



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
© Copyright by Albert Edward Parker III (2003)

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.

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

D378
P223

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

Tomáš Gedeon 04/17/2003
(Signature) Date

Approved for the Department of Mathematics

Kenneth L. Bowers

Kenneth L. Bowers 4/17/03
(Signature) Date

Approved for the College of Graduate Studies

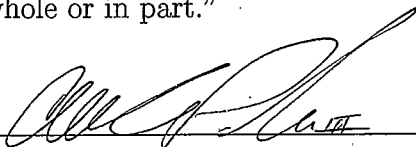
Bruce McLeod

Bruce L. McLeod 4-17-03
(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

4/17/03

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	viii
LIST OF FIGURES	x
1. INTRODUCTION	1
Neural Coding	16
Neural Coding through the Ages	21
Neural Encoding	21
Neural Decoding	29
The Information Distortion	38
Outline of Thesis	40
2. MATHEMATICAL PRELIMINARIES	45
Notation and Definitions	45
Information Theory	50
The Distortion Function $D(q)$	60
The Information Distortion Problem	61
The Information Distortion Measure	62
The Maximal Entropy Problem	64
Derivatives	65
Dealing with Complex Inputs	67
The Function $G(q)$	69
3. THE DYNAMICAL SYSTEM	73
The Optimization Problem	73
The Gradient Flow	80
4. KERNEL OF THE HESSIAN	84
General Form of a Vector in the Kernel	84
Determinant Forms of the Hessian	87
Generic Singularities	97
Singularities of the Information Bottleneck	100
5. GENERAL BIFURCATION THEORY WITH SYMMETRIES	104
Existence Theorems for Bifurcating Branches	108
Bifurcation Structure	114
Derivation of the Liapunov-Schmidt Reduction	126
Equivariance of the Reduction	134

6. SYMMETRY BREAKING BIFURCATION	137
Notation	138
M -uniform Solutions	139
The Group of Symmetries	141
The Group S_M	149
The Initial Solution q_0	152
Kernel of the Hessian at Symmetry Breaking Bifurcation	156
Liapunov-Schmidt Reduction	165
Equivariance of the Reduction	169
Isotropy Subgroups	180
Bifurcating Branches from M -uniform Solutions	195
Bifurcating Branches when $M \leq 4$	204
Bifurcation Structure of M -uniform Solutions	207
The Theory Applied to the Information Bottleneck	221
7. CONTINUATION	224
Parameter Continuation	225
Pseudoarclength Continuation	228
Branch Switching	231
Continuation of the Gradient Flow	232
Numerical Results	236
8. SADDLE-NODE BIFURCATION	247
Kernel of the Hessian at Non-symmetry Breaking Bifurcation	248
Necessary Conditions	252
A Sufficient Condition	253
9. OPTIMIZATION SCHEMES	257
Notation	257
Optimization Theory	258
Unconstrained Line Searches	260
Newton Conjugate Gradient Method	264
Constrained Line Searches	266
Augmented Lagrangian	269
Optimization Schemes	273
Annealing	273
Vertex Search	276
A New Numerical Algorithm	279
Numerical Results	281
Synthetic Data	282
Physiological Data	282
10. CONCLUSION	286
REFERENCES CITED	288

