^3 r_m(\mathbf{0},0)}{\partial x_i \partial x_j \partial x_m} [\boldsymbol{x}_0]_i [\boldsymbol{x}_0]_j$ is equal to $\theta(q^*, \beta^*, \boldsymbol{u}_k)$.

From (5.46), we have that $\frac{\partial^3 r_m(\mathbf{0},0)}{\partial x_i \partial x_j \partial x_m}$ is equal to

$$\langle \boldsymbol{w}_m, \partial^4 \mathcal{L}[\boldsymbol{w}_i, \boldsymbol{w}_j, \boldsymbol{w}_m] - LE^{-1}(\partial^3 \mathcal{L}[\boldsymbol{w}_j, \boldsymbol{w}_m] + \partial^3 \mathcal{L}[\boldsymbol{w}_i, \boldsymbol{w}_m] + \partial^3 \mathcal{L}[\boldsymbol{w}_i, \boldsymbol{w}_j]) \rangle. \quad (6.82)$$

The theorem now follows from the linearity of each of the multilinear forms in (6.82). We show this explicitly for the first term. To get $\sum_{i,j,m} \frac{\partial^3 r_m(\mathbf{0},0)}{\partial x_i \partial x_j \partial x_m} [\boldsymbol{x}_0]_i [\boldsymbol{x}_0]_j$, we first simplify $\sum_{ij} \langle \boldsymbol{w}_m, \partial^4 \mathcal{L}[\boldsymbol{w}_i, \boldsymbol{w}_j, \boldsymbol{w}_m] \rangle [\boldsymbol{x}_0]_i [\boldsymbol{x}_0]_j$, which is

$$\sum_{i,j} \sum_{r,s,t,u} \partial^4 \mathcal{L}_{rstu} [\boldsymbol{w}_m]_r [\boldsymbol{w}_i]_s [\boldsymbol{w}_j]_t [\boldsymbol{w}_m]_u [\boldsymbol{x}_0]_i [\boldsymbol{x}_0]_j. \quad (6.83)$$

Since

$$[\boldsymbol{u}_k]_s = \sum_i [\boldsymbol{w}_i]_s [\boldsymbol{x}_0]_s \text{ and } [\boldsymbol{u}_k]_t = \sum_j [\boldsymbol{w}_j]_t [\boldsymbol{x}_0]_t, \quad (6.84)$$

then the term (6.83) is $\langle \boldsymbol{w}_m, \theta_1 \rangle$. Using the same observation as in (6.84) for the second, third and fourth terms of (6.82), the theorem is proved □

The following Theorem shows that if a bifurcating branch corresponds to an eigenvalue of $\Delta\mathcal{L}(q^*)$ changing from negative to positive, then the branch consists of stationary points $(q^*, \beta^*)$ which are not solutions of (1.9). This is a nontrivial result. In general, if $(q^*, \lambda^*, \beta^*)$ is an equilibrium of (3.18) such that $\Delta\mathcal{L}(q^*)$ has a positive eigenvalue, then $q^*$ may or may not be a solution to the optimization problem (3.1) at $\beta = \beta^*$ (see Remark 27).

THEOREM 129. *Suppose that Assumption 81 holds. If*

$$\left( \begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}_k, \beta^* + \beta(t) \right)$$

*is a bifurcating branch, guaranteed by Theorem 110, then $\boldsymbol{u}_k$ is an eigenvector of $\Delta_{q,\lambda}\mathcal{L}(\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}_k, \beta^* + \beta(t))$ for sufficiently small $t$. Furthermore, if the corresponding eigenvalue is positive, then the branch consists of stationary points which are not solutions to (3.1).*

*Proof.* By assumption, bifurcation occurs at the $M$-uniform solution $(q^*, \lambda^*, \beta^*)$ which is fixed by $\Gamma_{\mathcal{U}}$, and $\boldsymbol{u}_k = \begin{pmatrix} \hat{\boldsymbol{u}}_k \\ \mathbf{0} \end{pmatrix} \in \ker \Delta\mathcal{L}(q^*)$ (see (6.56) and (6.57)). We first show that $\boldsymbol{u}_k$ is an eigenvector of $\Delta_{q,\lambda}\mathcal{L}(q^* + t\hat{\boldsymbol{u}}_k, \lambda^*, \beta + \beta(t))$ for small $t$. Let $Q = \begin{pmatrix} q \\ \lambda \end{pmatrix}$ and let

$$L(Q, \beta) := \nabla_{q,\lambda}\mathcal{L}(q^* + q, \lambda^* + \lambda, \beta^* + \beta).$$

Thus, bifurcation of solutions to

$$L(Q, \beta) = \mathbf{0}$$

occurs at $(\mathbf{0}, 0)$. By Lemma 100, $\boldsymbol{u}_k$ is the sole basis vector of $\mathrm{Fix}\langle T_k \rangle$, where $\langle T_k \rangle < \Gamma_{\mathcal{U}}$ is isomorphic to $S_{M-1}$. By Lemma 44,

$$L(t\boldsymbol{u}_k, \beta) = h(t, \beta)\boldsymbol{u}_k$$

for some scalar function $h(t, \beta)$. Taking the derivative of this equation with respect to $t$, we get

$$\partial_Q L(t\boldsymbol{u}_k, \beta)\boldsymbol{u}_k = \partial_t h(t, \beta))\boldsymbol{u}_k, \tag{6.85}$$

from which it follows that $\boldsymbol{u}_k$ is an eigenvector of $\Delta_{q,\lambda}\mathcal{L}(q^* + t\hat{\boldsymbol{u}}_k, \lambda^*, \beta + \beta(t))$, with corresponding eigenvalue

$$\xi = \partial_t h(t, \beta).$$

We now show that if $\xi > 0$, then the bifurcating branch consists of stationary points which are not solutions to (3.1). Using (3.8) and letting $\widehat{\Delta F} := \Delta F(q^* + t\hat{\boldsymbol{u}}_k, \beta + \beta(t))$, we see that (6.85) can be rewritten as

$$\begin{pmatrix} \widehat{\Delta F} & J^T \\ J & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{u}}_k \\ \mathbf{0} \end{pmatrix} = \xi \begin{pmatrix} \hat{\boldsymbol{u}}_k \\ \mathbf{0} \end{pmatrix},$$

which shows that

$$\begin{aligned}\widehat{\Delta F}\hat{\boldsymbol{u}}_k &= \xi\hat{\boldsymbol{u}}_k \\ J\hat{\boldsymbol{u}}_k &= \mathbf{0}.\end{aligned}$$

Thus, $\hat{\boldsymbol{u}}_k$ is an eigenvector of $\Delta F(q^* + t\hat{\boldsymbol{u}}_k, \beta + \beta(t))$ with corresponding positive eigenvalue $\xi$, and $\hat{\boldsymbol{u}}_k \in \ker J$. The desired result now follows from Theorem 20. □

The Theory Applied to the Information Bottleneck

For the Information Bottleneck problem (2.35),

$$\max_{q\in\Delta} F(q,\beta) = \max_{q\in\Delta}(I(Y;Y_N) + \beta I(X;Y_N)),$$

Assumptions 81.2 and 81.3 are never satisfied. This is due to the fact that $q$ is always in the kernel of $\Delta F(q,\beta)$ for every $\beta$ (Theorem 43). In particular, this shows that on the $N$-uniform solution branch $(q_{\frac{1}{N}}, \lambda, \beta)$, we have that the $K \times 1$ vector of $\frac{1}{N}$'s, is in the kernel of each of the identical blocks of $\Delta F(q(y_N|y),\beta)$ for every $\beta$. In this section, we review some results which deal with this scenario at bifurcation.

Consider any problem of the form (1.9),

$$\max_{q\in\Delta}(G(q) + \beta D(q)),$$

where Assumption 81.2 is replaced by the condition that for $B$, the blocks of the Hessian defined in (6.3), then

$$\ker B \text{ has dimension 2 with } K\times 1 \text{ basis vectors } \{\boldsymbol{v},\boldsymbol{z}\}. \tag{6.86}$$

Suppose that bifurcation of the problem (1.9) occurs at $(q_{\frac{1}{N}}, \lambda^*, \beta^*)$ when (6.86) holds. Observe that all of the blocks of the Hessian $\Delta F(q_{\frac{1}{N}})$ are identical, and so Assumptions 81.3 and 81.4 are not required. We review the following conditions, which must hold at $(q_{\frac{1}{N}}, \lambda^*, \beta^*)$, without proof:

1. The space $\ker \Delta F(q_{\frac{1}{N}})$ has dimension $2N$.
2. The basis of $\ker \Delta F(q_{\frac{1}{N}})$ is $\{\{\boldsymbol{v}_i\}_{i=1}^N, \{\boldsymbol{z}_i\}_{i=1}^N\}$, where $\boldsymbol{v}_i$ and $\boldsymbol{z}_i$ are defined as in (6.21).
3. The space $\ker \Delta\mathcal{L}(q_{\frac{1}{N}})$ has dimension $2N-2$.
4. The basis of $\ker \Delta\mathcal{L}(q_{\frac{1}{N}})$ is $\{\boldsymbol{w}_i\}$ where

$$\boldsymbol{w}_i = \begin{cases} \begin{pmatrix} \boldsymbol{v}_i \\ \mathbf{0}\end{pmatrix} - \begin{pmatrix} \boldsymbol{v}_N \\ \mathbf{0}\end{pmatrix} & \text{for } i = 1,...,N-1 \\ \begin{pmatrix} \boldsymbol{z}_{i-N+1} \\ \mathbf{0}\end{pmatrix} - \begin{pmatrix} \boldsymbol{z}_N \\ \mathbf{0}\end{pmatrix} & \text{for } i = N,...,2N-2 \end{cases}$$

5. The group $\mathcal{A} \cong S_N$, for which the Liapunov-Schmidt reduction $r(\boldsymbol{x}, \beta)$ is equivariant, is a subgroup of the group of all $(2N-2) \times (2N-2)$ matrices.

6. The subspaces $\text{span}(\{\boldsymbol{w}_i\}_{i=1}^{N-1})$ and $\text{span}(\{\boldsymbol{w}_i\}_{i=N}^{2N-2})$ are invariant subspaces of $\ker \Delta\mathcal{L}(q_{\frac{1}{N}})$, which shows that $\ker \Delta\mathcal{L}(q_{\frac{1}{N}})$ is not irreducible and that $\mathcal{A}$ is not absolutely irreducible.

7. The group $\langle T_k \rangle \cong \Gamma_{\mathcal{U}} < \Gamma$, which is isomorphic to $S_{N-1}$, is a maximal isotropy subgroup, and it has a two dimensional fixed point space.

8. The fixed point space $\text{Fix}\langle T_k \rangle$ has basis $\{\boldsymbol{a}_k, \boldsymbol{b}_k\}$, where

$$\boldsymbol{a}_k = \begin{pmatrix} -\boldsymbol{v} \\ \vdots \\ -\boldsymbol{v} \\ (N-1)\boldsymbol{v} \\ -\boldsymbol{v} \\ \vdots \\ -\boldsymbol{v} \\ \boldsymbol{0} \end{pmatrix}, \boldsymbol{b}_k = \begin{pmatrix} -\boldsymbol{z} \\ \vdots \\ -\boldsymbol{z} \\ (N-1)\boldsymbol{z} \\ -\boldsymbol{z} \\ \vdots \\ -\boldsymbol{z} \\ \boldsymbol{0} \end{pmatrix},$$

and $(N-1)\boldsymbol{v}$ and $(N-1)\boldsymbol{z}$ are in the $k^{th}$ row of $\boldsymbol{a}_k$ and $\boldsymbol{b}_k$ respectively.

# CHAPTER 7

## CONTINUATION

In chapter 6, we developed the theory which gives the existence, as well as the structure, of bifurcating branches from symmetry breaking bifurcation of $M$-uniform solutions. We would like to numerically confirm this theory. To do this, we employed *continuation* techniques [6, 26] to analyze the explicit behavior of the equilibria of the dynamical system (3.18)

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda}\mathcal{L}(q, \lambda, \beta)$$

for the Information Distortion problem. Continuation techniques are numerical methods for tracking equilibria of a general dynamical system (3.15),

$$\dot{\boldsymbol{x}} = \psi(\boldsymbol{x}, \beta),$$

by constructing a sequence of equilibria $\{(\boldsymbol{x}_k, \beta_k)\}$ which reside on some solution branch. The scalar $\beta \in \Re$ is considered a *continuation parameter*. This is analogous to the scenario when using Algorithm 1 to solve (1.9), where $\beta$ could be interpreted as an *annealing parameter*.

Recall that equilibria $(\boldsymbol{x}^*, \beta)$ of (3.15) satisfy

$$\psi(\boldsymbol{x}, \beta) = \mathbf{0}.$$

If $(\boldsymbol{x}_k, \beta_k)$ is some equilibrium on some branch, then continuation techniques compute the vector tangent to the curve $\psi(\boldsymbol{x}, \beta) = \mathbf{0}$ to give an initial guess $(\boldsymbol{x}_{k+1}^{(0)}, \beta_{k+1}^{(0)})$ for Newton's method, which computes the equilibrium $(\boldsymbol{x}_{k+1}, \beta_{k+1})$ on the same branch as $(\boldsymbol{x}_k, \beta_k)$ for some $\beta_{k+1}$ close to $\beta_k$. If bifurcation is detected, then one might choose to continue along some particular bifurcating branch. This is effected by a *branch switch*.

### Parameter Continuation

Parameter continuation is the simplest type of continuation, an algorithm for which is given at the end of this section (Algorithm 130). It uses the tangent vector $\partial_\beta \boldsymbol{x}_{k-1}$ at $(\boldsymbol{x}_{k-1}, \beta_{k-1})$ to compute an initial guess $(\boldsymbol{x}_k^{(0)}, \beta_k)$ for the equilibrium $(\boldsymbol{x}_k, \beta_k)$ by setting

$$\begin{pmatrix} \boldsymbol{x}_k^{(0)} \\ \beta_k \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{k-1} + \Delta\beta\partial_\beta \boldsymbol{x}_{k-1} \\ \beta_{k-1} + \Delta\beta \end{pmatrix} \tag{7.1}$$

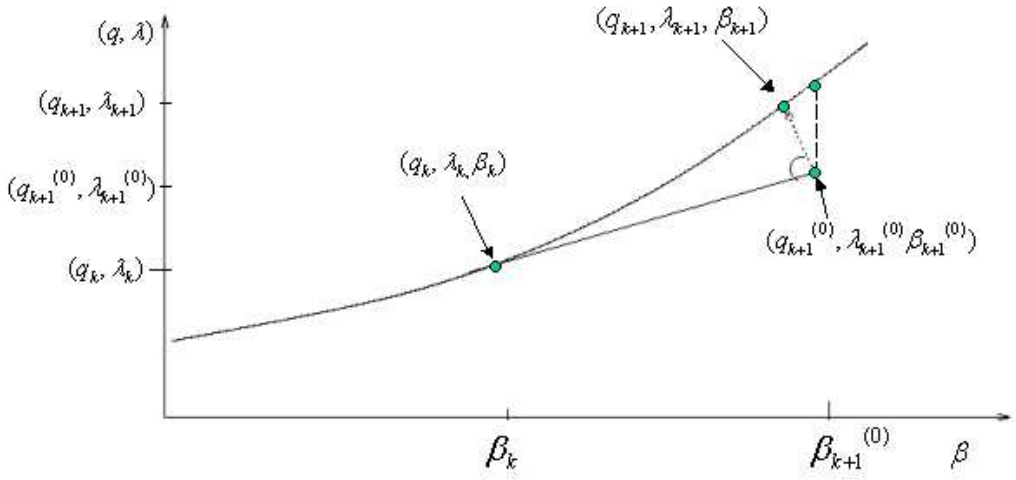



Figure 15. Conceptual figure depicting continuation along the curve $\nabla_{q,\lambda}\mathcal{L}(q,\lambda,\beta) = \mathbf{0}$. From the point $(q_{k+1}^{(0)}, \lambda_{k+1}^{(0)}, \beta_{k+1}^{(0)})$, the dashed line indicates the path taken by parameter continuation. The dotted line indicates the path taken by pseudoarclength continuation as the points $\{(q_{k+1}^{(i)}, \lambda_{k+1}^{(i)}, \beta_{k+1}^{(i)})\}_i$ converge to $(q_{k+1}, \lambda_{k+1}, \beta_{k+1}\}$.

for some $\Delta\beta > 0$. Given this guess, Newton's method is used to determine $(\boldsymbol{x}_k, \beta_k)$. Thus, $\beta_k$ is kept fixed as we search for $\boldsymbol{x}_k$ (see Figure 15).

We proceed by showing how to compute the vector, $\partial_\beta \boldsymbol{x}_k$, which is tangent to the curve $\psi(\boldsymbol{x}, \beta) = \mathbf{0}$ at the equilibrium $(\boldsymbol{x}_k, \beta_k)$ when $\partial_{\boldsymbol{x}}\psi(\boldsymbol{x}_k, \beta_k)$ is nonsingular. By the Implicit Function Theorem, we can take the total derivative of $\psi = \mathbf{0}$ with respect to $\beta$, which shows that

$$\frac{\partial}{\partial\beta}\psi(\boldsymbol{x}, \beta) = \mathbf{0}$$

so that

$$\partial_{\boldsymbol{x}}\psi(\boldsymbol{x}, \beta)\partial_\beta \boldsymbol{x}(\beta) + \partial_\beta\psi(\boldsymbol{x}, \beta) = \mathbf{0}. \tag{7.2}$$

Thus, the tangent vector at an equilibrium $(\boldsymbol{x}_k, \beta_k)$ is found by solving

$$\partial_{\boldsymbol{x}}\psi(\boldsymbol{x}_k, \beta_k)\partial_\beta \boldsymbol{x}(\beta_k) = -\partial_\beta\psi(\boldsymbol{x}_k, \beta_k) \tag{7.3}$$

which shows that

$$\partial_\beta \boldsymbol{x}(\beta_k) = -\partial_{\boldsymbol{x}}\psi(\boldsymbol{x}_k, \beta_k)^{-1}\partial_\beta\psi(\boldsymbol{x}_k, \beta_k).$$

In practice, the tangent vector

$$\partial_\beta \boldsymbol{x}_k := \partial_\beta \boldsymbol{x}(\beta_k) \tag{7.4}$$

is found by solving (7.3).

Newton's method is used to find the next equilibrium $(\boldsymbol{x}_k, \beta_k)$ since this method is not dependent on the stability of $(\boldsymbol{x}_k, \beta_k)$. Newton's method can be used to find solutions of any equation

$$\psi(\boldsymbol{x}, \beta) = \mathbf{0}$$

by considering a sequence of linear approximations $\{\hat{\psi}_i\}$ to $\psi$, and determining the solutions of

$$\hat{\psi}_i(\boldsymbol{x}, \beta) = \mathbf{0}$$

for each of these. By Taylor's Theorem, the linear approximation of $\psi$ about $\boldsymbol{x}_k^{(i)}$ for a fixed $\beta$ is

$$\hat{\psi}_i(\boldsymbol{x}, \beta) = \partial_{\boldsymbol{x}}\psi(\boldsymbol{x}_k^{(i)}, \beta)(\boldsymbol{x} - \boldsymbol{x}_k^{(i)}) + \psi(\boldsymbol{x}_k^{(i)}, \beta).$$

Thus, the solution, $\boldsymbol{x}_k^{(i+1)}$, of $\hat{\psi} = \mathbf{0}$ at $\beta_k$ is found by solving

$$\partial_{\boldsymbol{x}}\psi(\boldsymbol{x}_k^{(i)}, \beta_k)(\boldsymbol{x}_k^{(i+1)} - \boldsymbol{x}_k^{(i)}) = -\psi(\boldsymbol{x}_k^{(i)}, \beta_k). \tag{7.5}$$

In this way, if $\partial_{\boldsymbol{x}}\psi(\boldsymbol{x}_k^{(i)}, \beta_k)$ is nonsingular for each $i$, and if $\boldsymbol{x}_k^{(0)}$ is sufficiently close to $\boldsymbol{x}_k$, then [6]

$$\lim_{i\to\infty} \boldsymbol{x}_k^{(i)} \to \boldsymbol{x}_k.$$

We conclude the previous discussion with the following algorithm.