LIST OF TABLES

Table	Page
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 ν_1 are decoded as the stimulus x_2 with certainty .42. Any of the neural responses in class ν_3 are decoded as the stimulus x_3 with certainty .56.....	37
2. Bifurcation Location: Theorem 81 is used to determine the β values where bifurcations can occur from $(q_{\frac{1}{N}}, \beta)$ when $\Delta G(q_{\frac{1}{N}})$ is nonsingular. Using Corollary 115 and Remark 117.1 for the Information Distortion problem (2.33), we predict bifurcation from the branch $(q_{\frac{1}{4}}, \beta)$, at each of the 15 β values given in this table ...	236
3. The bifurcation discriminator: Numerical evaluations of the bifurcation discriminator $\zeta(q_{\frac{1}{N}}, \beta^* \approx 1.038706, \mathbf{u}_k)$ (6.85) as a function of N for the four blob problem (see Figure 1a) when F is defined as in (2.33). We interpret that $\zeta(q_{\frac{1}{2}}, 1.038706, \mathbf{u}_k) = 0$. Thus, further analysis is required to determine whether the bifurcating branches guaranteed by Theorem 114 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 $\zeta(q_{\frac{1}{N}}, \beta^*, \mathbf{u}_k) < 0$, predicting that bifurcating branches from $q_{\frac{1}{N}}$ are subcritical and unstable in these cases (Theorem 131).....	237
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.	283

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..... 285

LIST OF FIGURES

Figure	Page
1. <i>The Four Blob Problem</i> 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 ν 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 $I(X; Y_N)$	8
2. Conceptual bifurcation structure of solutions (q^*, β) to the problem (1.1) as a function of the parameter β . 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 $y \in Y$	9
3. [22, 29] Observed bifurcations of the solutions (q^*, β) 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 β	10
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, $Y = y_1, y_2, y_3, y_4$	18
5. A: Modelling a sensory system as a communication channel. B: The structure, $p(X, Y)$, of an optimal communication system	19

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, τ and t , to make the distinction that the stimuli X may occur during different intervals of time than do the neural responses Y 22
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] 23
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..... 24
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 26
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] 34

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 36

12. A hierarchical diagram showing how the singularities of $\Delta_{q,\lambda}\mathcal{L}$ and ΔF affect the bifurcation structure of equilibria of (3.18) 100

13. Partial lattice of the maximal isotropy subgroups $\langle \gamma^p \rangle < S_M$, from Theorem 101, when $M = 4$ and $p = 2$, and the corresponding basis vectors of the fixed point spaces of the corresponding groups from Theorem thm:gammapisotropy and Remark 113.3 186

14. The lattice of the maximal isotropy subgroups $S_M < S_N$ for $N = 4$ from Lemma 103 and the corresponding basis vectors of the fixed point spaces of the corresponding groups from Lemma 103 190

15. Panel (A) shows the full lattice of subgroups $S_2 < S_3$ for $N = 4$ and the corresponding basis vectors, from Theorem 100 and Lemma 103, 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 103, of the fixed point spaces of the corresponding groups 192

16. 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 $(q_{k+1}, \lambda_{k+1}, \beta_{k+1})$ 226

17. [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 $\zeta(q_{\frac{1}{4}}, 1.038706, \mathbf{u}_k) < 0$. Here, the bifurcation diagram is shown with respect to $\|q^* - q_{\frac{1}{4}}\|$. It is at the saddle node that this 3-uniform branch changes from being a stationary point to a local solution of the problem (2.33) 237

- 18. 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 86. 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 $\Delta \mathcal{L}(q^*)$ 238

- 19. Actual bifurcation structure of M -uniform solutions for (2.33) when $N = 4$. Figure 3 showed an incomplete bifurcation structure for this same scenario. Observe that Figure 18 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)..... 239

- 20. Symmetry breaking bifurcation from the 4-uniform branch $(q_{\frac{1}{N}}, \lambda, 1.038706)$, as in Figure 19, but now we investigate the bottom 2-uniform branch, panels (2)-(5) 239

- 21. Comparison of the observed bifurcation structure from the 4-uniform branch given in Figure 3 (triangles), and the actual bifurcation structure given in Figures 19 and 20 (dots) when $N = 4$ for the Four Blob problem. Qualitatively, the bifurcation structure is the same, except for the shift in β , which we explain in Remark 156 ... 240

- 22. A close up, from Figure 19, 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 23), and an eigenvalue of $\Delta F(q^*)$ becomes positive. As we saw in Figure 18, this positive eigenvalue of $\Delta F(q^*)$ crosses back at the asterisk shown, which does not correspond to a singularity of $\Delta \mathcal{L}(q^*)$ 241