Algorithm 130 (Parameter continuation). *[6, 26] Suppose that $(\boldsymbol{x}_0, \beta_0)$ is a given equilibria to (3.15). Let $\Delta\beta > 0$. For $k \geq 0$, iterate the following steps until $\beta_k = \mathcal{B}$ for some $\mathcal{B} > 0$.*

1. *Find the tangent vector $\partial_\beta \boldsymbol{x}_k$ from (7.4) by solving (7.3).*
2. *Get the initial guess $\boldsymbol{x}_{k+1}^{(0)}$ for $\boldsymbol{x}_{k+1}$ from (7.1) and set $\beta_{k+1} = \beta_k + \Delta\beta$.*
3. *Find the equilibrium $\boldsymbol{x}_{k+1}$ using the initial guess $\boldsymbol{x}_{k+1}^{(0)}$ by iterating Newton's method (7.5), giving $\{\boldsymbol{x}_{k+1}^{(i)}\}_i \to \boldsymbol{x}_{k+1}$.*

## Pseudoarclength Continuation

This method, due to Keller [39], uses Newton's method to find the next equilibrium $(\boldsymbol{x}_k, \beta_k)$ by allowing both $\boldsymbol{x}$ and $\beta$ to vary. The explicit algorithm is given at the end of this section (Algorithm 131). The advantage of this approach is twofold. First, the step size in $\beta$, $\Delta\beta_{k+1} = \beta_{k+1} - \beta_k$, changes depending on the "steepness" of the curve $\psi(\boldsymbol{x}_k, \beta_k) = 0$. Secondly, since $\beta$ is varying, this method allows for continuation of equilibria around a saddle-node bifurcation (see Figure 15).

Pseudoarclength continuation works in two steps. First, we parameterize $\boldsymbol{x}$ and $\beta$ with respect to some variable $s$, so that the tangent vector to $\psi = \mathbf{0}$ at $(\boldsymbol{x}_k, \beta_k)$, $\begin{pmatrix} \partial_s \boldsymbol{x}_k \\ \partial_s \beta_k \end{pmatrix}$, is found by taking the total derivative as in (7.2),

$$\partial_{\boldsymbol{x}} \psi(\boldsymbol{x}_k, \beta_k) \partial_s \boldsymbol{x}(s_k) + \partial_\beta \psi(\boldsymbol{x}_k, \beta_k) \partial_s \beta(s_k) = \mathbf{0}, \tag{7.6}$$

and solving for $\partial_s \boldsymbol{x}(s_k)$ when the scalar $\partial_s \beta(s_k) = 1$. Thus, we determine $\partial_s \boldsymbol{x}(s_k)$ as in (7.3)

$$\partial_{\boldsymbol{x}} \psi(\boldsymbol{x}_k, \beta_k) \partial_s \boldsymbol{x}(s_k) = -\partial_\beta \psi(\boldsymbol{x}_k, \beta_k). \tag{7.7}$$

Setting $\partial_s \beta(s_k) = 1$ is justified by the following argument. If we set $\partial_s \hat{\beta}(s_k) = a \neq 0$, then $\partial_s \hat{\boldsymbol{x}}(s_k) = -a \partial_{\boldsymbol{x}} \psi(\boldsymbol{x}_k, \beta_k)^{-1} \partial_\beta \psi(\boldsymbol{x}_k, \beta_k) = a \partial_s \boldsymbol{x}(s_k)$. Thus,

$$\begin{pmatrix} \partial_s \boldsymbol{x}(s_k) \\ 1 \end{pmatrix} = \frac{1}{a} \begin{pmatrix} \partial_s \hat{\boldsymbol{x}}(s_k) \\ \partial_s \hat{\beta}(s_k) \end{pmatrix}.$$

Therefore, these vectors are equivalent up to a scaling factor, which we may ignore since we will normalize in (7.9).

In order that subsequent tangent vectors $\left\{ \begin{pmatrix} \partial_s \boldsymbol{x}(s_k) \\ 1 \end{pmatrix}, \begin{pmatrix} \partial_s \boldsymbol{x}(s_{k-1}) \\ 1 \end{pmatrix} \right\}_k$ always have the same orientation, if we let

$$\theta_k = \angle \left( \begin{pmatrix} \partial_s \boldsymbol{x}(s_k) \\ 1 \end{pmatrix}, \begin{pmatrix} \partial_s \boldsymbol{x}(s_{k-1}) \\ 1 \end{pmatrix} \right), \tag{7.8}$$

then we require that

$$-\frac{\pi}{2} \leq \theta_k \leq \frac{\pi}{2}.$$

Thus, the normalized tangent vector $(\partial_s \boldsymbol{x}_k^T \;\; \partial_s \beta_k)^T$ at $(\boldsymbol{x}_k, \beta_k)$ which has the same orientation as $(\partial_s \boldsymbol{x}(s_k)^T \;\; 1)^T$ which we will use in all of computations that follow is

$$\begin{pmatrix} \partial_s \boldsymbol{x}_k \\ \partial_s \beta_k \end{pmatrix} := \frac{\operatorname{sgn}(\cos \theta_k)}{\sqrt{\|\partial_s \boldsymbol{x}(s_k)\|^2 + 1}} \begin{pmatrix} \partial_s \boldsymbol{x}(s_k) \\ 1 \end{pmatrix}. \tag{7.9}$$

Now we see that the initial guess for $(\boldsymbol{x}_{k+1}, \beta_{k+1})$ given an equilibrium $(\boldsymbol{x}_k, \beta_k)$ is

$$\begin{pmatrix} \boldsymbol{x}_{k+1}^{(0)} \\ \beta_{k+1}^{(0)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_k + d \partial_s \boldsymbol{x}_k \\ \beta_k + d \partial_s \beta_k \end{pmatrix} \tag{7.10}$$

for some

$$d > 0. \tag{7.11}$$

The second step of the pseudoarclength method finds the next equilibrium $(\boldsymbol{x}_{k+1}, \beta_{k+1})$ using (7.10) by creating a sequence of points, $\{(\boldsymbol{x}_{k+1}^{(i)}, \beta_{k+1}^{(i)})\}_i$, that converge to $(\boldsymbol{x}_{k+1}, \beta_{k+1})$

such that the norm of the projection of the vector $\begin{pmatrix} \boldsymbol{x}_{k+1}^{(i)} - \boldsymbol{x}_k \\ \beta_{k+1}^{(i)} - \beta_k \end{pmatrix}$ onto $\begin{pmatrix} \partial_s \boldsymbol{x}_k \\ \partial_s \beta_k \end{pmatrix}$ for every $i$ is always $d$ from (7.11). To effect this constraint, we use the fact that the projection of a vector $\boldsymbol{w}$ onto a vector $\boldsymbol{v}$ is given by [65]

$$\text{proj}_{\boldsymbol{v}}(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{v}}{\|\boldsymbol{v}\|^2} \boldsymbol{v}$$

from which it follows that

$$\|\text{proj}_{\boldsymbol{v}}(\boldsymbol{w})\| = \frac{\boldsymbol{w}^T \boldsymbol{v}}{\|\boldsymbol{v}\|}.$$

Thus

$$\|\text{proj}_{(\partial_s \boldsymbol{x} \quad \partial_s \beta)^T} \begin{pmatrix} \boldsymbol{x}_{k+1}^{(i)} - \boldsymbol{x}_k \\ \beta_{k+1}^{(i)} - \beta_k \end{pmatrix} \| = d$$

for every $i$ if and only if

$$P(\boldsymbol{x}_{k+1}^{(i)}, \beta_{k+1}^{(i)}) := \begin{pmatrix} \boldsymbol{x}_{k+1}^{(i)} - \boldsymbol{x}_k \\ \beta_{k+1}^{(i)} - \beta_k \end{pmatrix}^T \begin{pmatrix} \partial_s \boldsymbol{x}_k \\ \partial_s \beta_k \end{pmatrix} = d, \tag{7.12}$$

since $\| \begin{pmatrix} \partial_s \boldsymbol{x}_k \\ \partial_s \beta_k \end{pmatrix} \| = 1$ by (7.9). So now we have the vector function

$$\Psi(\boldsymbol{x}, \beta) := \begin{pmatrix} \psi(\boldsymbol{x}, \beta) \\ P(\boldsymbol{x}, \beta) - d \end{pmatrix} \tag{7.13}$$

for which we are interested in solutions to $\Psi = \mathbf{0}$ for some $d > 0$.

We use Newton's method to solve $\Psi = \mathbf{0}$ as in (7.5), but now we must differentiate $\Psi$ with respect to the vector $\begin{pmatrix} \boldsymbol{x} \\ \beta \end{pmatrix}$, which we write as $\partial_{\boldsymbol{x},\beta} \Psi$. Hence, one can find $(\boldsymbol{x}_k^{(i+1)}, \beta_k^{(i+1)})$ given $(\boldsymbol{x}_k^{(i)}, \beta_k^{(i)})$ by solving

$$\partial_{\boldsymbol{x},\beta} \Psi(\boldsymbol{x}_k^{(i)}, \beta_k^{(i)}) \left( \begin{pmatrix} \boldsymbol{x} \\ \beta \end{pmatrix} - \begin{pmatrix} \boldsymbol{x}_k^{(i)} \\ \beta_k^{(i)} \end{pmatrix} \right) = -\Psi(\boldsymbol{x}_k^{(i)}, \beta_k^{(i)})$$

for $\boldsymbol{x}$ and $\beta$, which is equivalent to solving

$$\begin{pmatrix} \partial_{\boldsymbol{x}} \psi(\boldsymbol{x}_k^{(i)}, \beta_k^{(i)}) & \partial_\beta \psi(\boldsymbol{x}_k^{(i)}, \beta_k^{(i)}) \\ \partial_s \boldsymbol{x}_k^T & \partial_s \beta_k \end{pmatrix} \left( \begin{pmatrix} \boldsymbol{x} \\ \beta \end{pmatrix} - \begin{pmatrix} \boldsymbol{x}_k^{(i)} \\ \beta_k^{(i)} \end{pmatrix} \right) = - \begin{pmatrix} \psi(\boldsymbol{x}_k^{(i)}, \beta_k^{(i)}) \\ P(\boldsymbol{x}_k^{(i)}, \beta_k^{(i)}) - d \end{pmatrix} \tag{7.14}$$

We conclude the previous discussion with the following algorithm.

ALGORITHM 131 (PSEUDOARCLENGTH CONTINUATION). *[6, 26] Suppose that $(\boldsymbol{x}_0, \beta_0)$ is a given equilibria to (3.15). For $k \geq 0$, iterate the following steps until $\beta_k = \mathcal{B}$ for some $\mathcal{B} > 0$.*

1. *Find the tangent vector* $(\partial_s \boldsymbol{x}_k^T \;\; \partial_s \beta_k)^T$ *by solving (7.7), and then normalize as in (7.9).*

2. *Get the initial guess* $(\boldsymbol{x}_{k+1}^{(0)}, \beta_{k+1}^{(0)})$ *from (7.10).*

3. *Find the equilibrium* $(\boldsymbol{x}_{k+1}, \beta_{k+1})$ *using the initial guess* $(\boldsymbol{x}_{k+1}^{(0)}, \beta_{k+1}^{(0)})$ *by iterating Newton's method (7.14), giving* $\{(\boldsymbol{x}_{k+1}^{(i)}, \beta_{k+1}^{(i)})\}_i \to (\boldsymbol{x}_{k+1}, \beta_{k+1})$.

REMARK 132. *We have used an algorithm, which is a simple merger of the parameter and pseudoarclength continuation methods, which we call* tangent continuation. *Using tangent continuation, the tangent vector and the initial guess are found by steps 1-2 as in pseudoarclength continuation (Algorithm 131), and* $\boldsymbol{x}_k$ *is found as in step 3 in parameter continuation (Algorithm 130).*

## Branch Switching

Suppose that a symmetry breaking bifurcation has been located at the equilibria $(\boldsymbol{x}^*, \beta^*)$ of (3.15)

$$\dot{\boldsymbol{x}} = \psi(\boldsymbol{x}, \beta)$$

such that the assumptions of the Equivariant Branching Lemma (Theorem 47) are satisfied. To proceed, one can use the explicit form of the bifurcating direction, $\boldsymbol{u}$, to search for a bifurcating solution of interest, say $(\boldsymbol{x}_{k+1}, \beta_{k+1})$, whose existence is guaranteed by Theorem 47. As an initial guess for $\boldsymbol{x}_{k+1}$, we implement a branch switch

$$\boldsymbol{x}_{k+1}^{(0)} = \boldsymbol{x}^* + d\boldsymbol{u}. \tag{7.15}$$

Now, either Parameter, Tangent (Remark 132), or Pseudoarclength continuation can be used to create a sequence $\{(\boldsymbol{x}_{k+1}^{(i)}, \beta_{k+1}^{(i)})\}$ which converges to $(\boldsymbol{x}_{k+1}, \beta_{k+1})$.

## Continuation of the Gradient Flow

We now show how to apply Algorithms 130 and 131 to the gradient flow (3.18)

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda} \mathcal{L}(q, \lambda, \beta)$$

when $\Delta_{q,\lambda}\mathcal{L}$ is nonsingular. We use Algorithm 131 to provide the numerical results at the end of this chapter. To determine the tangent vector in step 1 of either algorithm, one needs to solve a matrix equation of the form

$$\Delta_{q,\lambda}\mathcal{L}(q_k, \lambda_k, \beta_k) \begin{pmatrix} \partial_s q(s_k) \\ \partial_s \lambda(s_k) \end{pmatrix} = -\partial_\beta \nabla_{q,\lambda}\mathcal{L}(q_k, \lambda_k, \beta_k).$$

Thus, by (3.4) and (3.6), the normalized tangent vector, $(\partial_s q_k^T \;\; \partial_s \lambda_k^T \;\; \partial_s \beta_k)^T$, to the curve $\nabla_{q,\lambda}\mathcal{L} = \mathbf{0}$ at $(q_k^T \;\; \lambda_k^T)^T$ which preserves orientation is found by solving

$$\Delta_{q,\lambda}\mathcal{L}(q_k, \lambda_k, \beta_k) \begin{pmatrix} \partial_s q(s_k) \\ \partial_s \lambda(s_k) \end{pmatrix} = - \begin{pmatrix} \nabla D(q_k) \\ \mathbf{0} \end{pmatrix} \tag{7.16}$$

and then normalizing as in (7.9). This shows that

$$\begin{pmatrix} \partial_s q_k \\ \partial_s \lambda_k \\ \partial_s \beta_k \end{pmatrix} = \frac{\operatorname{sgn}(\cos\theta)}{\sqrt{||\partial_s q_k||^2 + ||\partial_s \lambda_k||^2 + 1}} \begin{pmatrix} \partial_s q(s_k) \\ \partial_s \lambda(s_k) \\ 1 \end{pmatrix} \tag{7.17}$$

where

$$\theta = \angle \left( \begin{pmatrix} \partial_s q(s_k) \\ \partial_s \lambda(s_k) \\ 1 \end{pmatrix}, \begin{pmatrix} \partial_s \boldsymbol{x}(s_{k-1}) \\ \partial_s \lambda(s_{k-1}) \\ 1 \end{pmatrix} \right) \tag{7.18}$$

as in (7.8).

Remark 133. *Equation (3.8) shows that (7.16) can be written as*

$$\begin{pmatrix} \Delta F(q,\beta) & J^T \\ J & \mathbf{0} \end{pmatrix} \begin{pmatrix} \partial_s q(s_k) \\ \partial_s \lambda(s_k) \end{pmatrix} = - \begin{pmatrix} \nabla D(q_k) \\ \mathbf{0} \end{pmatrix}$$

*which shows that the vector $\partial_s q_k \in \ker J$, where $J$ is the Jacobian of the constraints from (3.7).*

To begin any continuation algorithm, one needs a starting point $(q_0, \lambda_0, \beta_0)$ at $k = 0$. To find this initial equilibrium , we consider the case where $q_0 = q_{\frac{1}{N}}$ and $\beta_0 = 0$, as in the case for the Information Distortion and the Information Bottleneck cost functions (2.34) and (2.35) respectively. First, we decompose

$$\nabla F(q,\beta) = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_N \end{pmatrix},$$

for some $K \times 1$ vectors $\{g_\nu\}$. By (3.5), we see that at any equilibrium $(q^*, \lambda^*, \beta)$,

$$\lambda^* = g_\nu$$

for any $\nu \in \mathcal{Y}_N$. In other words,

$$\nabla F(q^*, \beta^*) = \begin{pmatrix} g \\ g \\ \vdots \\ g \end{pmatrix},$$

and in particular,

$$\nabla F(q_{\frac{1}{N}},\beta) = \begin{pmatrix} g \\ g \\ \vdots \\ g \end{pmatrix}$$

for some $K \times 1$ vector $g$. Thus, the vector of Lagrange multipliers corresponding to $(q_{\frac{1}{N}}, 0)$ is

$$\lambda_0 = g. \tag{7.19}$$

As we will see in the numerical analysis section of this chapter, there are many saddle-node bifurcations of equilibria of (3.18). Thus, it is imperative to track equilibria by implementing pseudoarclength continuation which can navigate about such structures. We now give the Jacobian necessary to implement the Newton step (7.14) of Algorithm 131:

$$\partial_{q,\lambda,\beta} \begin{pmatrix} \nabla_{q,\lambda}\mathcal{L}_k^i \\ P_k^i - d \end{pmatrix} = \begin{pmatrix} \Delta_{q,\lambda}\mathcal{L}_k^i & \nabla D(q) \\ (\partial_s q_k^T \;\; \partial_s \lambda_k^T)^T & \partial_s \beta_k \end{pmatrix}. \tag{7.20}$$

where the notation $f_k^i$ for a function $f(q,\lambda,\beta)$ indicates $f(q_k^{(i)}, \lambda_k^{(i)}, \beta_k^{(i)})$.

As we use a continuation method to create a sequence of equilibria $\{(q_k, \lambda_k, \beta_k)\}$ along a solution branch of (3.18), it is possible that bifurcation of equilibria occurs at some $(q^*, \lambda^*, \beta^*)$ for some $\beta^* \in (\beta_k, \beta_{k+1})$ (or in $(\beta_{k+1}, \beta_k)$, if $\beta_k > \beta_{k+1}$, such as when continuing along a subcritical branch of equilibria). To determine whether a symmetry breaking bifurcation has occurred from an $M$-uniform solution, we assume that Assumption 81 holds, and rely on Corollary 89. Thus, we detect for symmetry breaking bifurcation by comparing the determinant of an unresolved block of $\Delta F(q_k, \beta_k)$ with the determinant of an unresolved block of $\Delta F(q_{k+1}, \beta_{k+1})$. This is important computationally, because we have reduced the problem of taking the determinant of the $(NK + K) \times (NK + K)$ Hessian $\Delta_{q,\lambda}\mathcal{L}$, to that of taking the determinant of a $K \times K$ block of $\Delta F$. If a bifurcation is detected, then one can use the explicit form of the bifurcating directions, $\{\boldsymbol{u}_m\}_{m=1}^M$ from (6.57) to search for the bifurcating solution of interest, say $(q_{k+1}, \lambda_{k+1}, \beta_{k+1})$, whose existence is guaranteed by Theorem 110 and Corollary 111. To do this, let $\boldsymbol{u} = \boldsymbol{u}_m$ for some $m \leq M$, then implement a branch switch as in (7.15)

$$\begin{pmatrix} q_{k+1}^{(0)} \\ \lambda_{k+1}^{(0)} \end{pmatrix} = \begin{pmatrix} q_k \\ \lambda_k \end{pmatrix} + d \cdot \boldsymbol{u}$$

for some $d > 0$. Knowledge of the explicit bifurcating directions is important computationally because, in practice, attempting to find equilibria after a bifurcation can incur significant computational cost [6, 29, 61].