23. Panel (A) shows a close up, from Figure 19, 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 19, 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 242
24. Panel (A) is a log-log plot of 3-uniform branches, some of which are shown in Figure 21, which bifurcate from the $q_{\frac{1}{N}}$ branch at the β 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 $(q_{\frac{1}{N}}, 1.390994)$ 243
25. In panel (A) we show a 3-uniform branch, from Figure 24, which bifurcates from $(q_{\frac{1}{N}}, 4.28766)$ and some of the particular quantizers. Panel (B) shows the 3-uniform solutions, from Figure 24, which bifurcate from $q_{\frac{1}{N}}$ when $\beta \in \{5.413846, 31.12109, 46.29049\}$, and some of the associated quantizers as well 244
26. The bifurcating branches from the 4-uniform solution branch at the values $\beta \in \{1.038706, 1.133929, 1.390994\}$ as predicted by the Smoller-Wasserman Theorem and Theorem 112 when $N = 4$. The isotropy group for all of the solution branches shown is $\langle \gamma_{(1324)}^2 \rangle < \Gamma$. The element $\gamma_{(1324)}$ of order 4 in Γ is represented by the 4-cycle $(1324) \in \mathcal{S}$ (see (6.13)). Thus, γ^2 is represented by the element $(1324)^2 = (12)(34) \in \mathcal{S}$. The group $\langle \gamma_{(1324)}^2 \rangle$ only fixes the quantizers which are "twice" 2-uniform: 2-uniform on the classes $\mathcal{U}_1 = \{1, 2\}$, and 2-uniform on the classes $\mathcal{U}_2 = \{3, 4\}$... 245

27. 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 Δ_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 ν . 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 ν 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 278

28. [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 ν (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)..... 284

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

$$\max_{q \in \Delta} (G(q) + \beta D(q)) \quad (1.1)$$

where $\beta \in [0, \infty)$, Δ is a subset of \mathfrak{R}^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 rigorously 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 ν_i with a certain probability, $q(\nu_i|y_j)$. Observe that we may view q as a vector in some subspace Δ of \mathfrak{R}^{NK} . The subspace Δ is the space of valid discrete conditional probabilities in \mathfrak{R}^{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 ν_i . This optimization problem can be written as

$$\max_{C, q \in \Delta} H(q, C) \quad \text{constrained by} \quad (1.2)$$

$$D(q, C) \leq D_0,$$

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 ν . 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

$$\max_{q \in \Delta} H(q) \quad \text{constrained by} \quad (1.3)$$

$$D_I(q) \leq D_0$$

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

$$\max_{q \in \Delta} H(q) \quad \text{constrained by}$$

$$D_{eff}(q) \geq I_0$$

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

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

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*

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

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

1. Perform β -step: Let $\beta_{k+1} = \beta_k + d_k$ where $d_k > 0$.
2. Take $q_{k+1}^{(0)} = q_k + \eta$, where η is a small perturbation, as an initial guess for the solution q_{k+1} at β_{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, $Y_N = \{1, \dots, N\}$. When $N = 2$ as in panel (b), the optimal clustering q^* yields an incomplete description of the relationship

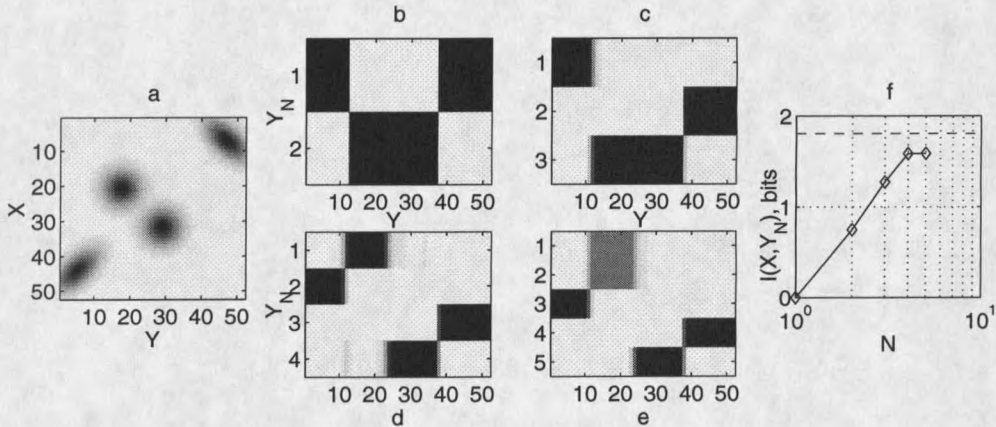


Figure 1. *The Four Blob Problem* from [22, 29]. (a) A joint probability for the relation $p(X, Y)$ between a stimulus set X and a response set Y , each with 52 elements. (b–d) The optimal clusterings $q^*(Y_N|Y)$ for $N = 2, 3$, and 4 classes respectively. These panels represent the conditional probability $q(\nu|y)$ of a class ν 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 $I(X; Y_N)$.

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.

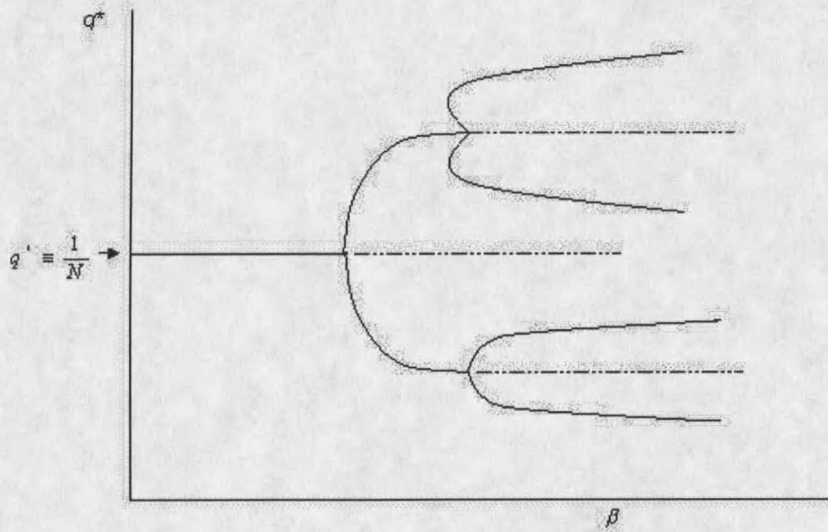


Figure 2. Conceptual bifurcation structure of solutions (q^*, β) to the problem (1.1) as a function of the parameter β . In this instance, the first solution is denoted as $q_{\frac{1}{N}}$, the clustering of the data such that $q(Y_N|Y) = \frac{1}{N}$ for every $\nu \in Y_N$ and every $y \in Y$.

It has been observed that the solutions (q, β) of (1.1), which contain the sequence $\{(q_k, \beta_k)\}$ found in step 3 of Algorithm 1, undergo *bifurcations* or *phase transitions* as $\beta \rightarrow \mathcal{B}$ [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 β can be seen in the top panel. Some of the solutions $\{(q_k, \beta_k)\}$ for different values of β_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 \rightarrow 2, 2 \rightarrow 3) are similar to the ones described in [61] for a different distortion function. One also observes