In chapter 9, we apply these ideas to Algorithm 1 which gives a numerical algorithm (Algorithm 157) to find *solutions* of the problem (1.9).

| | | | |
|---|---|---|---|
| 1.038706e+000 | 1.133929e+000 | 1.390994e+000 | 4.287662e+000 |
| 5.413846e+000 | 3.112109e+001 | 4.629049e+001 | 3.827861e+002 |
| 5.961492e+002 | 7.165659e+003 | 1.010679e+004 | 1.866824e+005 |
| 2.052584e+005 | 4.683332e+006 | 6.366756e+006 | |

Table 2. Bifurcation Location: Theorem 80 is used to determine the $\beta$ values where bifurcations can occur from $(q_{\frac{1}{N}}, \beta)$ when $\Delta G(q_{\frac{1}{N}})$ is nonsingular. Using Corollary 111 and Remark 113.1 for the Information Distortion problem (2.34), we predict bifurcation from the branch $(q_{\frac{1}{4}}, \beta)$, at each of the 15 $\beta$ values given in this table.

| N | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| $\zeta(q_{\frac{1}{N}}, \beta^*, \boldsymbol{u}_k)$ | 6.04393e-4 | -5.06425e+1 | -5.40219e+2 | -2.53231e+3 | -8.10344e+3 |

Table 3. The bifurcation discriminator: Numerical evaluations of the bifurcation discriminator $\zeta(q_{\frac{1}{N}}, \beta^* \approx 1.038706, \boldsymbol{u}_k)$ (6.81) as a function of $N$ for the four blob problem (see Figure 1a) when $F$ is defined as in (2.34). We interpret that $\zeta(q_{\frac{1}{2}}, 1.038706, \boldsymbol{u}_k) = 0$. Thus, further analysis is required to determine whether the bifurcating branches guaranteed by Theorem 110 are supercritical or subcritical (numerical evidence indicates that the branches in this case are supercritical). For $N = 3, 4, 5$ and 6, we have that $\zeta(q_{\frac{1}{N}}, \beta^*, \boldsymbol{u}_k) < 0$, predicting that bifurcating branches from $q_{\frac{1}{N}}$ are subcritical and unstable in these cases (Theorem 127).

## Numerical Results

We created software in MATLAB which implemented pseudoarclength continuation (Algorithm 131) to numerically confirm the bifurcation structure guaranteed by the theory of chapter 6. All of the results presented here are for the Information Distortion problem (2.34),

$$\max_{q \in \Delta}(H(q) + \beta D_{eff}(q))$$

and for the Four Blob Problem introduced in chapter 1 and Figure 1.

When $\Delta G(q_0)$ is nonsingular, Theorem 80 determines the $\beta$ values at which singularity occurs on the branch of equilibria $(q_0, \lambda^*, \beta)$ of (3.18). In Table 2, we compute the location of singularities from the solution branch $(q_{\frac{1}{N}}, \lambda, \beta)$ of (3.18). Since $G = H(Y_N|Y)$ is strictly concave, then Corollary 111 and Remark 113.1 predict symmetry breaking bifurcation from $(q_{\frac{1}{N}}, \beta^*)$ for every $\beta^*$ value in Table 2.

Theorem 127 shows that the bifurcation discriminator, $\zeta(q^*, \beta^*, \boldsymbol{u}_k)$, can determine whether the bifurcating branches guaranteed by Theorem 110 are subcritical

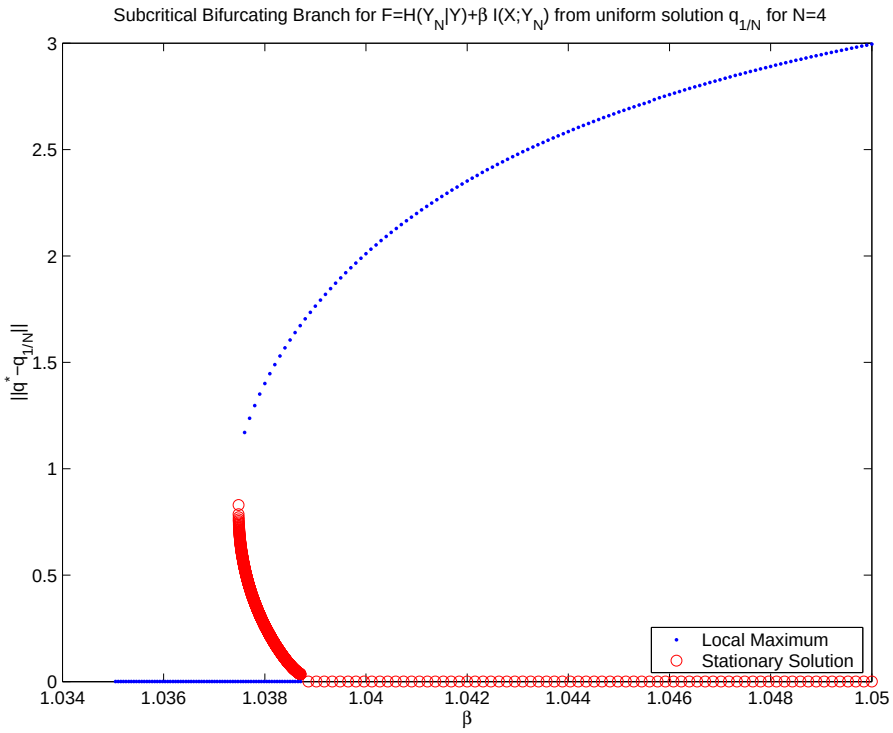



Figure 16. [54] The subcritical bifurcation from the 4-uniform solution $(q_{\frac{1}{4}}, \beta^* \approx 1.038706)$ to a 3-uniform solution branch as predicted by the fact that $\zeta(q_{\frac{1}{4}}, 1.038706, \boldsymbol{u}_k) < 0$. Here, the bifurcation diagram is shown with respect to $||q^* - q_{\frac{1}{N}}||$. It is at the saddle node that this 3-uniform branch changes from being a stationary point to a local solution of the problem (2.34).

($\zeta < 0$) or supercritical ($\zeta > 0$). The numerical results obtained by calculating $\zeta(q_{\frac{1}{N}}, \beta^*, \boldsymbol{u}_k)$ for $N = 2, 3, 4, 5$ and 6 at $\beta^* \approx 1.038706$ are shown in Table 3. The subcritical bifurcation predicted by the discriminator for the Information Distortion problem (2.34) at $\beta^* \approx 1.038706$ is shown in Figures 16 and 17.

The Figures 16–24 show numerical confirmation of symmetry breaking bifurcation from $S_M$ to $S_{M-1}$ for $N = 4$ and $M \in \{1, 2, 3, 4\}$, as guaranteed by Theorem 110 and Corollary 111. We have used both the mutual information $I(X; Y_N)$ and the norm $\|q^* - q_{\frac{1}{N}}\|$ as the vertical axis in the bifurcation diagrams. Figure 20 shows a comparison of the observed bifurcation structure given in Figure 3 (triangles), and the actual bifurcation structure given in Figures 18 and 19 (dots). Observe the shift in $\beta$, which we explain in Remark 152.

Figure 25 is numerical confirmation of symmetry breaking bifurcation from $S_N$ to the subgroups $\langle \gamma^p \rangle < S_N$ when $N = 4$ and $\gamma$ is an element of order $N$ in $S_N$, as guaranteed by Theorem **??**.

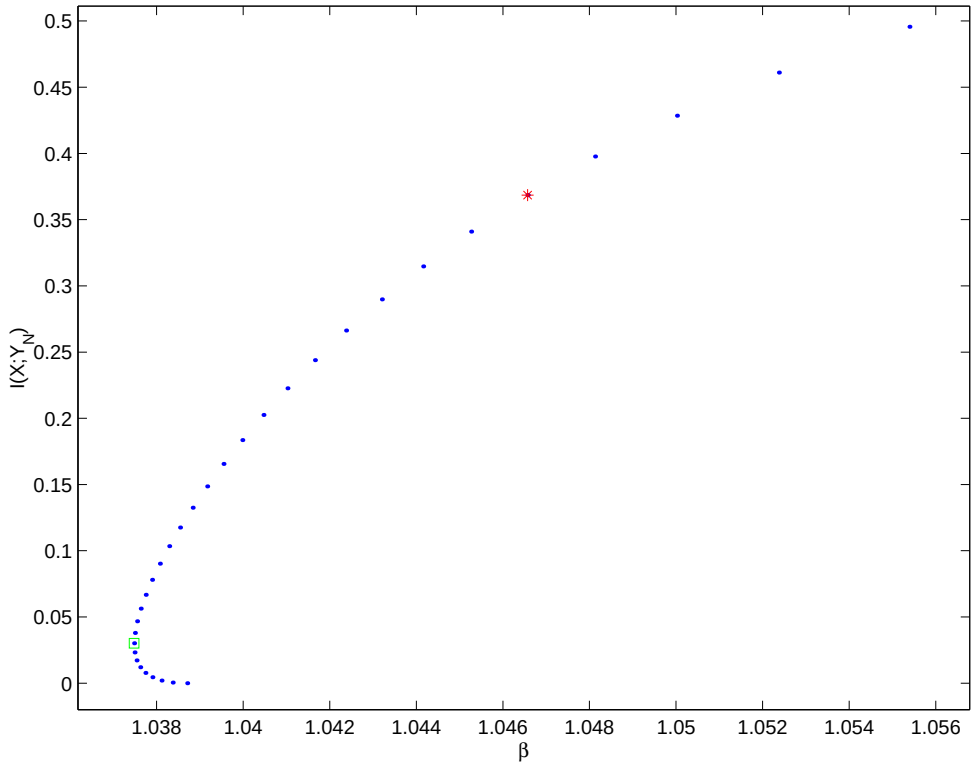



Figure 17. At symmetry breaking bifurcation from $(q_{\frac{1}{4}}, \beta^* \approx 1.038706)$, $\dim\ker \Delta F(q_{\frac{1}{N}}) = 4$ and $\dim\ker \Delta\mathcal{L}(q_{\frac{1}{N}}) = 3$ as predicted by Theorem 85. Along the subcritical branch, shown here with respect to the mutual information $I(X, Y_N)$, one eigenvalue of $\Delta F(q^*)$ is positive. The (first) block of $\Delta F(q^*)$, which by necessity also has a positive eigenvalue, is the resolved block of $\Delta F(q^*)$. Observe the saddle-node at $\beta \approx 1.037485$, where $\Delta\mathcal{L}(q^*)$ is singular, but where $\Delta F(q^*)$ is nonsingular. Later on, however, (at the asterisk) the single positive eigenvalue of $\Delta F(q^*)$ crosses again, which does not correspond to a singularity of $\Delta\mathcal{L}(q^*)$.

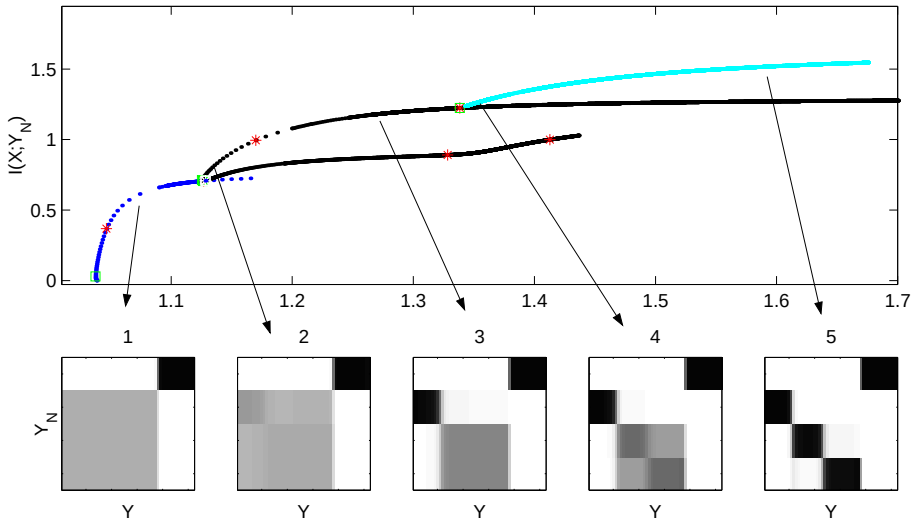



Figure 18. Actual bifurcation structure of $M$-uniform solutions for (2.34) when $N = 4$. Figure 3 showed an incomplete bifurcation structure for this same scenario. Observe that Figure 17 is a closeup of the subcritical branch which bifurcates from $(q^*, \lambda^*, 1.038706)$. Symmetry breaking bifurcation from the 4-uniform branch $(q_{\frac{1}{N}}, \lambda, 1.038706)$, to the 3-uniform branch whose quantizer is shown in panel (1), to the 2-uniform branch whose quantizer is shown in panels (2) and (3), and finally, to the 1-uniform solution branch whose quantizer is shown in panels (4) and (5).

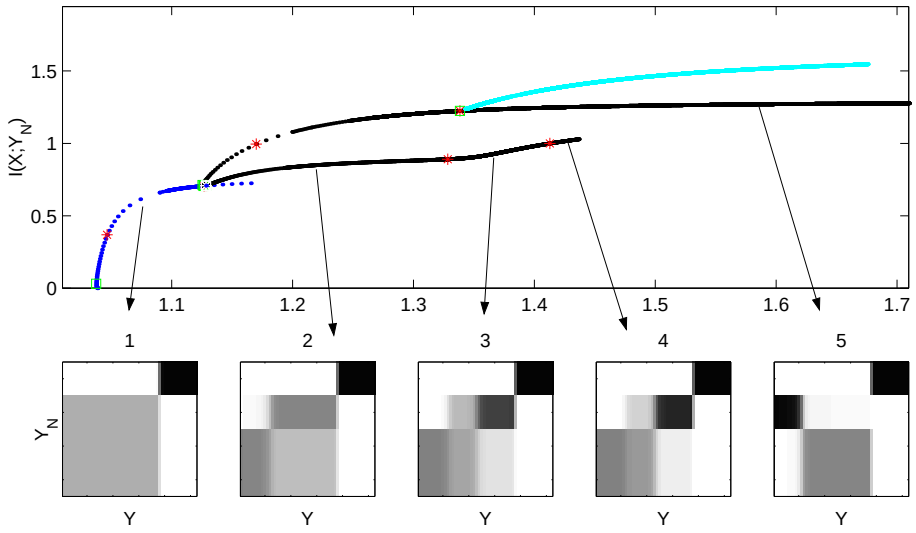



Figure 19. Symmetry breaking bifurcation from the 4-uniform branch $(q_{\frac{1}{N}}, \lambda, 1.038706)$, as in Figure 18, but now we investigate the bottom 2-uniform branch, panels (2)-(5).

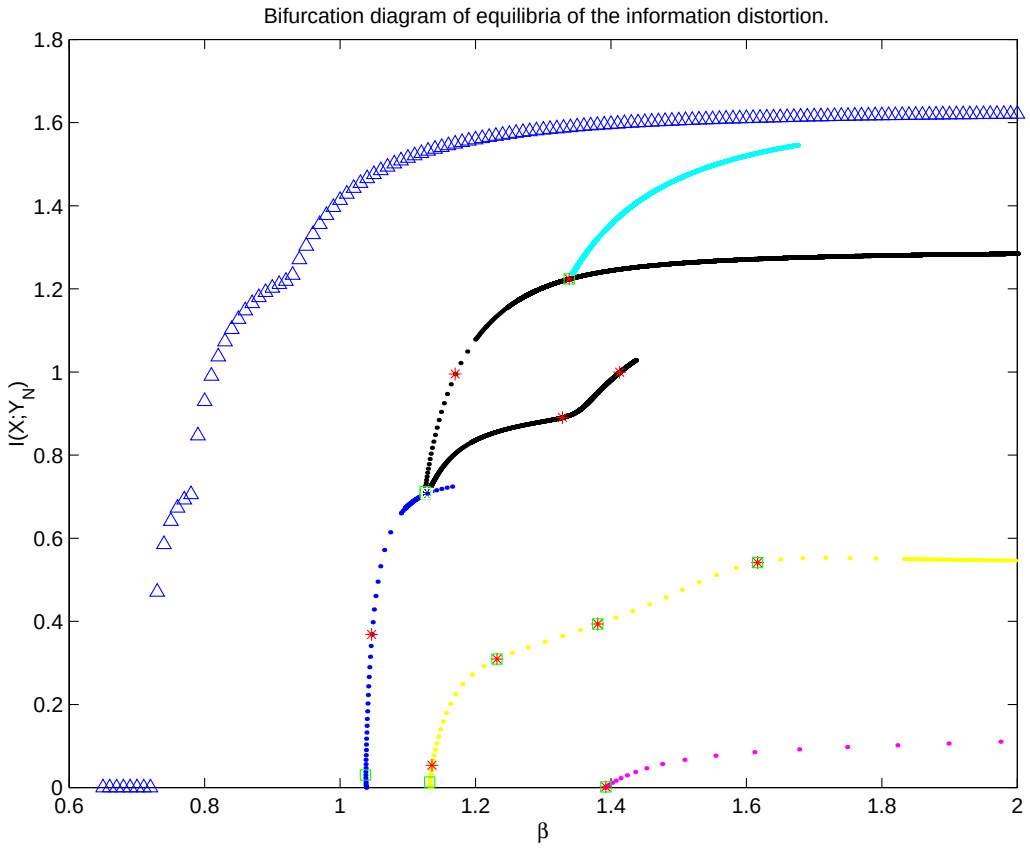



Figure 20. Comparison of the observed bifurcation structure from the 4-uniform branch given in Figure 3 (triangles), and the actual bifurcation structure given in Figures 18 and 19 (dots) when $N = 4$ for the Four Blob problem. Qualitatively, the bifurcation structure is the same, except for the shift in $\beta$, which we explain in Remark 152.

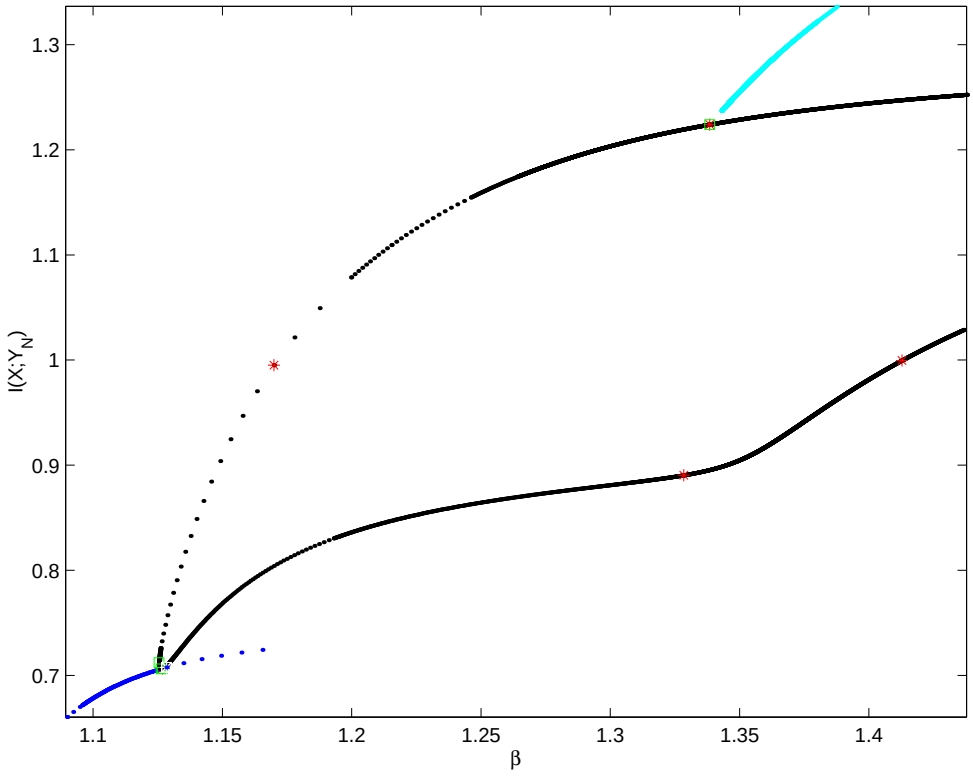



Figure 21. A close up, from Figure 18, of the 2-uniform branch which connects the 3 uniform branch below to the 1-uniform solution above. The bifurcating branch from symmetry breaking bifurcation of the 3 uniform solution is subcritical (see Figure 22), and an eigenvalue of $\Delta F(q^*)$ becomes positive. As we saw in Figure 17, this positive eigenvalue of $\Delta F(q^*)$ crosses back at the asterisk shown, which does not correspond to a singularity of $\Delta \mathcal{L}(q^*)$.

In each of the Figures 16–24, a "*" indicates a singularity point of $\Delta F(q^*)$, and a square indicates a singularity point of $\Delta \mathcal{L}(q^*)$. These pictures show that there are points where both $\Delta \mathcal{L}(q^*)$ and $\Delta F(q^*)$ are singular (at symmetry breaking bifurcations), points where just $\Delta F(q^*)$ is singular (explained by Theorem 114), and points where just $\Delta \mathcal{L}(q^*)$ is singular (at the saddle-node bifurcations). These three types of singularities are depicted in Figure 12.

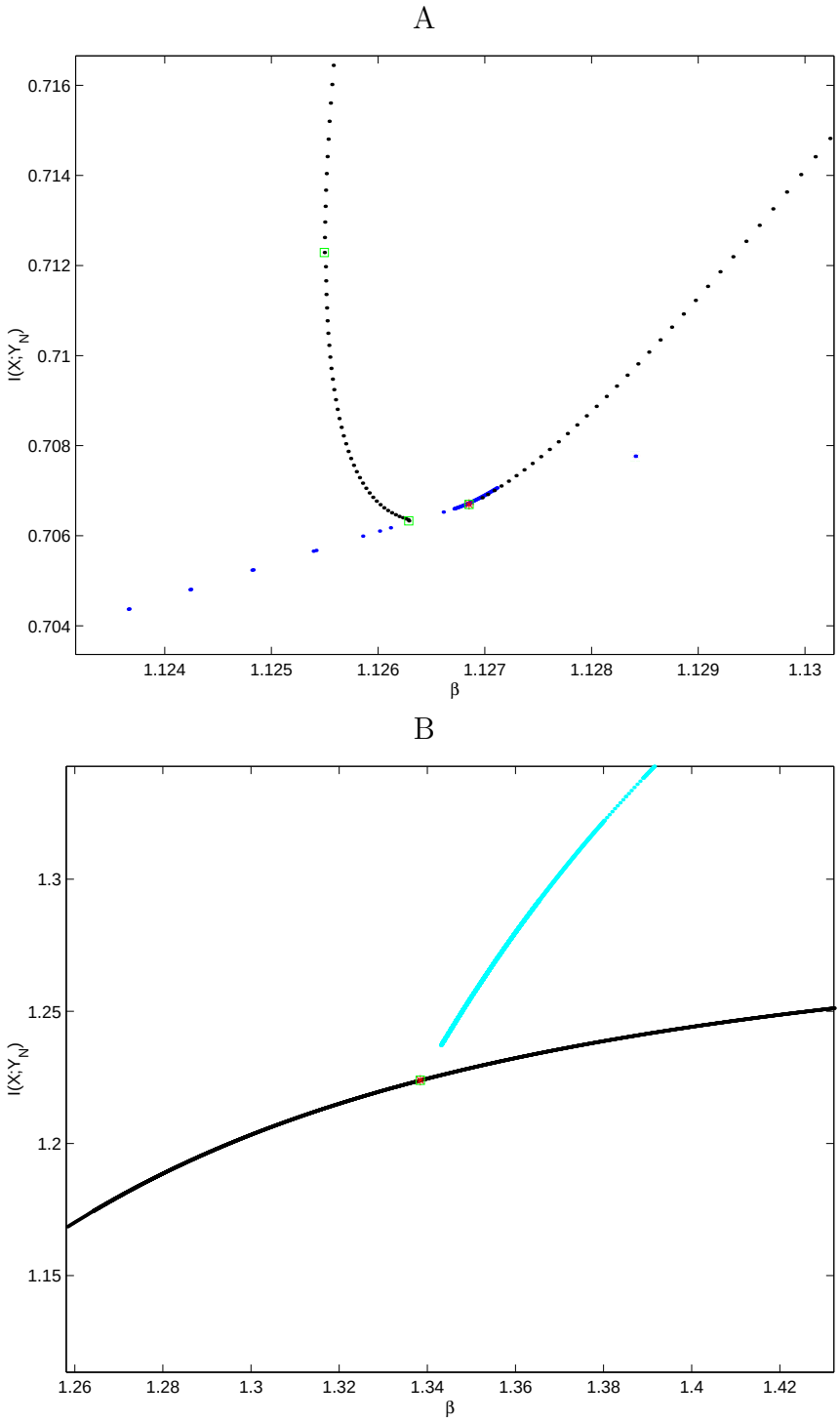



Figure 22. Panel (A) shows a close up, from Figure 18, of the subcritical bifurcation from the 3-uniform branch to the 2-uniform branch. Observe that at the saddle node, which occurs at $\beta \approx 1.1254$, only $\Delta\mathcal{L}(q^*)$ is singular. In panel (B), we show a close up, from Figure 18, where the 1-uniform branch bifurcates from symmetry breaking bifurcation of the 2-uniform solution. It is not clear whether this branch is subcritical or supercritical.

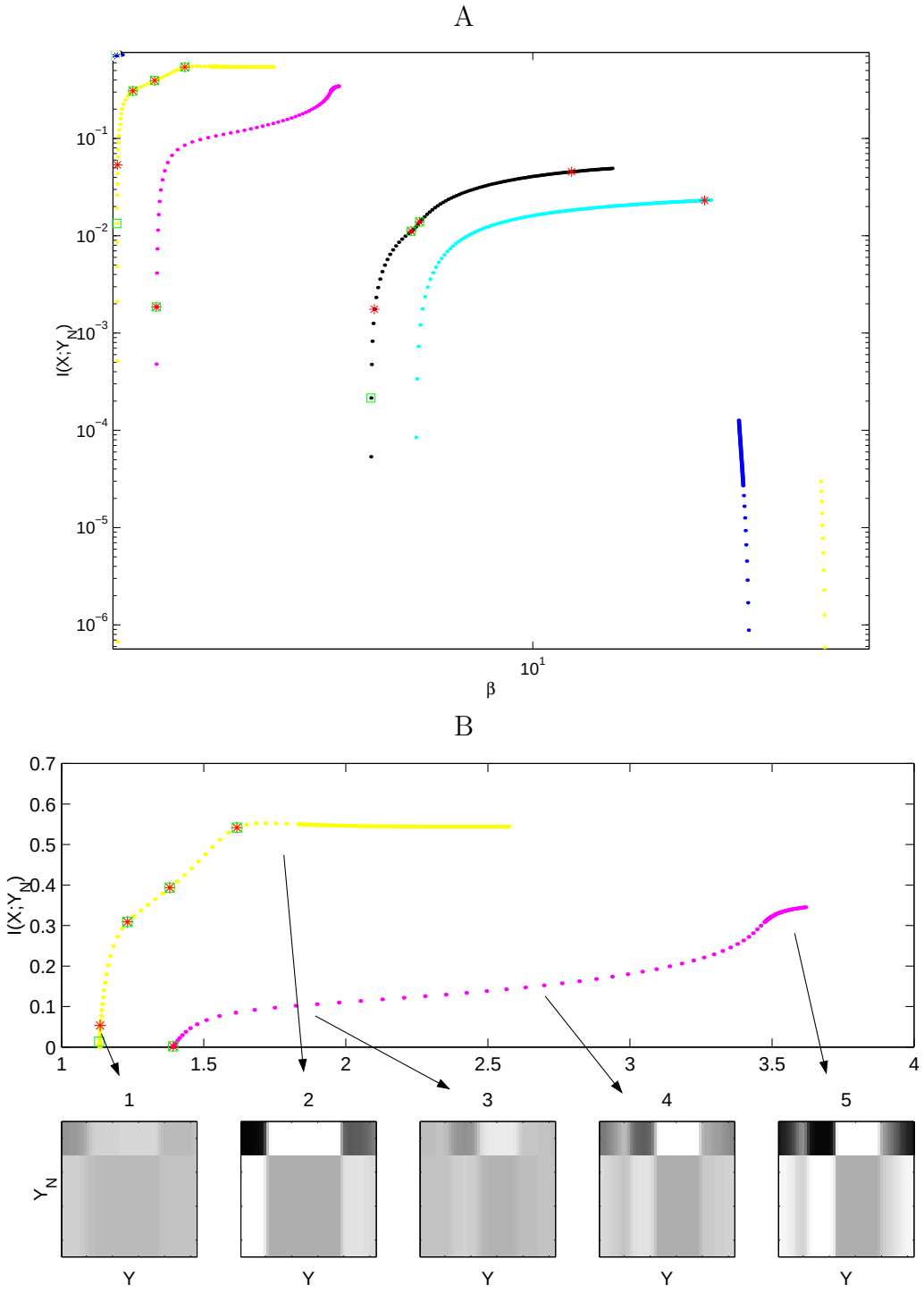



Figure 23. Panel (A) is a log-log plot of 3-uniform branches, some of which are shown in Figure 20, which bifurcate from the $q_{\frac{1}{N}}$ branch at the $\beta$ values $\{1.133929, 1.390994, 4.287662, 5.413846, 31.12109, 46.29049\}$ shown in Table 2. Panel (B) shows some of the particular quantizers along the 3-uniform branches which bifurcate from $(q_{\frac{1}{N}}, 1.133929)$ and $(q_{\frac{1}{N}}, 1.390994)$.

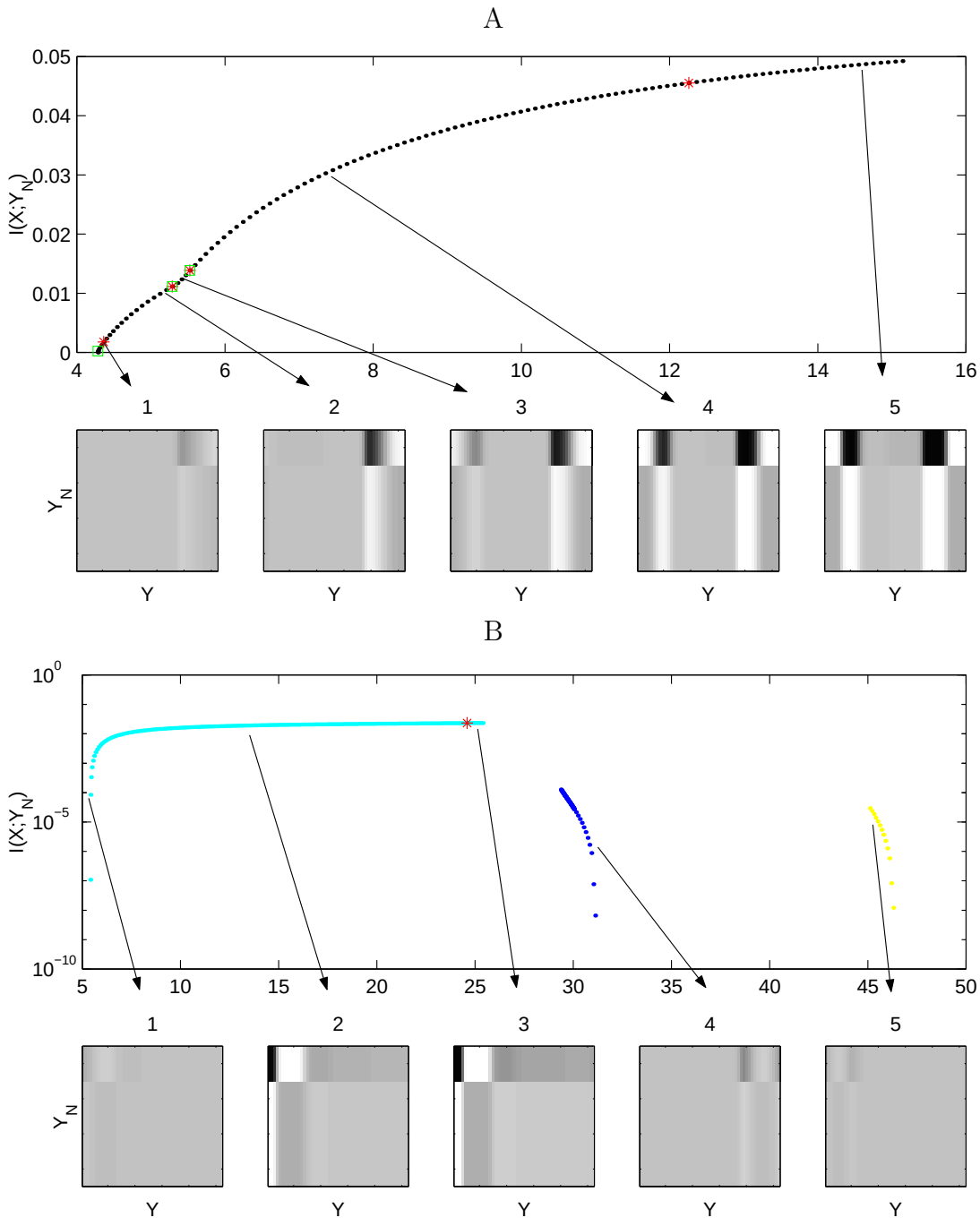



Figure 24. In panel (A) we show a 3-uniform branch, from Figure 23, which bifurcates from $(q_{\frac{1}{N}}, 4.28766)$ and some of the particular quantizers. Panel (B) shows the 3-uniform solutions, from Figure 23, which bifurcate from $q_{\frac{1}{N}}$ when $\beta \in \{5.413846, 31.12109, 46.29049\}$, and some of the associated quantizers as well.

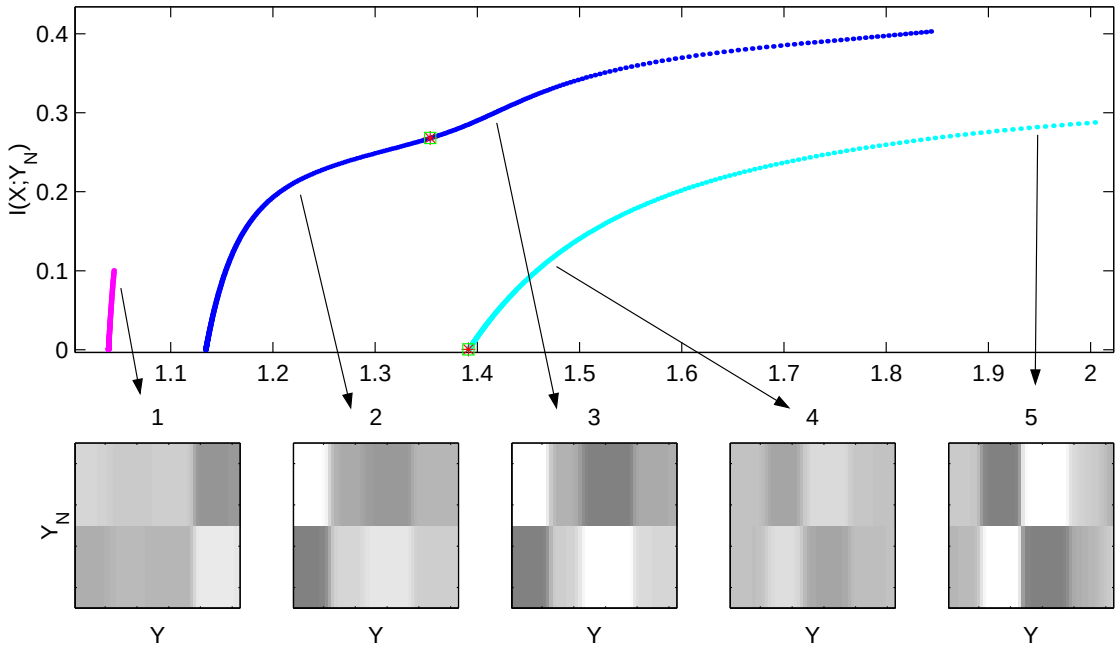



Figure 25. Bifurcating branches from the 4-uniform solution branch at the values $\beta \in \{1.038706, 1.133929, 1.390994\}$ in addition to those explained by Theorem 110. when $N = 4$. The isotropy group for all of the solution branches shown is $\langle \gamma_{(12)}, \gamma_{(34)} \rangle$ which is isomorphic to $S_2 \times S_2$. This group fixes the quantizers which are "twice" 2-uniform: 2-uniform on the classes $\mathcal{U}_1 = \{1, 2\}$, and 2-uniform on the classes $\mathcal{U}_2 = \{3, 4\}$.

# CHAPTER 8

## SADDLE-NODE BIFURCATION

This chapter examines bifurcations, which are not symmetry breaking bifurcations, in the bifurcation structure of equilibria of (3.18),

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda}\mathcal{L}(q,\lambda,\beta).$$

We show that generically, these types of bifurcations are saddle-node bifurcations, which we confirmed numerically in chapter 7.

We will call bifurcations which are not symmetry breaking bifurcations *non-symmetry breaking bifurcations*. We derive an explicit basis of $\ker \Delta_{q,\lambda}\mathcal{L}$ at non-symmetry breaking bifurcations. We also show necessary and sufficient conditions for the existence of a saddle-node bifurcation.

Suppose that a bifurcation of equilibria of (3.18) occurs at $(q^*,\lambda^*,\beta^*)$, with a bifurcating branch $\left(\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + \boldsymbol{u}(t), \beta^* + \beta(t)\right)$. Furthermore, let $n(\beta)$ be the number of equilibria of (3.18). We use the following definition for a saddle-node bifurcation.

DEFINITION 134. *A bifurcation at* $(q^*,\lambda^*,\beta^*)$ *is a* saddle-node *bifurcation if* $\beta'(0) = 0$, $n(\beta^*) = 1$, *and if either*

$$n(\beta) = \begin{cases} 0 & \text{for } \beta < \beta^* \\ 2 & \text{for } \beta > \beta^* \end{cases}$$

*or*

$$n(\beta) = \begin{cases} 0 & \text{for } \beta > \beta^* \\ 2 & \text{for } \beta < \beta^* \end{cases}.$$

Let the $K \times K$ matrices $B$ and $\{R_\nu\}_{\nu\in\mathcal{R}}$ be defined as in (6.3) and (6.4). We assume that generically, only one of the matrices $B$, $\{R_\nu\}_{\nu\in\mathcal{R}}$, or $B\sum_\nu R_\nu^{-1} + MI_K$ is singular at a given point $(q,\beta) \in \Delta \times \Re$ (see Definition 40 and Remark 41).

### Kernel of the Hessian at Non-symmetry Breaking Bifurcation

The Hessian $\Delta_{q,\lambda}\mathcal{L}$ plays a pivotal role in determining the bifurcation structure of $M$-uniform equilibria $(q^*,\lambda^*,\beta)$ of (3.18) since bifurcation at $\beta = \beta^*$ happens when $\ker \Delta_{q,\lambda}\mathcal{L}(q^*,\lambda^*,\beta^*)$ is nontrivial (Theorem 24). Furthermore, as we have seen in chapter 6 for symmetry breaking bifurcation, the bifurcating branches are tangent to certain linear subspaces of $\ker \Delta_{q,\lambda}\mathcal{L}(q^*,\lambda^*,\beta^*)$ (Theorem 110). Theorems 36 and 114,

and Corollary 89, show that the Hessian $\Delta F$ plays a part in predicting bifurcation as well (see Figure 12). In this section, we examine the singularities of $\Delta_{q,\lambda}\mathcal{L}$ and $\Delta F$ which give rise to non-symmetry breaking bifurcations, which we observed as saddle-node bifurcations in chapter 7.

We begin by deriving an explicit basis for $\ker \Delta\mathcal{L}(q^*)$ when $\Delta F(q^*)$ is nonsingular. Theorem 139 shows that, under the genericity assumption given in Remark 41, this is the basis for $\ker \Delta\mathcal{L}(q^*)$ at a non-symmetry breaking bifurcation. The next theorem was presented in chapter 4 as Theorem 39.