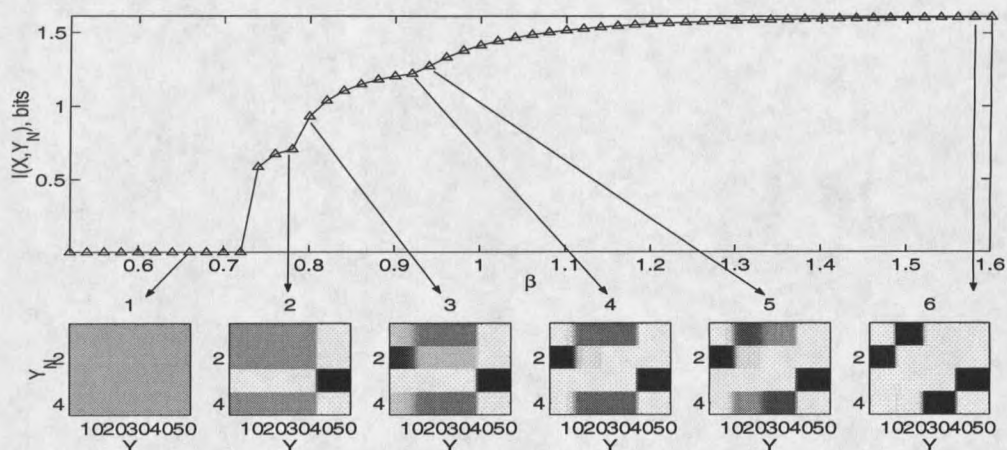


Figure 3. [22, 29] Observed bifurcations of the solutions (q^*, β) to the Information Distortion problem (1.4). For the data set in Figure 1a, the behavior of $D_{eff} = I(X; Y_N)$ (top) and the solutions $q(Y_N|Y)$ (bottom) as a function of β .

transitions ($4 \rightarrow 5$) 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 $q \in \mathfrak{R}^{4K} = \mathfrak{R}^{208}$. 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 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

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

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 $q \in \mathfrak{R}^{NK}$ 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 $Y_N = \{1, \dots, N\}$, 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 74). 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 ν , 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 82), 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"

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

The first solution branch, (q_0, β) , where $q_{\frac{1}{N}}$ 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 115 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 114 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 118).

We have shown that all symmetry breaking bifurcations from S_M to S_{M-1} are pitchfork-like (Theorem 124 and see Figures 17–25). Furthermore, we have ascertained the existence of other types of bifurcating branches from symmetry breaking bifurcation (Theorem 112 and the Smoller-Wasserman Theorem) which we did not expect (see Figure 26).

In fact, we have shown that the observed bifurcation structure given in Figure 3, although qualitatively correct, is "shifted" in β (see Figure 21 and Remark 156).

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 131 and 132). 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 17,

18 and 25). We have also numerically confirmed that subcritical bifurcations occur on other branches as well (Figure 23).

It is a well known fact that subcritical bifurcating branches are unstable (Theorem 131). We have also provided a condition which ascertains the stability of supercritical branches (Theorem 132). We have shown that, in some instances, unstable branches can not contain solutions to (1.9) (Theorem 133). For example, the subcritical bifurcating branches in Figure 17 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 17, 18, 23, and 25). 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

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

where

$$F(q, \beta) = G(q) + \beta D(q), \quad (1.10)$$

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 Δ is the linear constraint space of valid conditional probabilities,

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

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 ν 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 Δ .

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.

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

$$X : \Omega_X \rightarrow \mathcal{X}, \quad (1.12)$$

and Y is a random variable of outputs

$$Y : \Omega_Y \rightarrow \mathcal{Y} \quad (1.13)$$

(see Figure 5).