THEOREM 135. *Suppose that $\Delta F(q^*)$ is nonsingular. Then $\Delta_{q,\lambda}\mathcal{L}$ is singular if and only if $B\sum_\nu R_\nu^{-1} + MI_K$ is singular.*

The next two lemmas prove this theorem. Lemma 136 gives a basis of $\ker \Delta_{q,\lambda}\mathcal{L}$ with respect to the matrix $\sum_\nu R_\nu^{-1}B + MI_K$. Lemma 137 relates this result with the matrix $B\sum_\nu R_\nu^{-1} + MI_K$

LEMMA 136. *Suppose that $\Delta F(q^*)$ is nonsingular. Then $\Delta_{q,\lambda}\mathcal{L}$ is singular if and only if $\sum_\nu R_\nu^{-1}B + MI_K$ is singular. Furthermore, $\boldsymbol{v}$ is in the kernel of $\sum_\nu R_\nu^{-1}B + MI_K$ if and only if $\boldsymbol{k}$ is in the kernel of $\Delta\mathcal{L}(q^*)$ where*

$$\boldsymbol{k} = \begin{pmatrix} \hat{\boldsymbol{k}} \\ -B\boldsymbol{v} \end{pmatrix} \tag{8.1}$$

*and*

$$[\hat{\boldsymbol{k}}]_\eta \;=\; \begin{cases} R_\nu^{-1}B\boldsymbol{v} & \text{if } \eta \text{ is the } \nu^{th} \text{ resolved class of } \mathcal{R} \\ \boldsymbol{v} & \text{otherwise (i.e. if } \eta \in \mathcal{U}) \end{cases} . \tag{8.2}$$

*Proof.* We first prove sufficiency. Let $\boldsymbol{v} \in \ker(\sum_\nu R_\nu^{-1}B + MI_K)$. Constructing a vector $\boldsymbol{k}$ as in (8.1) and (8.2), and rewriting $\Delta\mathcal{L}(q^*)$ as in (3.8), we see that

$$\Delta\mathcal{L}(q^*)\boldsymbol{k} \;=\; \begin{pmatrix} \Delta F & J^T \\ J & \mathbf{0} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{k}} \\ -B\boldsymbol{v} \end{pmatrix} = \begin{pmatrix} \Delta F(q^*)\hat{\boldsymbol{k}} - J^TB\boldsymbol{v} \\ J\hat{\boldsymbol{k}} \end{pmatrix}, \tag{8.3}$$

which is the left hand side of (4.2). Multiplying out (8.3) (see (4.7) and (4.8)), we see that

$$\begin{pmatrix} \Delta F(q^*)\hat{\boldsymbol{k}} - J^TB\boldsymbol{v} \\ J\hat{\boldsymbol{k}} \end{pmatrix} = \begin{pmatrix} B\boldsymbol{v} \\ B\boldsymbol{v} \\ \vdots \\ B\boldsymbol{v} \\ (\sum_\nu R_\nu^{-1}B + MI_K)\boldsymbol{v} \end{pmatrix} - \begin{pmatrix} B\boldsymbol{v} \\ B\boldsymbol{v} \\ \vdots \\ B\boldsymbol{v} \\ \mathbf{0} \end{pmatrix} = \mathbf{0}.$$

To prove necessity, let $\boldsymbol{k} \in \ker \Delta\mathcal{L}(q^*)$ and decompose it as in (4.1) and (4.6). Then by (4.7) we have

$$\begin{pmatrix} B_1\boldsymbol{x}_1 \\ B_2\boldsymbol{x}_2 \\ \vdots \\ B_N\boldsymbol{x}_N \end{pmatrix} = - \begin{pmatrix} \boldsymbol{k}_J \\ \boldsymbol{k}_J \\ \vdots \\ \boldsymbol{k}_J \end{pmatrix}.$$

This equation implies

$$\begin{aligned} B\boldsymbol{x}_\eta &= -\boldsymbol{k}_J \text{ for } \eta \in \mathcal{U} \\ R_\nu\boldsymbol{x}_\nu &= -\boldsymbol{k}_J \text{ for } \nu \in \mathcal{R} \quad . \end{aligned}$$

Since $\Delta F(q^*)$ is nonsingular, then

$$\boldsymbol{x}_\eta = \boldsymbol{x} = -B^{-1}\boldsymbol{k}_J \tag{8.4}$$

for every $\eta \in \mathcal{U}$, from which it follows that

$$\boldsymbol{x}_\nu = R_\nu^{-1}B\boldsymbol{x} \tag{8.5}$$

for every $\nu \in \mathcal{R}$. This shows that if $\boldsymbol{k} \in \ker \Delta\mathcal{L}(q^*)$, then it has the form specified by (8.1) and (8.2) for some vector $\boldsymbol{x} \in \Re^K$. To show that $\boldsymbol{x} \in \ker(\sum_i R_i^{-1}B + MI_K)$, we use the relationship (4.8),

$$\sum_{\mu \in \mathcal{Y}_N} \boldsymbol{x}_\mu = \boldsymbol{0},$$

which implies that

$$\sum_{\nu \in \mathcal{R}} \boldsymbol{x}_\nu + \sum_{\eta \in \mathcal{U}} \boldsymbol{x}_\eta = \boldsymbol{0},$$

and so (8.4) and (8.5) give

$$\sum_{\nu \in \mathcal{R}} R_\nu^{-1}B\boldsymbol{x} + \sum_{\eta \in \mathcal{U}} \boldsymbol{x} = \sum_{\nu \in \mathcal{R}} R_\nu^{-1}B\boldsymbol{x} + M\boldsymbol{x} = \boldsymbol{0}$$

which shows that $\boldsymbol{x}$ is in the kernel of $\sum_i R_i^{-1}B + MI_K$. □

The previous lemma explicitly considered the matrix $\sum_i R_i^{-1}B + MI_K$. To rephrase the result of Lemma 136 in terms of $B\sum_i R_i^{-1} + MI_K$, we prove the following lemma.

LEMMA 137. *Suppose that $\Delta F(q^*)$ is nonsingular. Then $\sum_i R_i^{-1}B + MI_K$ is singular with a single basis vector $\boldsymbol{v}$ if and only if $B\sum_i R_i^{-1} + MI_K$ is singular with a single basis vector $\boldsymbol{x} = B\boldsymbol{v}$.*

*Proof.* Let $\boldsymbol{v}$ be the basis vector for $\ker(\sum_i R_i^{-1}B + MI_K)$. Then

$$\begin{aligned} & (\sum_i R_i^{-1}B + MI_K)\boldsymbol{v} = \mathbf{0} \\ \Leftrightarrow \quad & (\sum_i R_i^{-1} + MB^{-1})B\boldsymbol{v} = \mathbf{0} \\ \Leftrightarrow \quad & (B\sum_i R_i^{-1} + MI_K)B\boldsymbol{v} = \mathbf{0}. \end{aligned}$$

To show that $B\boldsymbol{v}$ is the basis vector of $B\sum_i R_i^{-1}+MI_K$, consider some $\boldsymbol{x} \in \ker(B\sum_i R_i^{-1}+MI_K)$. Then

$$\begin{aligned} & (B\sum_i R_i^{-1} + MI_K)\boldsymbol{x} = \mathbf{0} \\ \Leftrightarrow \quad & B(\sum_i R_i^{-1} + MB^{-1})\boldsymbol{x} = \mathbf{0} \\ \Leftrightarrow \quad & (\sum_i R_i^{-1}B + M)B^{-1}\boldsymbol{x} = \mathbf{0}. \end{aligned}$$

Thus, $B^{-1}\boldsymbol{x} \in \ker(\sum_i R_i^{-1}B + MI_K)$. Since $\boldsymbol{v}$ is the basis vector of $\ker(\sum_i R_i^{-1}B + MI_K)$, then $\boldsymbol{v} = cB^{-1}\boldsymbol{x}$ for some $c \in \Re$, which shows that $B\boldsymbol{v}$ is a basis vector for $\ker(B\sum_i R_i^{-1} + MI_K)$. □

## Necessary Conditions

We are ready to prove some necessary conditions which must be satisfied generically at a non-symmetry breaking bifurcation of an $M$-uniform solution, which includes saddle-node bifurcations. The next theorem shows that $\Delta F(q^*)$ is generically nonsingular at a bifurcation which is not symmetry breaking.

THEOREM 138. *At a non-symmetry breaking bifurcation of an $M$-uniform solution $(q^*, \lambda^*, \beta^*)$, $\Delta F(q^*)$ is generically nonsingular.*

*Proof.* If $\Delta F(q^*)$ is singular, then, generically, either $R_\nu$ is singular for some resolved block of $\Delta F(q^*)$, or the unresolved block $B$ of $\Delta F(q^*)$ is singular. If the former holds, then generically, $B\sum_i R_i^{-1} + MI_K$ is nonsingular, and now Theorem 114 shows that $\Delta\mathcal{L}(q^*)$ is nonsingular, which is impossible since we assume that we are at a bifurcation. If $B$ is singular, then generically, $B\sum_i R_i^{-1} + MI_K$ is nonsingular, which we showed in chapter 6 leads to symmetry breaking bifurcation (Theorem 110 and Corollary 111). Thus, we must have that $\Delta F(q^*)$ is nonsingular. □

The next theorem shows that, generically, the kernel of $\Delta\mathcal{L}(q^*)$ at a non-symmetry breaking bifurcation has dimension 1. Thus, we are able to give an explicit bifurcating direction.

THEOREM 139. *At a generic non-symmetry breaking bifurcation* $(q^*, \lambda^*, \beta^*)$ *of an* $M$*-uniform solution,* $\dim\ker\Delta\mathcal{L}(q^*) = 1$ *and the bifurcating direction* $\boldsymbol{u}$ *is given by*

$$\boldsymbol{u} \;=\; \begin{pmatrix} \hat{\boldsymbol{u}} \\ -B\boldsymbol{v} \end{pmatrix},$$

*where*

$$[\hat{\boldsymbol{u}}]_\eta \;=\; \begin{cases} R_\nu^{-1}B\boldsymbol{v} & \text{if } \eta \text{ is the } \nu^{th} \text{ resolved class of } \mathcal{R} \\ \boldsymbol{v} & \text{otherwise (i.e. if } \eta \in \mathcal{U}) \end{cases},$$

*and* $\boldsymbol{v}$ *is in the kernel of* $\sum_\nu R_\nu^{-1}B + MI_K$.

*Proof.* By genericity, we can apply by Theorem 138, and Lemmas 136 and 137, showing that $\dim\ker\Delta\mathcal{L}(q^*) = 1$. Since bifurcating directions are in $\ker\Delta\mathcal{L}(q^*)$ (see (5.35)), then the basis vector given in Lemma 136 must be the bifurcating direction. □

At a non-symmetry breaking bifurcation, the whole kernel of $\Delta\mathcal{L}(q^*)$ is fixed by the isotropy group of $(q^*, \lambda^*, \beta^*)$.

THEOREM 140. *At a generic non-symmetry breaking bifurcation* $(q^*, \lambda^*, \beta^*)$ *of an* $M$*-uniform solution,* $\mathrm{Fix}(\Gamma_\mathcal{U}) \cap \ker\Delta\mathcal{L}(q^*) = \ker\Delta\mathcal{L}(q^*)$.

*Proof.* By genericity, we can apply Theorem 138 and Lemma 136 to get the explicit form of $\boldsymbol{k}$, the basis vector of $\Delta\mathcal{L}(q^*)$, from (8.1) and (8.2). The desired result now follows by Theorem 71 and the definition of the group $\Gamma_\mathcal{U}$ from (6.8). □

## A Sufficient Condition

In this section we provide a sufficient condition for the existence of saddle-node bifurcations. Observe that the first assumption given in the following theorem is satisfied generically at any non-symmetry breaking bifurcation (Theorem 139), and that the second assumption is a crossing condition.

THEOREM 141. *Suppose that* $(q^*, \lambda^*, \beta^*)$ *is a bifurcation point of (3.18) such that:*

1. *The dimension of* $\ker\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*)$ *is* 1 *with basis vector* $k$.
2. *The dot product* $< \boldsymbol{k}, \begin{pmatrix} \nabla D(q^*) \\ \mathbf{0} \end{pmatrix} > \;\neq 0$.

*Then* $(q^*, \lambda^*, \beta^*)$ *is a saddle-node bifurcation.*

*Proof.* Since $\dim\ker\Delta_{q,\lambda}\mathcal{L}(q^*,\lambda^*,\beta^*)=1$, then a bifurcating branch must be of the form

$$\left(\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}, \beta^* + \beta(t)\right)$$

for $\boldsymbol{u}\in\ker\Delta_{q,\lambda}\mathcal{L}(q^*,\lambda^*,\beta^*)$. We prove the theorem by showing that $\beta'(0)=0$ and that the number of equilibria, $n(\beta)$, changes from 0 to 2 about bifurcation at $\beta=\beta^*$ (see Definition 134).

Since we have chosen $\boldsymbol{k}$ as the single basis vector of $\ker\Delta_{q,\lambda}\mathcal{L}(q^*,\lambda^*,\beta^*)$, then

$$\boldsymbol{u}=x_0\boldsymbol{k}$$

for some nonzero scalar $x_0\in\Re$. Furthermore, by definition of the Liapunov-Schmidt reduction given in (5.36), we have that

$$\begin{aligned} r(x,\beta) &:= \boldsymbol{k}^T(I-E)\mathcal{F}(\boldsymbol{k}x+U(\boldsymbol{k}x,\beta),\beta) \\ r &: \Re\times\Re\to\Re, \end{aligned} \tag{8.6}$$

where $\mathcal{F}(q,\lambda,\beta)=\nabla_{q,\lambda}\mathcal{L}(q+q^*,\lambda+\lambda^*,\beta+\beta^*)$ and $q=\boldsymbol{k}x+U(\boldsymbol{k}x,\beta)$. Thus,

$$r(tx_0,\beta)=h(t,\beta)x_0 \tag{8.7}$$

for some scalar function $h(t,\beta)$. From (8.6) we have that $r(0,0)=0$, and now (8.7) implies that

$$h(0,0)=0. \tag{8.8}$$

From (8.7) we see that $\partial_\beta r(tx_0,\beta)=\partial_\beta h(t,\beta)x_0$ from which it follows that

$$\partial_\beta r(0,0)=\partial_\beta h(0,0)x_0.$$

To show that $\partial_\beta h(0,0)\neq 0$, we appeal to equations (5.39) and (8.6), which show that

$$\begin{aligned} \partial_\beta r(0,0) &= \boldsymbol{k}^T(I-E)\partial_\beta\nabla_{q,\lambda}\mathcal{L}(q^*,\lambda^*,\beta^*) \\ &= \boldsymbol{k}^T\begin{pmatrix} \nabla D(q^*) \\ \mathbf{0} \end{pmatrix}, \end{aligned}$$

where the last equality follows from the fact that $\partial_\beta\nabla_{q,\lambda}\mathcal{L}=(\nabla D^T\ \ \mathbf{0}^T)^T$, and that $<\boldsymbol{k},(I-E)V>=<\boldsymbol{k},V>$ for any vector $V$ since $\boldsymbol{k}\perp EV$. By the assumption that $<\boldsymbol{k},\begin{pmatrix} \nabla D(q^*) \\ \mathbf{0} \end{pmatrix}>\ \neq 0$, we have that $\partial_\beta h(0,0)\neq 0$. This and (8.8) show that the Implicit Function Theorem can be applied to solve

$$h(t,\beta)=0 \tag{8.9}$$

uniquely in $\Re$ for $\beta=\beta(t)$ about $(t=0,\beta=0)$. Thus, there is only one bifurcating branch in $\ker\Delta_{q,\lambda}\mathcal{L}(q^*,\lambda^*,\beta^*)$ for small $t$

$$\left(\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}, \beta^* + \beta(t)\right).$$

Thus, $n(\beta)$ must change from 0 to 2 about bifurcation at $\beta = \beta^*$, since there is one bifurcating branch for positive $t$, and a second branch for negative $t$. The only other possibility is that $n(\beta)$ is 1 for all $\beta$ about $\beta^*$, which violates the assumption that bifurcation occurs at $\beta = \beta^*$.

To show that $\beta'(0) = 0$, we find the total derivative of (8.9), giving

$$\partial_t h(t, \beta) + \partial_\beta h(t, \beta)\beta'(t) = 0$$

from which it follows that

$$\beta'(0) = -\frac{\partial_t h(0, 0)}{\partial_\beta h(0, 0)}.$$

By (8.7) we see that

$$\partial_x r(tx_0, \beta)x_0 = \partial_t h(t, \beta)x_0, \tag{8.10}$$

and so (8.6) and the fact that $\ker(I - E) = \text{range}\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*)$ show that

$$\partial_t h(0, 0) = \partial_x r(0, 0) = \boldsymbol{k}^T(I - E)\Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*) = 0.$$

Thus $\beta'(0) = 0$. □

# CHAPTER 9

## OPTIMIZATION SCHEMES

Up until now, we have studied the structure of all of the stationary points of (1.9),

$$\max_{q\in\Delta}(G(q)+\beta D(q)).$$

by working with (3.1)

$$\max_{q\in\Delta_{\mathcal{E}}}(G(q)+\beta D(q)).$$

In this chapter, we derive three methods to find *solutions* of (1.9), which are stationary points $(q^*,\beta^*)$ of (1.9) for which $\Delta F(q^*)$ is non-positive definite on $\ker J$, the kernel of the Jacobian of the constraints (3.7) (Theorem 20).

We begin by reviewing the theory which justifies our use of numerical optimization techniques to solve (1.9). We use the Augmented Lagrangian method (Algorithm 149) with a Newton Conjugate Gradient line search (Algorithm 145). We also present an implicit solution method (9.20). Both of these methods are used in conjunction with the method of annealing, Algorithm 1. When $D(q)$ is convex and $\beta\to\infty$, the maximizer of (1.9) lies generically at a vertex of $\Delta$ (Theorem 153). Thus, we use an algorithm, called Vertex Search (Algorithm 155), to solve (1.9) in this instance, by searching over the vertices of $\Delta$. We conclude the chapter with numerical results of these applications on synthetic and physiological data sets.

### Notation

The following notation will be used throughout the chapter:

$n := NK$, the number of coordinates of $q\in\Delta$.

$F(q) := F(q,\beta)$ for a fixed $\beta$.

### Optimization Theory

The goal of numeric optimization techniques is to efficiently compute the optimizer of a given cost function subject to given constraints. In the case of solving (1.9), this means that we search for

$$\arg\max_{q\in\Delta}F(q,\mathcal{B}), \tag{9.1}$$

where $\mathcal{B} \in [0, \infty)$, and $F$ is defined as in (1.10),

$$F(q, \beta) = G(q) + \beta D(q).$$

Using Algorithm 1 to find (9.1), we see that in step 3 of the $m^{th}$ iteration, one finds

$$q_{m+1} = \arg\max_{q \in \Delta} F(q, \beta_m). \tag{9.2}$$

In other words, at the $m^{th}$ iteration, one is interested in solving (1.9) for a fixed $\beta = \beta_m$. One of the main topics of this chapter is solving (9.2) for such a fixed $\beta$. Since $\beta$ is fixed, we will write $F(q)$ instead of $F(q, \beta)$ throughout much of this chapter.

As in step 3 of Algorithm 1, we wish to find a local solution $q_m$ (for $m \geq 0$) of (1.9) at $\beta = \beta_m$. We let

$$q^* = q_m.$$

Thus, $(q^*, \beta)$ is a local solution of (1.9) (and of (3.1) - see Remark 19). Furthermore, by Theorem 16, there exists a vector of Lagrange multipliers $\lambda^*$ such that $(q^*, \lambda^*, \beta)$ is an equilibria of (3.19) (and of (3.18) - see Remark 28).

Let $\hat{\mathcal{L}}$ be the Lagrangian of (1.9)

$$\hat{\mathcal{L}}(q, \lambda, \xi, \beta) = F(q, \beta) + \sum_{k=1}^{K} \lambda_k (\sum_{\nu=1}^{N} q_{\nu k} - 1) + \sum_{k=1}^{K} \sum_{\nu=1}^{N} \xi_{\nu k} q_{\nu k}$$

(compare with (3.3) and (3.13)). The goal of constrained numerical optimization techniques is to find $q^*$ by building a sequence $\{q_k\}_{k=1}^{\infty}$ which converges to $q^*$ such that

$$F \text{ is increased for each } k : F(q_{k+1}) \geq F(q_k) \text{ for all } k. \tag{9.3}$$

$$\textit{global convergence: } ||\nabla_{q,\lambda} \hat{\mathcal{L}}(q_k)|| \to 0 \text{ as } k \to \infty. \tag{9.4}$$

$$q_k \in \Delta \text{ for each } k. \tag{9.5}$$

One way to stipulate (9.5) is to define constraint functions $\{c_i\}_{i \in \mathcal{E} \cup \mathcal{I}}$ as in Remark 17.

When no constraints are present, then *unconstrained* numerical optimization techniques are used to find the unconstrained maximizer $\hat{q}$ of $F(q, \beta)$ by building a sequence $\{q_k\}_{k=1}^{\infty}$ which converges to $q^*$ such that

$$F \text{ is increased for each } k : F(q_{k+1}) \geq F(q_k) \text{ for all } k. \tag{9.6}$$

$$\textit{global convergence: } ||\nabla F(q_k)|| \to 0 \text{ as } k \to \infty. \tag{9.7}$$

We review unconstrained numerical techniques as an introduction to the methods used in the constrained regime.

Unconstrained Line Searches

Unconstrained *line searches* can be used to find a sequence $\{q_k\}_{k=1}^{\infty}$ which converges to

$$q^* = \text{argmax}_{q \in \Re^n} F(q, \beta),$$

where each element of the sequence satisfies (9.6) and (9.7).

Algorithm 142 (Unconstrained Line Search). *Choose some $q_0$ and let $k \geq 0$. At $q_k$ compute $q_{k+1}$ as follows:*

1. *Compute an* ascent direction $p_k$ *at* $q_k$.
2. *Compute the* step length

$$\alpha_k \approx \arg\max_{\alpha > 0} F(q_k + \alpha p_k).$$

3. *Define* $q_{k+1} = q_k + \alpha_k p_k$.

Recall that $\nabla F \in \Re^n$. An ascent direction is a vector $p_k \in \Re^n$ for which

$$\nabla F(q_k)^T p_k > 0. \tag{9.8}$$

Such a $p_k$ guarantees that $F$ can be increased along $p_k$ for some step $\alpha$, since applying Taylor's Theorem about $\alpha = 0$ shows that

$$F(q_k + \alpha p_k) = F(q_k) + \alpha p_k \nabla F(q_k) + \mathcal{O}(\alpha^2)$$

which implies that

$$F(q_k + \alpha p_k) - F(q_k) > 0$$

for $\alpha$ sufficiently small. Geometrically, letting $\theta_k$ be the angle between $\nabla F(q_k)$ and $p_k$, (9.8) is equivalent to requiring that

$$\|\nabla F(q_k)\| \|p_k\| \cos\theta_k > 0,$$

which implies that $-\frac{\pi}{2} < \theta_k < \frac{\pi}{2}$. To compute the step length $\alpha_k$ given an ascent direction $p_k$, one might only require that

$$F(q_k + \alpha_k p_k) > F(q_k).$$

This naive condition is not strong enough. Rather, one must find $\alpha_k$ such that the following two conditions, called the *Wolfe Conditions*, are satisfied

$$F(q_k + \alpha_k p_k) \geq F(q_k) + c_1 \alpha_k \nabla F(q_k)^T p_k \text{ for some } c_1 \in (0, 1) \tag{9.9}$$

$$\nabla F(q_k + \alpha_k p_k)^T p_k \leq c_2 \nabla F(q_k)^T p_k \text{ for some } c_2 \in (c_1, 1). \tag{9.10}$$

Condition (9.9) requires *sufficient decrease* of $F$ and (9.10) is called the *curvature condition*. The following theorem shows that enacting a line search with $\alpha_k$ and $p_k$ which satisfy the Wolfe Conditions yields $q_k$ such that (9.6) and (9.7) are satisfied.

THEOREM 143. *(p. 45-6 [50]) Let $F$ be defined as in (3.2) with Assumptions 15. If for every $k \geq 0$ in Algorithm 142, $p_k$ is an ascent direction, and $\alpha_k$ satisfies (9.9) and (9.10), then* $\lim_{k\to\infty} \|\nabla F(q_k)\| = 0$.

Other conditions on $p_k$ and $\alpha_k$ which also yield global convergence of the cost function are the Goldstein and Strong Wolfe Conditions [50].

We now review three common ways to compute an ascent direction $p_k$. The first is called the *method of steepest ascent*, where

$$p_k = \nabla F(q_k), \tag{9.11}$$

which clearly satisfies (9.8). Convergence to $q^*$ in this case is linear, but the computational cost incurred calculating $\nabla F(q_k)$ is low compared to *Newton* and *Quasi-Newton* methods [50, 40]. A Newton or quasi-Newton direction is found by considering the quadratic model for $F$ at $q_k$, given by

$$m(p) = F(q_k) + p^T \nabla F(q_k) + \frac{1}{2} p^T B_k p \approx F(q_k + p) - \mathcal{O}(p^3),$$

where $B_k \approx \Delta F(q_k)$. If $B_k$ is negative definite, then $m(p)$ is maximized at $p^*$ such that $\nabla_p m(p^*) = \mathbf{0}$. That is

$$\nabla F(q_k) + B_k p^* = \mathbf{0}, \tag{9.12}$$

from which it follows that $p^* = -B_k^{-1} \nabla F(q_k)$ is the unique maximizer of $m(p)$. Checking (9.8), we see that

$$\nabla F(q_k)^T p^* = -\nabla F(q_k)^T B_k^{-1} \nabla F(q_k),$$

which is guaranteed to be positive when $B_k$ is negative definite. Letting $B_k = \Delta F(q_k)$ in (9.12), we see that the Newton direction is found by solving

$$\Delta F(q_k) p_k = -\nabla F(q_k) \tag{9.13}$$

for $p_k$. In this case, convergence is quadratic, but the computational cost incurred determining $\Delta F(q_k)$ and then solving (9.13) can be very high [50]. A compromise between convergence and cost can be accomplished by using a quasi-Newton direction, which is found by solving

$$B_k p_k = -\nabla F(q_k) \tag{9.14}$$

for $p_k$, where $B_k$ is an approximation of $\Delta F(q_k)$. Observe that the method of steepest ascent can be interpreted as a quasi-Newton direction with $B_k = -I$. For a general quasi-Newton direction, if the approximation is negative definite and close enough to $\Delta F(q_k)$, then convergence to $q^*$ can be shown to be superlinear [50, 40], while the cost of computing $B_k$ and then solving (9.14) can be much less than computing $H_k$ and then solving (9.13). As we see in the next section, there are algorithms, such as the Newton Conjugate Gradient method (Algorithm 145), which compute $B_k$ and solve (9.14) simultaneously.

We now state the following Corollary to Theorem 143.

COROLLARY 144. *(p. 45 [50]) Let $F$ be defined as in (3.2) with Assumptions 15. Suppose that for every $k \geq 0$ in Algorithm 142, $\alpha_k$ satisfies (9.9) and (9.10). If $p_k$ is the steepest ascent direction (9.11) for every $k$, then $\lim_{k\to\infty} \|\nabla F(q_k)\| = 0$. If $p_k$ is a Newton or quasi-Newton direction as in (9.13) or (9.14) for every $k$, then $\lim_{k\to\infty} \inf \|\nabla F(q_k)\| = 0$. Furthermore, if there is some $M > 0$ such that*

$$\|B_k\|\|B_k^{-1}\| \leq M$$

*with $B_k$ negative definite for all $k$, then $\lim_{k\to\infty} \|\nabla F(q_k)\| = 0$.*

To deal with the case when $B_k$ is not negative definite for some $k$, many schemes have been devised to impose this condition [50]. For example, small multiples of the identity are added to $\Delta F(q_k)$, or one simply creates a negative definite approximation to $\Delta F(q_k)$ by flipping the signs of the positive eigenvalues of $\Delta F(q_k)$. Other *diagonal modifications* include changing positive eigenvalues of $\Delta F(q_k)$ for ones with small negative eigenvalues, or increasing the diagonal elements encountered during the Cholesky factorization (where necessary) of $\Delta F(q_k)$ to ensure that its eigenvalues are sufficiently positive (p. 143-145 [50]).

Newton Conjugate Gradient Method

One quasi-Newton method used to solve (9.14) for $p_k$, while simultaneously computing the Hessian approximation $B_k$, is the Newton Conjugate Gradient (CG) method. Determining a search direction $p_k$ by solving $B_k p_k = -\nabla F(q_k)$ can be expensive. The goal of Newton CG is to efficiently solve

$$Bp = -g$$

where $B \in \Re^{n\times n}$ and $p, g \in \Re^n$. The Newton Conjugate Gradient method accomplishes this goal by creating a sequence $\{p_j\}$ which converges to $p^* = -B^{-1}g$ in at most $n$ iterations when $B$ is negative definite. Newton CG is implemented by the following algorithm, which minimizes the quadratic

$$\phi(p) = \frac{1}{2}p^T Bp + g^T p$$

when $B$ is symmetric negative definite, using the line search Algorithm 142.

ALGORITHM 145 (NEWTON CONJUGATE GRADIENT METHOD). *(p. 108 and 111 [50]) Implement Algorithm 142:*

1. *The ascent direction, $d_j$, at the $j^{th}$ step is*

$$d_j = \nabla\phi_{j-1} - \frac{< \nabla\phi_{j-1}, d_{j-1} >_B}{||d_{j-1}||_B^2} d_{j-1}.$$

*2. The step length at the $j^{th}$ step is found by solving*

$$\tau_j = \arg\max_{\tau>0} \phi(p_j + \tau d_j).$$

*3.* $p_{j+1} = p_j + \tau_j d_j$.

Algorithm 145 is a Gram-Schmidt process with respect to $- < \cdot, \cdot >_B$, which is an inner product when $B$ is negative definite. Thus, when $B$ is negative definite, $\{d_j\}$ form a $(-B)$-orthogonal set in $\Re^n$. Furthermore, the algorithm depends on $B$ only in $< \cdot, \cdot >_B$. Thus, Algorithm 145 does not require that the full matrix $B$ be computed. Rather, only the vector-matrix multiplications $Bd_{j-1}$ need to be computed in step 1.

THEOREM 146. *(p. 103 [50]) If $B$ is symmetric negative definite, then for any initial $p_0 \in \Re^n$, $p_j \to p^*$ in at most $n$ steps.*

To deal with the case when $B$ is not negative definite, one stops Algorithm 145 when either of the following occur:

1. CG residual $||Bp_j + g|| \leq \epsilon$, where $\epsilon$ denotes a stopping tolerance.
2. Positive curvature detected, i.e., $d_j^T B d_j > 0$.

These criteria are called *Steihaug's Stopping Criteria* (p. 75-6 [50]).

Constrained Line Searches

Now we address the types of line searches that can be used to solve the constrained system (1.9). The goal of constrained line searches is to build a sequence $\{q_k\}_{k=1}^{\infty}$ of approximates to $q^*$ such that (9.3), (9.4) and (9.5) are satisfied for each $k$. The idea is that at $q_k$, one computes an ascent direction $p_k$, and then projects (or "bends") it so that $p_k$ is still an ascent direction and so that $q_{k+1} = q_k + \alpha p_k$ remains feasible. That is, $\nabla F(q_k)^T p_k > 0$, and the constraints must be satisfied at $q_{k+1}$: $c_i(q_{k+1}) \geq 0$ for inequality constraints ($i \in \mathcal{I}$), and $c_i(q_{k+1}) = 0$ for equality constraints ($i \in \mathcal{E}$) (see Theorem 16).

In chapter 6, we argued that we could ignore the equality constraints, since stationary points of (1.9),

$$\max_{q \in \Delta} F(q, \beta),$$

in the interior of $\Delta$ are stationary points of (3.1)

$$\max_{q \in \Delta_\mathcal{E}} F(q, \beta)$$

(see Remark 19). Along this same line of reasoning, let us consider a solution $(q^*, \beta)$ of (1.9) in the interior of $\Delta$. Then, by Theorem 20, $\Delta F(q^*, \beta)$ is non-positive definite

on $\ker J$, where $J$ is defined in (3.7). However, $(q^*, \beta^*)$ is guaranteed to be a solution of (3.1) only if $\Delta F(q^*, \beta)$ is negative definite on $\ker J$. Furthermore, the functions $G$ and $D$ may very well not even be defined on $\Delta_{\mathcal{E}}$ (as is the case for the Information Distortion and the Information Bottleneck cost functions (2.34) and (2.35) respectively). This is of course not a problem for the theorist, but a definite problem for a numerical algorithm. For these reasons, when looking for solutions of (1.9), we use constrained optimization techniques which enforce the negativity constraints.

A constraint $c_i(q_k)$ is said to be *active* if $c_i(q_k) = 0$. $c_i(q_k)$ is *inactive* if $c_i(q_k) > 0$. Thus, equality constraints are always active.

REMARK 147. *Once the active constraints are identified, then Theorem 143 can be used to assure that constrained line searches, under the assumptions of the theorem, procure a stationary point (p 95-6 [40],[50]).*

A computational problem is that the projection can be expensive. So projected line searches work best for simple inequality constraints, such as the non-negativity constraints imposed by (1.11):

$$q(\nu|y) \geq 0 \quad \forall \ y \in Y \quad \text{and} \quad \forall \ \nu \in Y_N. \tag{9.15}$$

We now review three common ways to compute a projected ascent direction which deals with the nonnegativity constraints (9.15). The first is the *projected gradient method*, where one finds the steepest ascent direction, then projects if necessary

$$p_k = q_k - \max(q_k - \nabla F(q_k), \eta),$$

where $\eta \in \Re^n$, with components greater than zero. As with the steepest ascent method, convergence in this case is linear (p. 95-6 [40]), and the computational cost is low [50, 40].

Projected Newton and quasi-Newton methods find an ascent direction by solving the system

$$B_{k_{\text{Red}}} p_k = -\nabla F(q_k) \tag{9.16}$$

where $B_{k_{\text{Red}}}$ is an approximation of the *reduced Hessian*, $H_{k_{\text{Red}}}$, a non-negative definite matrix defined by

$$[H_{k_{\text{Red}}}]_{ij} := \begin{pmatrix} \delta_{ij} & \text{if either } c_i(q_k) \text{ or } c_j(q_k) \text{ are active} \\ [\Delta F(q_k)]_{ij} & \text{otherwise.} \end{pmatrix}$$

Convergence in this regime is superlinear (p.565-6 [50], p.90 [40]). For the simple non-negativity constraints, $c_i = [q]_j \geq 0$ for every $j$, $1 \leq j \leq n$ and $i \in \mathcal{I}$, Newton and quasi-Newton projection methods behave like steepest ascent on the active constraints and like Newton and Quasi-Newton methods on the inactive constraints. This claim becomes evident by rewriting the quantizer $q$ as

$$q = \begin{pmatrix} q_I \\ q_A \end{pmatrix},$$

where the subscript $A$ denotes the components of $q$ which are zero (i.e. those $j$ for which $c_i(q_k) = [q]_j = 0$ for some $i \in \mathcal{I}$), and the subscript $A$ denotes those components of $q$ which are strictly larger than zero (i.e. those $j$ for which $c_i(q_k) = [q]_j > 0$ for some $i \in \mathcal{I}$). Similarly rewriting $\nabla F(q_k)$ and $B_{k_{\text{Red}}}$ using this convention,

$$\nabla F(q_k) = \begin{pmatrix} \nabla F_I \\ \nabla F_A \end{pmatrix}$$

and

$$B_{k_{\text{Red}}} = \begin{pmatrix} B_I & 0 \\ 0 & I \end{pmatrix},$$

we see that

$$p_k = -B_{k_{\text{Red}}}^{-1} \nabla F(q_k) = \begin{pmatrix} -B_I^{-1} \nabla F_I \\ -\nabla F_A \end{pmatrix}.$$

Augmented Lagrangian

We want a fast, rigorous quasi-Newton algorithm which takes into account all the constraints imposed by $\Delta$ (1.11). Many optimization methods consider either all equality constraints or all inequality constraints. The Augmented Lagrangian algorithm is one method which takes into account both kinds of constraints. It is similar to other *quadratic penalty methods* [50] in that the constraints to the problem are subtracted from $F$ to create a new cost function to maximize, such as,

$$P(q, \mu) := F(q) - \frac{1}{2\mu} \sum_j (c_j(q))^2,$$

where $c_j(q) := \sum_{\mathcal{Y}_N} q(y_N|y) - 1$, is the constraint imposed for every $y_j \in \mathcal{Y}$. The more infeasible the constraints $c_j(q)$ (when $\sum_{\mathcal{Y}_N} q(y_N|y) - 1 >> 0$), the harsher the penalty in $P$. $P$ is *ill conditioned* as $\mu \to \infty$.

The Augmented Lagrangian, however, avoids the ill-conditioning of other penalty methods (as $\mu \to \infty$) by introducing explicit approximations of the Lagrange multipliers into the cost function at each optimization iteration (p.494-5,498,513-14 [50]) (Theorem 148). These approximations are constructed in such a way so that the solution to the algorithm satisfies the *KKT* conditions [50] (Lemma 150).

The new cost function to maximize, the Augmented Lagrangian $\mathcal{L}_A$, is defined as

$$\mathcal{L}_A(q, \lambda^l, \mu_l) := \mathcal{F}(q) + \sum_{j \in \mathcal{E}} \lambda_j^l c_j(q) - \frac{1}{2\mu_l} \sum_{j \in \mathcal{E}} c_j(q)^2,$$

which deals with the equality constraints $c_j(q) = \sum_{\mathcal{Y}_N} q(y_N \mid y) - 1 = 0$ (p. 514 [50]). To deal with the non-negativity constraints, a Newton CG projected line search is used.

The next theorem shows that we don't need $\mu_l \to 0$ to determine $q^*$.

THEOREM 148. *(p. 519 [50]) If* $q^* = \arg\max_{q\in\Delta} F$ *such that* $\Delta F(q^*)$ *is negative definite on* $\ker J$*, then there exists* $\bar{\mu} > 0$ *such that* $q^* = \arg\max \mathcal{L}_A(q, \lambda^*, \mu)$ *for* $\mu \in (0, \bar{\mu}]$.

ALGORITHM 149 (AUGMENTED LAGRANGIAN METHOD). *(p. 515,523 [50]) There are three nested iterations. The first is the Augmented Lagrangian or outer iteration, subscripted by* $l$*. The second is the optimization or inner iteration, subscripted by* $k$*. The third is the line search iteration implicit in step 1.*

*Choose* $q_0 \in \Delta, \mu_0 > 0, 0 < \tau, \epsilon, s < 1$*, and set* $l = 0$.

1. *Solve* $q_l = \arg\max \mathcal{L}_A(q, \lambda^l, \mu_l)$ *using a projected line search which satisfies the Wolfe Conditions, and Newton CG is used to compute the ascent direction* $p_k$ *by solving*
$$B_{k_{\text{Red}}} p_k = -\nabla \mathcal{L}_A(q^k, \lambda^l, \mu_l).$$
2. $\lambda_i^{l+1} = \lambda_i^l - c_i(q_l)\mu_l$
3. $\mu_{l+1} = s\mu_l$
4. *Stop if both of the following occur:*
$$||P_{[\eta,\infty)} \nabla \mathcal{L}_A(q^k, \lambda^l, \mu_l)|| \leq \tau$$
$$||c_y(q)|| < \epsilon$$
5. *Let* $l = l + 1$ *and repeat steps 1-4.*

LEMMA 150. *Step 2 of Algorithm 149 assures that* $(q_l, \lambda_l)$ *satisfies the KKT conditions for every* $l$.