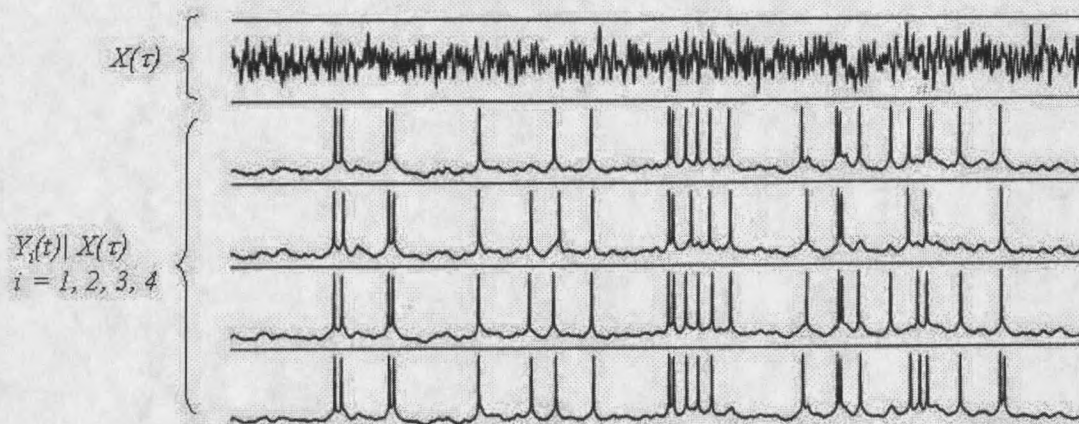
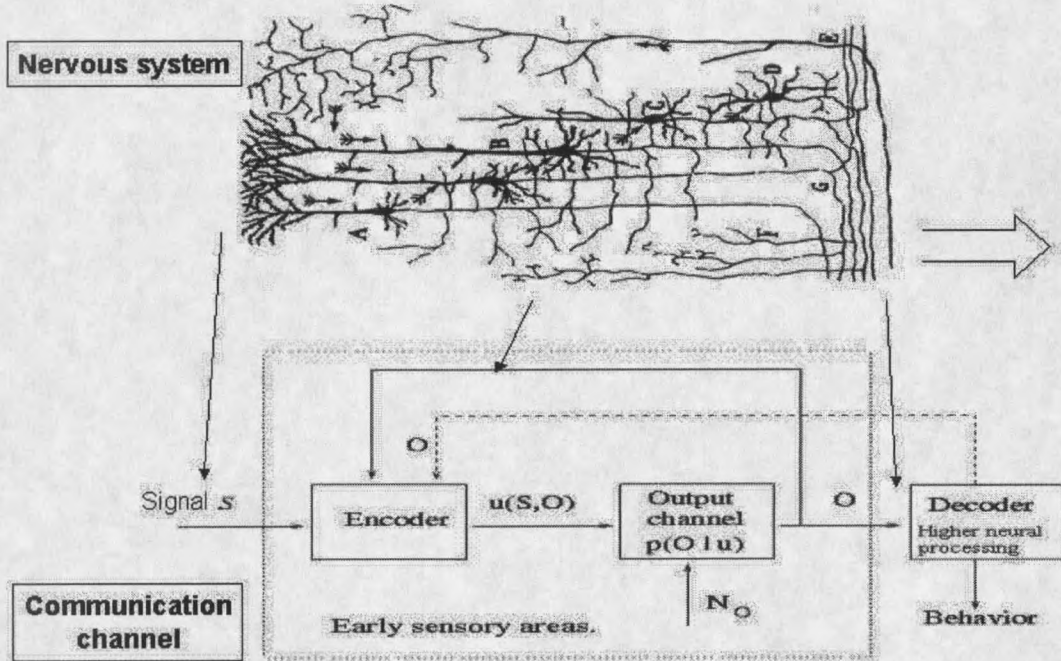


Figure 4. The neural response to a static stimulus is stochastic. Presenting an identical stimulus, $X(\tau) = x$, four separate times to a biological sensory system produces four distinct neural responses, $Y = y_1, y_2, y_3, y_4$.

When translating the structure of an information channel to neural systems, the output space Ω_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 Ω_Y as the finite dimensional measurable space $\mathcal{Y} = \{0, 1\}^k$. Thus, each neural response is

Analysis Framework:



B

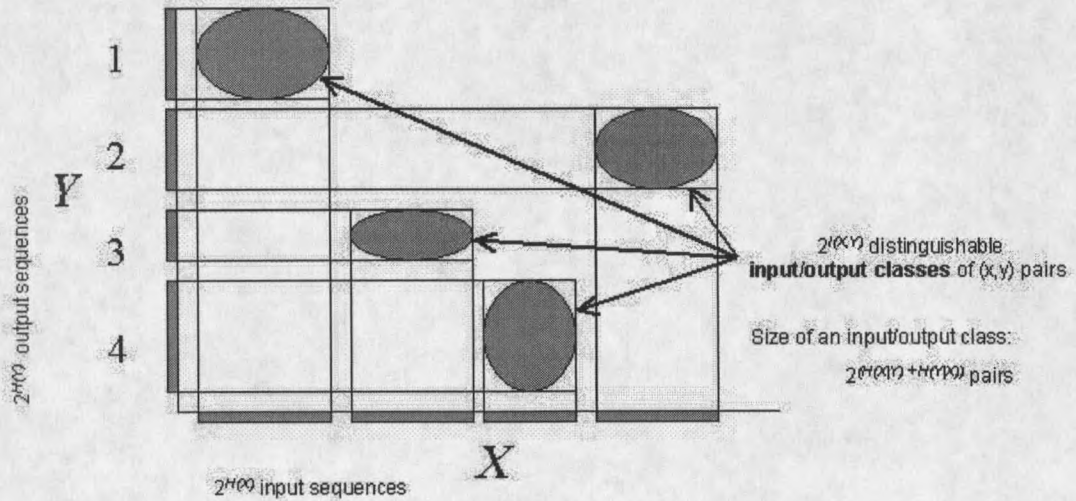


Figure 5. A: Modelling a sensory system as a communication channel. B: The structure, $p(X, Y)$, of an optimal communication system.

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 $k = 100$. 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

$$\tilde{Y} : \mathcal{Y} \rightarrow \tilde{\mathcal{Y}}, \quad (1.14)$$

where $\tilde{\mathcal{Y}}$ is the space of real numbers \mathfrak{R} . \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 – 500ms in the neurophysiological literature [59, 67].

The input space Ω_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 $\mathcal{X} = \mathfrak{R}^K$, where \mathfrak{R}^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 $\mathcal{X} = \mathfrak{R}^{km} = (\mathfrak{R}^m)^k$. In this