*Proof.* $\nabla_q \mathcal{L}_A = \nabla F - \sum_j \left(\lambda^l - \frac{c_j(q)}{\mu_l}\right) \nabla c_j(q)$. Since $\nabla \mathcal{L}_A(q_l) = 0$, then it follows that

$$\nabla F = \sum_j \left(\lambda_j^l - \frac{c_j(q)}{\mu_l}\right) \nabla c_j(q)$$

if and only if the Lagrange multipliers corresponding with constraint $c_j(q)$ is

$$\lambda^* = \lambda^l - \frac{c(q)}{\mu_l}.$$

□

The following theorem gives conditions under which there is a maximizer $q_l$ of $\mathcal{L}_A$ that lies close to $q^*$, and gives error bounds on $q_l$ obtained from performing Algorithm 149 at iteration $l$.

THEOREM 151. *([50] p.521) Let $q^*$ be a solution of (1.9), with corresponding vector of Lagrange multipliers $\lambda^*$, such that $\Delta F(q^*)$ is negative definite on* $\ker J$*, and let $\bar{\mu}$ be chosen as in Theorem 148. Then there exists $\delta, \epsilon, m > 0$ such that for all $\lambda^l$ and $\mu_l$ satisfying*

$$\|\lambda_l - \lambda^*\| \leq \frac{\delta}{\mu_l}$$

*for $\mu_l \leq \bar{\mu}$, the problem*

$$\begin{array}{c} \max_q \mathcal{L}_A(q, \lambda^l, \mu_l) \quad \textit{subject to} \\ \|q - q^*\| \leq \epsilon \end{array}$$

*has a unique solution $q_l$. Furthermore, we have*

$$\|q_l - q^*\| \leq m\mu_l \|\lambda_l - \lambda^*\|.$$

## Optimization Schemes

In this section, we investigate and compare three different approaches to solving the optimization problem (1.9) for $\beta = \mathcal{B} \in [0, \infty)$. Two of them use the method of *annealing* to find extrema by starting in the interior of the feasible region $\Delta$ and incrementing $\beta$ in sufficiently small steps until $\beta = \mathcal{B}$. The third method is based on the observation (Theorem 153) that an optimal solution of (1.9) for $\mathcal{B} = \infty$ lies generically at a vertex of the feasible region if $D(q)$ is convex. As a consequence of this fact, in Theorem 154 we formulate an equivalent problem to (1.9), and pose Algorithm 155, called *vertex search*, to solve it. This algorithm finds an optimal solution of (1.9) when $D(q) = D_{eff}$ under mild conditions (Theorem 156).

When searching for the extrema of a general optimization problem, there is no known theory indicating whether using continuous, gradient-type algorithms is cheaper than searching over a finite, large set which contains the extrema. We compare these methods in section 9 of this chapter on synthetic data.

### Annealing

A basic annealing algorithm is given by Algorithm 1. In this regime, one tracks the optimal solutions, $(q_k, \beta_k)$, of (1.9) for $\beta_k$ values incremented in small steps from $\beta_0 = 0$ to $\beta_{\max} = \mathcal{B}$ in order to find $q^* = \arg\max_{q \in \Delta}(G + \mathcal{B}D)$. At $\beta_0 = 0$, the optimal solution to (1.9) is a maximum of $G$. When $G$ is strictly concave, this solution is unique. For the Information Distortion problem (2.34), the optimal solution at $\beta_0 = 0$ is the unique uniform solution $q(Y_N|Y) = q_{\frac{1}{N}}$ (Lemma 79). The Information Bottleneck problem (2.35) also has $q(Y_N|Y) = q_{\frac{1}{N}}$ as a solution for $\beta_0 = 0$, but it is not unique since $G = -I(Y; Y_N)$ is not strictly concave.

We have implemented two annealing algorithms which differ in the optimization techniques implemented in step 3 of Algorithm 1. The first uses an Augmented

Lagrangian algorithm (Algorithm 149). The second is an *implicit solution* algorithm, introduced in [22, 29], which we describe next.

The implicit solution algorithm is based on the observation that extrema of $F$ can be found by setting the gradient of the Lagrangian (3.3) with respect to the quantizer $q(Y_N|Y)$ to zero [22]

$$\begin{aligned} 0 &= (\nabla_q\Big(F + \sum_j \lambda_j \sum_\nu q_{\nu j} - 1\Big))_{\nu k} \\ &= (\nabla_q H)_{\nu k} + \beta(\nabla_q D_{eff})_{\nu k} + \lambda_k \\ &= -p(y_k)\Big(\frac{\ln q_{\nu k}}{\ln 2} + \frac{1}{\ln 2}\Big) + \beta(\nabla_q D_{eff})_{\nu k} + \lambda_k \Leftrightarrow \\ 0 &= \ln q_{\nu k} - \beta \ln 2 \frac{(\nabla D_{eff})_{\nu k}}{p(y_k)} - \mu_k \end{aligned} \tag{9.17}$$

where $\mu_k = \frac{\lambda_k \ln 2}{p(y_k)} - 1$. Using this,

$$\begin{aligned} \ln q_{\nu k} &= \beta \ln 2 \frac{(\nabla D_{eff})_{\nu k}}{p(y_k)} + \mu_k \Leftrightarrow \\ q_{\nu k} &= e^{\mu_k} e^{\beta \ln 2 \Big(\frac{(\nabla D_{eff})_{\nu k}}{p(y_k)}\Big)} \end{aligned} \tag{9.18}$$

The constraint on $q$ requires that

$$\begin{aligned} 1 &= \sum_\nu q_{\nu k} \Rightarrow \\ 1 &= e^{\mu_k} \sum_\nu e^{\beta \ln 2 \Big(\frac{(\nabla D_{eff})_{\nu k}}{p(y_k)}\Big)} \Leftrightarrow \\ e^{\mu_k} &= \frac{1}{\sum_\nu e^{\beta \ln 2 \Big(\frac{(\nabla D_{eff})_{\nu k}}{p(y_k)}\Big)}} \end{aligned} \tag{9.19}$$

We can substitute this in equation (9.18) and obtain an implicit expression for the optimal $q(y_\nu|y_k)$,

$$q_{\nu k} = \frac{e^{\beta \ln 2 \Big(\frac{(\nabla D_{eff})_{\nu k}}{p(y_k)}\Big)}}{\sum_\nu e^{\beta \ln 2 \Big(\frac{(\nabla D_{eff})_{\nu k}}{p(y_k)}\Big)}}. \tag{9.20}$$

For a fixed value of $\beta$ we use a fixed point iteration

$$q_{n+1} := f(q_n),$$

where $f$ is the right hand side of expression (9.20), to generate a sequence $\{q_n\}$ to find a solution $q^*$ for the optimization problem (1.9).

We do not have a complete theoretical understanding of the convergence of the implicit solution algorithm.

REMARK 152. *The "solutions" $\{(q_k, \beta_k)\}$ found by the authors in [22, 29] and given in Figure 1 are shifted in $\beta$ when $\beta$ is small (see Figure 20). This discrepancy is due to the fact that the authors incorrectly used the expression*

$$q_{\nu k} = \frac{e^{\beta\left(\frac{(\nabla D_{eff})_{\nu k}}{p(y_k)}\right)}}{\sum_\nu e^{\beta\left(\frac{(\nabla D_{eff})_{\nu k}}{p(y_k)}\right)}}$$

*instead of (9.20) when implementing the implicit solution method.*

Vertex Search

We now describe a method which can solve (1.9) when $D(q)$ is convex and $\beta \to \infty$, as is the case with the Information Distortion problem (2.34). The method simply searches over the vertices of the constraint space $\Delta$, which is a product of simplices, for a solution. This approach is justified by the following theorem

THEOREM 153. *[29] Let $D(q)$ from (3.2) be convex, and let $E$ be the set of vertices of $\Delta$. Then*

$$\max_E D(q) \geq \max_\Delta D(q).$$

This result allows us to reformulate the problem (1.9) as follows

THEOREM 154. *[29] Suppose that $D(q)$ is convex and let $E$ be the set of vertices of $\Delta$. The optimal solution of the problem (1.9) with maximal possible value of $D(q)$ can be found by the following algorithm:*

1. *Find a vertex $e \in E$ such that*

$$D(e) := \max_E D(q)$$

2. *Assume that $e$ is a strict maximum of $D(q)$ on the set $E$. That is, for all neighboring vertices $e_i \in E$ we have $D(e_i) < D(e)$. Then $e$ is an optimal solution of (2.16) with maximal possible value of $D(q)$.*

3. *Assume that $e = e_1$ is not a strict maximum. Then there are neighboring vertices $e_1, \ldots, e_k$ such that $D^* := D(e_i) = D(e_j)$ for all $1 \leq i, j \leq k$. Consider the region $Q_{y_1} \times \ldots \times Q_{y_s}$, where $Q_{y_j} \subset \Delta_{y_j}$ is the simplex spanned by the projection of these vertices to $\Delta_{y_j}$. For all $j$, take $D_{y_j} \subset Q_{y_j}$ to be the maximal sub-simplex with the property that $D(x) = D^*$ for all $x \in D_{y_1} \times \ldots \times D_{y_s}$. Then the solution of (2.16) is the product of the barycenters of $D_{y_i}$.*

Theorem 154 justifies the following algorithm (see Figure 26).

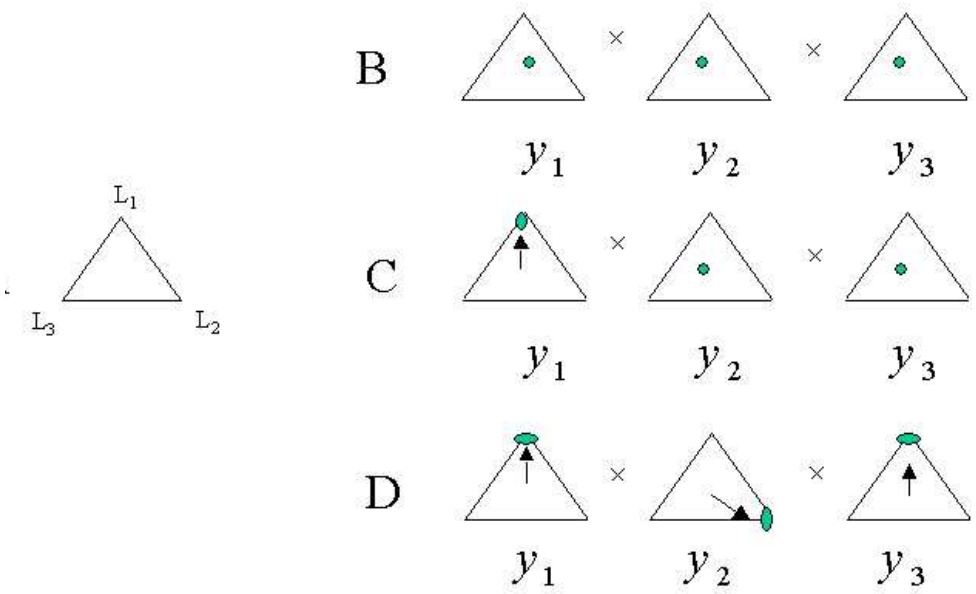



Figure 26. The vertex search algorithm, used to solve (1.9) when $D(q)$ is convex and $\mathcal{B} = \infty$, shown here for $N = 3$, $\mathcal{Y}_N = \{1, 2, 3\}$, and $K = 3$. A: A simplex $\Delta_y$. Each vertex $\nu \in \mathcal{Y}_N$ corresponds to the value $q(\nu|y) = 1$. B: The algorithm begins at some initial $q(\nu|y)$, in this case with $q(\nu|y) = 1/3$ for all $y$ and $\nu$. C: Randomly assign $y_1$ to a class $\nu = 1$. D: Assign $y_2$ consecutively to each class of $\mathcal{Y}_N = \{1, 2, 3\}$, and for each such assignment evaluate $D(q)$. Assign $y_2$ to the class $\nu$ which maximizes $D(q)$. Repeat the process for $y_3$. Shown here is a possible classification of $y_1$, $y_2$ and $y_3$: $y_1$ and $y_3$ are assigned to class 1, and $y_2$ is assigned to class 2. Class 3 remains empty.

Algorithm 155 (Vertex Search).

1. *We start the search from the uniform solution* $q_{\frac{1}{N}}$.

2. *Randomly select* $y_1$ *and evaluate the function* $D(q)$ *at all the vertices of* $\Delta_{y_1}$, *such that* $q(\nu|y_1) = 1$ *for some class* $\nu \in \mathcal{Y}_N$ *and* $q(\eta|y_1) = 0$ *for all other classes* $\eta \in (\mathcal{Y}_N \setminus \{\nu\})$. *Perform this calculation for each* $\nu \in \mathcal{Y}_N$. *Select the assignment of* $y_1$ *to a class which gives the maximal value of* $D(q)$.

3. *Repeat step 2 with* $y_2, y_3, \ldots$ *until all of the* $K$ *elements* $y_k \in \mathcal{Y}$ *are assigned classes. The resulting deterministic quantizer is a vertex* $e$ *of* $\Delta$.

4. *Starting from the vertex* $e$ *found in step 3, we repeat steps 1-3 until a local maximum in the set* $E$ *is found.*

5. *The steps 1-4 are repeated many times to avoid local maxima.*

The vertex search algorithm converges to a local maximum under certain conditions when $D(q) = D_{eff}$. The notation

$$y \in C(\nu)$$

means that the element $y \in \mathcal{Y}$ has been assigned to class $\nu \in \mathcal{Y}_N$. That is, $q(\nu|y) = 1$.

THEOREM 156. *[29] The point e, obtained by a vertex search, is a local maximum of $D_{eff}$ if for each k, when $q(\nu|y_k)$ is determined, we have*

$$p(x, y_k) << \sum_{y_i \in C(\nu), i \neq k} p(x, y_i), \quad p(y_k) << \sum_{y_i \in C(\nu), i \neq k} p(y_i)$$

*for each class $\nu \in \mathcal{Y}_N$.*

## A New Numerical Algorithm

In chapter 7, we were interested in finding the stationary points of (1.9). In this chapter, we address the issue of finding *solutions* of (1.9), which are stationary points such that $\Delta F$ is negative definite on $\ker J$ (Theorem 20). We now incorporate the ideas from both approaches. We apply continuation methods and our knowledge of the bifurcation structure into Algorithm 1, which can potentially aid in the search for solutions of (1.9) by minimizing the arbitrariness of the choice of the algorithm's parameters. We apply one of the optimization schemes from this chapter to perform the optimization.

Implementing continuation techniques (Algorithm 131) minimizes the arbitrariness of the choice of the parameters in Algorithm 1. Specifically, these techniques determine $d_k$ in step 1, and choose an initial guess $q_{k+1}^{(0)}$ in step 2. This alleviates the need for the perturbation $\eta$. Furthermore, continuation methods provide explicit estimates of the Lagrange multipliers,

$$\lambda_{k+1}^{(0)} = \lambda_k + d\partial_s \lambda_k,$$

for the equality constraints, which could improve the performance of methods, such as the Augmented Lagrangian method (Algorithm 149) in step 3 of Algorithm 1, which depend on explicit approximations to $\lambda$. And lastly, applying bifurcation theory in the presence of symmetries indicates how to detect bifurcation of the branch on which the solutions $\{(q_k, \beta_k)\}$ reside, and where to search for a desired solution branch once bifurcation is detected. This knowledge yields an initial guess $q_{k+1}^{(0)}$ in step 2 once a bifurcation is detected. The cost savings of these changes can be significant, especially when continuation is used in conjunction with a Newton type optimization scheme which explicitly uses the Hessian $\Delta F(q_k, \beta_k)$ (see (7.5), (7.14), (9.13), (9.14), and (9.16)). Otherwise, the CPU time incurred from solving (7.16) may outweigh this benefit.

We now provide an algorithm which incorporates the annealing algorithm (Algorithm 1), the bifurcation theory from chapter 6, the continuation ideas from chapter 7, and potentially an optimization scheme from this chapter.

ALGORITHM 157. *Let $q_0$ be the maximizer of $\max_{q\in\Delta} G$, $\lambda_0$ be defined as in (7.19), $\beta_0 = 1$, and $d > 0$. Iterate the following steps until $\beta_{\mathcal{K}} = \mathcal{B}$ for some $\mathcal{K} > 0$.*

1. *Perform $\beta$-step: solve (7.16) and (7.17) for $(\partial_s q_k^T \;\; \partial_s \lambda_k^T)^T$ and select $\beta_{k+1} = \beta_k + \Delta\beta_k$, where $\Delta\beta_k = d\frac{\mathrm{sgn}(\cos\theta)}{\sqrt{||\partial_s q_k||^2 + ||\partial_s \lambda_k||^2 + 1}}$ and $\theta$ is defined as in (7.18).*

2. *The initial guess at $\beta_{k+1}$ is $\begin{pmatrix} q_{k+1}^{(0)} \\ \lambda_{k+1}^{(0)} \end{pmatrix} = \begin{pmatrix} q_k + d\partial_s q_k \\ \lambda_k + d\partial_s \lambda_k \end{pmatrix}$.*

3. *Optimization: solve*

$$\begin{array}{c} \max_{q\in\Delta} G(q) + \beta_{k+1} D(q) \quad \textit{constrained by} \\ P(q_k, \lambda_k, \beta_k) - d = \mathbf{0} \end{array}$$

   *to get the maximizer $q_{k+1}$ and vector of Lagrange multipliers $\lambda_{k+1}$, using the initial guess $(q_{k+1}^{(0)}, \lambda_{k+1}^{(0)})$. The function $P$ is defined in (7.12).*

4. *Check for bifurcation: compare the sign of the determinant of an identical block of each of*

$$\Delta[G(q_k) + \beta_k D(q_k)] \textit{ and } \Delta[G(q_{k+1}) + \beta_{k+1} D(q_{k+1})].$$

   *If a bifurcation is detected, then set $q_{k+1}^{(0)} = q_k + d_k \cdot \boldsymbol{u}$ where $\boldsymbol{u}$ is defined as in (6.57) for some $m \le M$, and repeat step 3.*

One might remark why we use an optimization scheme in step 3. Obviously, this method will not be attracted to the stationary points which are not solutions of (1.9), as may happen when all of the bifurcating branches are subcritical for example. We observe in practice that searching for a solution in the bifurcating direction in this scenario may still have significant cost benefit over simply perturbing the solution as is done in Algorithm 1.

We have not fully explored this algorithm numerically.

## Numerical Results

All of the results presented here are for the Information Distortion problem (2.34),

$$\max_{q\in\Delta}(H(q) + \beta D_{eff}(q)).$$

| Algorithm | Cost in MFLOPs | | | $I(X;Y_N)$ in bits | | |
|---|---|---|---|---|---|---|
| N | 2 | 3 | 4 | 2 | 3 | 4 |
| Lagrangian | 431 | 822 | 1,220 | 0.8272 | 1.2925 | 1.6269 |
| Implicit Solution | 38 | 106 | 124 | 0.8280 | 1.2942 | 1.6291 |
| Vertex Search | 6 | 18 | 21 | 0.8280 | 1.2942 | 1.6291 |

Table 4. [29] Comparison of the optimization schemes on synthetic data. The first three columns compare the computational cost in FLOPs. The last three columns compare the value of $D_{eff} = I(X;Y_N)$, evaluated at the optimal quantizer obtained by each optimization algorithm..

We created software in MATLAB to implement the Augmented Lagrangian (Algorithm 149), the Vertex Search (Algorithm 155), and the implicit solution algorithm (9.20) to both synthetic and physiological data sets to determine solutions of (2.34).

Synthetic Data

We analyze the performance of the three optimization schemes on the Four Blob Problem introduced in chapter 1 and Figure 1. Table 4 gives a comparison of the Augmented Lagrangian and the implicit solution optimization algorithms for this data set. For $N = 2$, 3 and 4, left side of the table shows computational cost of each and the right side indicates the maximal value of $D_{eff}$ procured by each algorithm. The vertex search was the fastest and the Augmented Lagrangian the slowest of the three with an order of magnitude difference between each two algorithms. The values of the cost function are almost identical. Each algorithm has its advantages, though, as the Augmented Lagrangian (Algorithm 149) gives a point that satisfies the $KKT$ conditions (Corollary 144 and Lemma 150) and the Vertex Search (Algorithm 155) does so under certain conditions (see Theorem 156). Although we do not have a complete theoretical understanding of the convergence of the implicit solution algorithm (9.20), in particular, the fact that we do not understand the solutions we get for $0 < \beta << \infty$, it works very well in practice as $\beta \to \infty$.

Physiological Data

A biological system that has been used very successfully to address aspects of neural coding [7, 15, 44, 48, 76] is the cricket's cercal sensory system. It provides the benefits of being simple enough so that all output signals can be recorded, yet sufficiently elaborate to address questions about temporal and collective coding schemes. The cricket's cercal system is sensitive to low frequency, near-field air displacement stimuli [38]. During the course of the physiological recording, the system was stimulated with air current stimuli, drawn from a band-limited (5-500Hz) Gaussian white

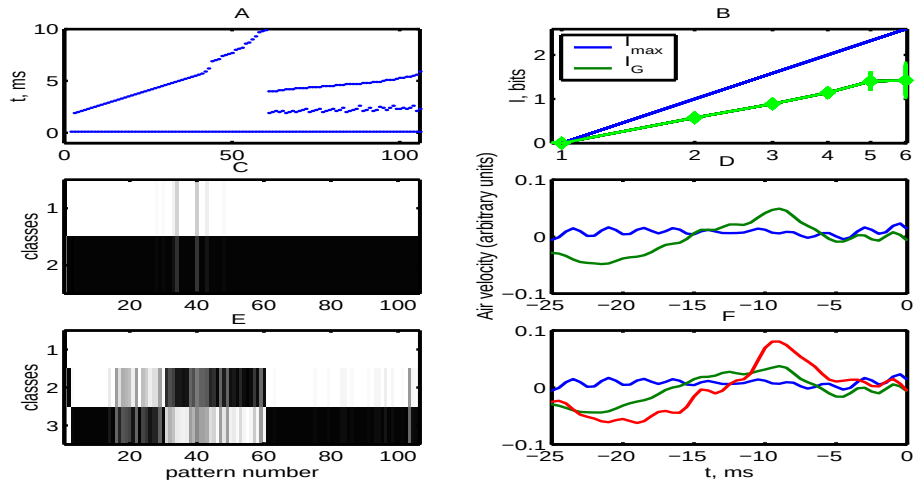



Figure 27. *[29] Results from the information distortion method. A: All the response spike patterns that were analyzed. Each dot represents the occurrence of a single spike. Each column of dots represents a distinct sequence of spikes. The y axis is the time in ms after the occurrence of the first spike in the pattern. The x axis here and below is an arbitrary number, assigned to each pattern. B: The lower bound of I (dashed line) obtained through the Gaussian model can be compared to the absolute upper bound $I = \log_2 N$ for an N class reproduction (solid line). C: The optimal quantizer for $N = 2$ classes. This is the conditional probability $q(\nu|y)$ of a pattern number y from (A) (horizontal axis) belonging to class $\nu$ (vertical axis). White represents zero, black represents one, and intermediate values are represented by levels of gray. D: The means, conditioned on the occurrence of class 1 (dotted line) or 2 (solid line). E: The optimal quantizer for $N = 3$ classes. F: The means, conditioned on the occurrence of class 1 (dotted line), 2 (solid line) or 3 (dashed line)..*

noise (GWN) source [75]. We apply the method to intra-cellular recordings from identified inter-neurons in this system.

When applying the method to this data, the joint stimulus/response probability $p(x, y)$ needs to be estimated. We use $\tilde{D}_{eff}$ (2.28) in place of $D_{eff}$, and the optimization scheme (2.29). Figure 27 illustrates the data set and optimal quantizers for this system. Sequences 2 through 105 in A were obtained by choosing 10 ms sequences from the recording which started with a spike (at time 0 here). Sequences in which the initial spike was preceded by another spike closer than 10 ms were excluded. Sequence 2 contains a single spike. Sequences 3-59 are doublets. Sequences 60-105 are triplets. Sequence 1 is a well isolated empty codeword (occurrences were chosen to be relatively far from the other patterns). Each pattern was observed multiple times (histogram not shown).

Panels C–F show the results of applying the information distortion approach to this data set. The optimal quantizer for the $N = 2$ reproduction is shown in panel

| Algorithm | Cost in GFLOPs | | | | $I(X, Y_N)$ in bits | | |
|---|---|---|---|---|---|---|---|
| N | 3 | 4 | 5 | | 3 | 4 | 5 |
| Lagrangian | 13 | 29 | 59 | | 0.18 | 0.18 | 0.16 |
| Implicit Solution | 7 | 11 | 9 | | 0.43 | 0.80 | 1.14 |
| Vertex Search | 31 | 84 | 141 | | 0.44 | 0.85 | 1.81 |

Table 5. [29] Comparison of the optimization schemes on physiological data. The first four columns compare the computational cost in gigaFLOPs. The last four columns compare the value of $D_{eff} = I(X; Y_N)$, evaluated at the optimal quantizer obtained by each optimization algorithm..

C. It isolates the empty codeword in one class (class $\nu = 1$) and all other patterns in another class (class $\nu = 2$). The mean of the stimuli conditioned with the zero codeword (panel D, dotted line), does not significantly deviate from a zero signal. Panels E and F show the results of extending the analysis to a reproduction of $N = 3$ classes. The zero codeword remains in class 1. The former class 2 is split into two separate classes: class 2, which contains the single spike codeword and codewords with an inter-spike interval $ISI > 5ms$, and class 3, which contains all doublets with $ISI < 2ms$ and all triplets. The mean in (D, solid line) is split into two separate class conditioned means (F, solid and dashed line).

In table 5 we compare the three algorithms on the physiological data set. We see that the cost is lowest for the implicit solution algorithm, but the vertex search finds the "best" solution, measured in terms of the value of $D_{eff}$.

# CHAPTER 10

# CONCLUSION

Our explicit goal in this thesis was to solve problems of the form

$$\max_{q(Y_N|Y)\in\Delta} (G(q) + \beta D(q)) \tag{10.1}$$

at some $\beta = \mathcal{B} \in (0, \infty)$ when $G$ and $D$ have symmetry: renaming classes of $Y_N$ leaves the values of $G(q(Y_N|Y))$ and $D(q(Y_N|Y))$ unchanged. The major ingredient to our approach was to build a mathematical theory which describes the bifurcation structure of stationary points of (10.1) for each $\beta \in [0, \mathcal{B}]$. As we have seen, the symmetry dictates the bifurcation structure of solutions to the problem (10.1). Our understanding of the bifurcation structure of these solutions lends itself to the computational problem of solving (10.1) since we know how to detect symmetry breaking bifurcation, and, once this of type of bifurcation is detected, we know in which direction the new branches bifurcate. We presented an algorithm (Algorithm 157) which uses these ideas.

For the Information Distortion method, which concerns itself with the biological problem of deciphering the neural code, we numerically confirmed the bifurcation structure predicted by the theory by implementing continuation techniques. We also presented optimization schemes, such as the Augmented Lagrangian, implicit solution and the vertex search method, to find solutions of the problem (10.1).

Determining the bifurcation structure of stationary points of (10.1), and implementing an efficient algorithm to solve (10.1) are two different things. The former illuminates *how* one might create the latter. Although we have presented Algorithm 157 which incorporates these ideas, we have not yet fully explored the method numerically, which holds the tantalizing prospect of an efficient algorithm to find local solutions of the problem (10.1).

# REFERENCES CITED

[1] L. F. Abbott. http://www.gatsby.ucl.ac.uk/ dayan/book/teaching.html, 2001.

[2] E. D. Adrian and Y. Zotterman. The impulses produced by sensory nerve endings: Part ii: The response of a single end organ. *Journal of Physiology (London)*, 61:151–171, 1926.

[3] E. D. Adrian and Y. Zotterman. The impulses produced by sensory nerve endings: Part iii: Impulses set up by pulse and pressure. *Journal of Physiology (London)*, 61:465–483, 1926.

[4] H. B. Barlow. Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith, editor, *Sensory Communications*. MIT Press, Cambridge, MA, 1961.

[5] M. J. Berry and M. Meister. Refractoriness and neural precision. *Journal of Neuroscience*, 18:2200–2211, 1998.

[6] W. J. Beyn, A. Champneys, E. Doedel, W. Govaerts, Y. A. Kuznetsov, and B. Sandstede. Numerical continuation and computation of normal forms. In *Handbook of Dynamical Systems III*.

[7] D. A. Bodnar, J. Miller, and G. A. Jacobs. Anatomy and physiology of identified wind-sensitive local interneurons in the cricket cercal sensory system. *J. Comp. Physiol. A*, 168:553–564, 1991.

[8] H. Boerner. *Representations of Groups*. Elsevier, New York, 1970.

[9] A. Borst and F. Theunissen. Information theory and neural coding. *Nature Neorscience*, 2:947–957, November 1999.

[10] L. Breiman. *Probability*. Addison-Wesley Publishing Companyy, Menlo Park, CA, 1968.

[11] E. Brown, L. Frank, D. Tang, M. Quirk, and M. Wilson. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18:7411–25, 1999.

[12] G. Cicogna. Symmetry breakdown from bifurcation. *Lettere Al Nuevo Cimento*, 31:600–602, 1981.

[13] G. Cicogna. Bifurcation and symmetries. *Bollettino Un. Mat. Ital.*, pages 787–796, 1982.

[14] G. Cicogna. Bifurcation from topology and symmetry arguments. *Bollettino Un. Mat. Ital.*, pages 131–138, 1984.

[15] H. Clague, F. Theunissen, and J. P. Miller. The effects of adaptation on neural coding by primary sensor interneurons in the cricket cercal system. *J. Neurophysiol.*, 77:207–220, 1997.

[16] J. Conway, R. T. Curtis, S. P. Norton, R. A. Parker, and R. Wilson. *Atlas of Finite Groups*. Clarendon Press, Oxford, 1985. p 236.

[17] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Series in Communication, New York, 1991.

[18] T. Cover and J. Thomas. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1998.

[19] R. R. de Ruyter van Steveninck, G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek. Reproducibility and variability in neural spike trains. *Science*, 275:1805–1808, March 1997.

[20] A. G. Dimitrov and J. P. Miller. Analyzing sensory systems with the information distortion function. In R. B. Altman, editor, *Pacific Symposium on Biocomputing 2001*. World Scientific Publushing Co., 2000.

[21] A. G. Dimitrov and J. P. Miller. Natural time scales for neural encoding. *Neurocomputing*, 32-33:1027–1034, 2000.

[22] A. G. Dimitrov and J. P. Miller. Neural coding and decoding: communication channels and quantization. *Network: Computation in Neural Systems*, 12(4):441–472, 2001.

[23] A. G. Dimitrov, J. P. Miller, and Z. Aldworth. Neural coding and decoding. New Orleans, November 2000. Society for Neuroscience Annual Meeting.

[24] A. G. Dimitrov, J. P. Miller, Z. Aldworth, T. Gedeon, and A. E. Parker. Coding schemes based on spike patterns in a simple sensory system. *Journal of Neuroscience*, 2002.

[25] A. G. Dimitrov, J. P. Miller, Z. Aldworth, and A. Parker. Spike pattern-based coding schemes in the cricket cercal sensory system. *Neurocomputing*, 2002. *(to appear)*.

[26] E. Doedel, H. B. Keller, and J. P. Kernevez. Numerical analysis and control of bifurcation problems in finite dimensions. *International Journal of Bifurcation and Chaos*, 1:493–520, 1991.

[27] D. S. Dummit and R. M. Foote. *Abstract Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[28] R. Durrett. *Probability: Theory and Examples*. Duxbery Press, New York, 1997.

[29] T. Gedeon, A. E. Parker, and A. G. Dimitrov. Information distortion and neural coding. *Canadian Applied Mathematics Quarterly*, 2002.

[30] A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986.

[31] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

[32] B. Girish, J. C. Roddey, and J. P. Miller. A metric for assessing the linearity of neural encoding. In J. Bower, editor, *Annual Computational Neuroscience Meeting, proceedings*, volume to appear, 1997.

[33] M. Golubitsky and D. G. Schaeffer. *Singularities and Groups in Bifurcation Theory I*. Springer Verlag, New York, 1985.

[34] M. Golubitsky, I. Stewart, and D. G. Schaeffer. *Singularities and Groups in Bifurcation Theory II*. Springer Verlag, New York, 1988.

[35] R. M. Gray. *Entropy and Information Theory*. Springer-Verlag, 1990.

[36] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proc. IEEE*, 70:939–952, 1982.

[37] D. H. Johnson, C. M. Gruner, K. Baggerly, and C. Seshagiri. Information-theoretic analysis of the neural code. *J. Comp. Neurosci*, 10(1):47–70, 2001.

[38] G. Kamper and H.-U. Kleindienst. Oscillation of cricket sensory hairs in a low frequency sound field. *J. Comp. Physiol. A.*, 167:193–200, 1990.

[39] H. B. Keller. Numerical solutions of bifurcation and nonlinear eigenvalue problems. In *Applications of Bifurcation Theory*.

[40] C. T. Kelley. *Iterative Methods for Optimization*. SIAM, Philadelphia, 1999.

[41] T. W. Kjaer, J. A. Hertz, and B. J. Richmond. Decoding cortical neuronal signals: Network models, information estimation and spatial tuning. *J. Comp. Neurosci*, 1(1-2):109–139, 1994.

[42] M. J. Korenberg and I. A. Hunter. The identification of nonlinear biological systems: Wiener kernel approaches. *Ann. of Biomed. Eng.*, 18:629–654, 1990.

[43] S. Kullback. *Information Theory and Statistics*. J Wiley and Sons, New York, 1959.

[44] M. A. Landolfa and J. P. Miller. Stimulus-response properties of cricket cercal filiform hair receptors. *J. Com. Physiol. A.*, 177:749–757, 1995.

[45] S. B. Laughlin. Efficiency and complexity in neural coding. Complexity in biological information processing. Wiley, Chichester (Novartis Foundation Symposium 239, 2001. p 177-192.

[46] M. W. Liebeck, C. E. Praeger, and J. Saxl. A classification of the maximal subgroups of the finite alternating and symmetric groups. *Journal of Algebra*, pages 365–383, 1987.

[47] P. Marmarelis and V. Marmarelis. *Analysis of physiological systems. The white noise approach.* Plenum Press, New York, 1978.

[48] J. P. Miller, G. A. Jacobs, and F. E. Theunissen. Representation of sensory information in the cricket cercal sensory system. I. Response properties of the primary interneurons. *J. Neurophys*, 66:1680–1689, 1991.

[49] S. Nirenberg, S. M. Carcieri, A. L. Jacobs, and P. E. Latham. Retinal ganglion cells act largely as independent encoders. *Nature*, 411:698–701, June 2001.

[50] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2000.

[51] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 2003. *at press*.

[52] S. Panzeri, R. S. Petersen, S. R. Schultz, M. Lebedev, and M. E. Diamond. The role of spike timing in the coding of stimulus location in rat somatosensory cortex. *Neuron*, 29:769–777, March 2001.

[53] S. Panzeri, S. R. Schultz, A. Treves, and E. T. Rolls. Correlations and the encoding of information in the nervous system. *Proc. R. Soc. Lond. B*, 266:1001–1012, 1999.

[54] A. Parker, T. Gedeon, and A. Dimitrov. Annealing and the rate distortion problem. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2003. *at press*.

[55] R. S. Petersen, S. Panzeri, and M. Diamond. Population coding of stimulus location in rat somatosensory cortex. *Neuron*, 32:503–514, November 2002.

[56] D. S. Reich, F. Mechler, K. Purpura, and J. D. Victor. Interspike intervals, receptive fields, and information encoding in primary visual cortex. *The Journal or Neuroscience*, 20:1964–1974, 2000.

[57] D. S. Reich, F. Mechler, and J. D. Victor. Temporal coding of contrast in primary visual cortex. *Journal or Neurophysiology*, 85:1039–1050, 2001.

[58] P. Reinagel and R. Reid. Temporal coding of visual information in the thalamus. *J. Neurosci.*, 20(14):5392–5400, 2000.

[59] F. Rieke, D. Warland, R. R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the neural code*. The MIT Press, 1997.

[60] J. C. Roddey, B. Girish, and J. P. Miller. Assessing the performance of neural encoding models in the presence of noise. *Journal of Computational Neuroscience*, 8:95–112, 2000.

[61] K. Rose. Deteministic annealing for clustering, compression, classification, regerssion, and related optimization problems. *Proc. IEEE*, 86(11):2210–2239, 1998.

[62] R. V. Rullen and S. J. Thorpe. Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex. *Neural Computation*, 13:1255–1283, 2001.

[63] E. Salinas and L. F. Abbott. Vector reconstruction from firing rates. *J. Comp. Neurosci*, 1(1-2):89–107, 1994.

[64] E. Schneidman, N. Slonim, N. Tishby, R. R. de Ruyter van Steveninck, and W. Bialek. Analyzing neural codes using the information bottleneck method. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2003. *at press*.

[65] J. R. Schott. *Matrix Analysis for Statistics*. John Wiley and Sons, New York, 1997.

[66] S. R. Schultz and S. Panzeri. Temporal correlations and neural spike train entropy. *Phys. Rev. Lett.*, 86(25):5823–5826, 2001.

[67] M. N. Shadlen and W. Newsome. Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol*, 4:569–579, 1994.

[68] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:623–656, 1948.

[69] N. Slonim. The information bottleneck: Theory and applications. Doctoral Thesis, Hebrew University, 2002.

[70] N. Slonim and N. Tishby. Agglomerative information bottleneck. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 617–623. MIT Press, 2000.

[71] J. Smoller and A. G. Wasserman. Bifurcation and symmetry breaking. *Inventiones mathematicae*, 100:63–95, 1990.

[72] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Let.*, 80(1):197–200, 1998.

[73] F. Theunissen and J. P. Miller. Temporal encoding in nervous systems: A rigorous definition. *J. Comp. Neurosci*, 2:149–162, 1995.

[74] F. Theunissen, J. C. Roddey, S. Stufflebeam, H. Clague, and J. P. Miller. Information theoretic analysis of dynamical encoding by four primary interneurons in the cricket cercal system. *J. Neurophysiol.*, 75:1345–1364, 1996.

[75] F. Theunissen, J. C. Roddey, S. Stufflebeam, H. Clague, and J. P. Miller. Information theoretic analysis of dynamical encoding by four primary sensory interneurons in the cricket cercal system. *J. Neurophy.*, 75:1345–1359, 1996.

[76] F. E. Theunissen and J. P. Miller. Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system

accuracy and optimal tuning curve width of four primary interneurons. *J. Neurophysiol.*, 66:1690–1703, 1991.

[77] S. J. Thorpe, A. Delorme, and R. V. Rullen. Spike based strategies for rapid processing. *Neural Networks*, 14:715–725, 2001.

[78] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. The 37th annual Allerton Conference on Communication, Control, and Computing, 1999.

[79] H. Tuckwell. *Introduction to Theoretical Neurobiology.* Cambridge University Press, 1988.

[80] H. Tuckwell. *Stochastic Processes in the Neurosciences, Philadelphia.* SIAM, 1989.

[81] J. H. van Hateren and H. P. Snippe. Information theoretical evaluation of parametric models of gain control in blowfly photorecoptor cells. *Vision Research*, 41:1851–1865, 2001.

[82] A. Vanderbauwhede. Local bifurcation and symmetry. Habilitation Thesis, Rijksuniversiteit Gent., 1980.

[83] J. D. Victor. How the brain uses time to represent and process visual information. *Brain Research*, 886:33–46, 2000.

[84] J. D. Victor and K. Purpura. Metric-space analysis of spike trains: theory, algorithms, and application. *Network: Computation in Neural Systems*, 8:127–164, 1997.

[85] V. Volterra. *Theory of Functionals and of Integral and Integro-differential Equations.* Blackwell Scientific, London, 1930.

[86] A.-K. Warzecha and M. Egelhaaf. Variability of spike trains during constant and dynamic stimulation. *Science*, 283:1927–1930, March 1999.

[87] N. Wiener. *Nonlinear Problems in Random Theory.* MIT Press, Cambridge, MA, 1958.