context, X can be written as $X = W^k$ where W is a random variable

$$W : \Omega_X \rightarrow \mathbb{R}^m,$$

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

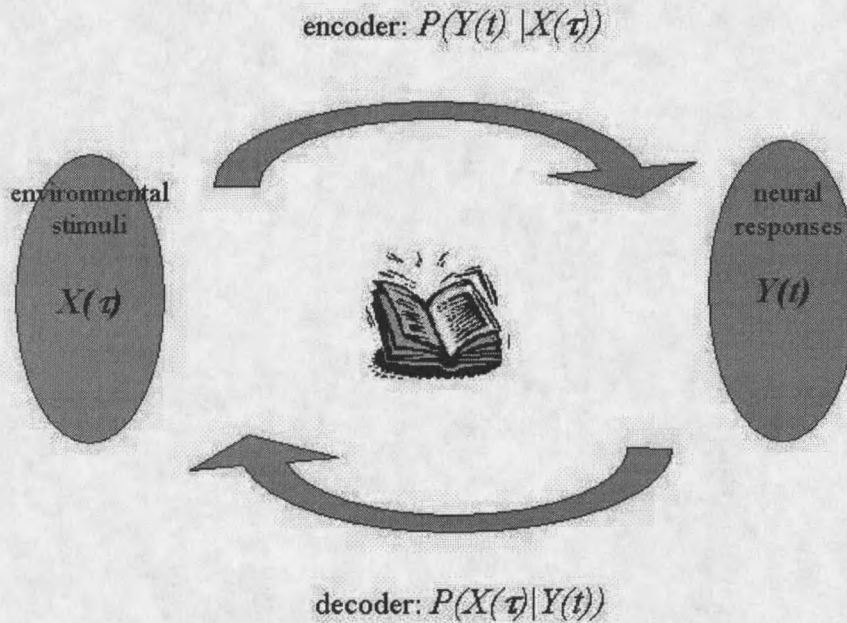


Figure 6. Probability framework, showing the spaces produced by $X(\tau)$ and $Y(t)$, and the stochastic mappings $p(Y|X)$ and $p(X|Y)$ between them. Discovering either of these mappings defines a dictionary between classes of stimuli and classes of responses, where the classes are defined by $p(X, Y)$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

different masses from a muscle, and measured the activity of a stretch receptor neuron 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.

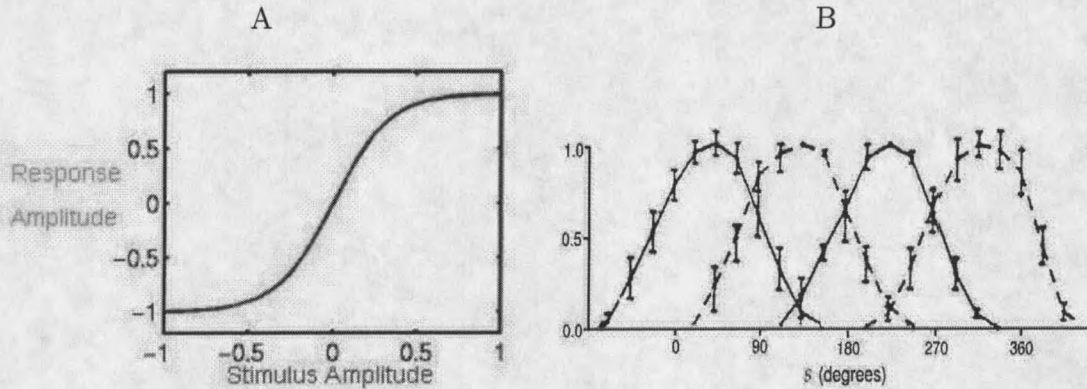


Figure 7. A: The response tuning curve. In *spike count* or *rate* coding, the response amplitude is \tilde{Y} , which we define as the number of spikes present in some time window. The stimulus amplitude is represented by some scalar. B: The Directional Tuning Curve. Another example of spike count coding. 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].

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

$$p(\tilde{Y} = \tilde{y}) = \sum_x p(\tilde{y}|x)p(x).$$

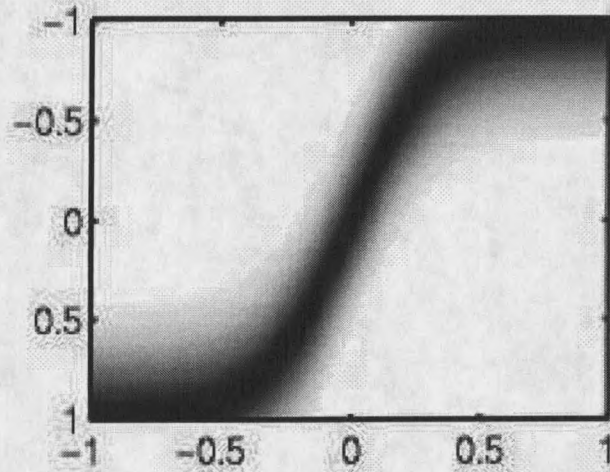


Figure 8. An estimate of the encoder $p(\tilde{Y}|X)$, using spike count coding, by repeating each stimulus $x \in \mathcal{X}$ many times, creating a histogram for each $\tilde{y}|X$, and then normalizing.

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, Ω_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 λ . This model presupposes that the spikes are independent from each other given a stimulus $X = x$. Determining λ 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].

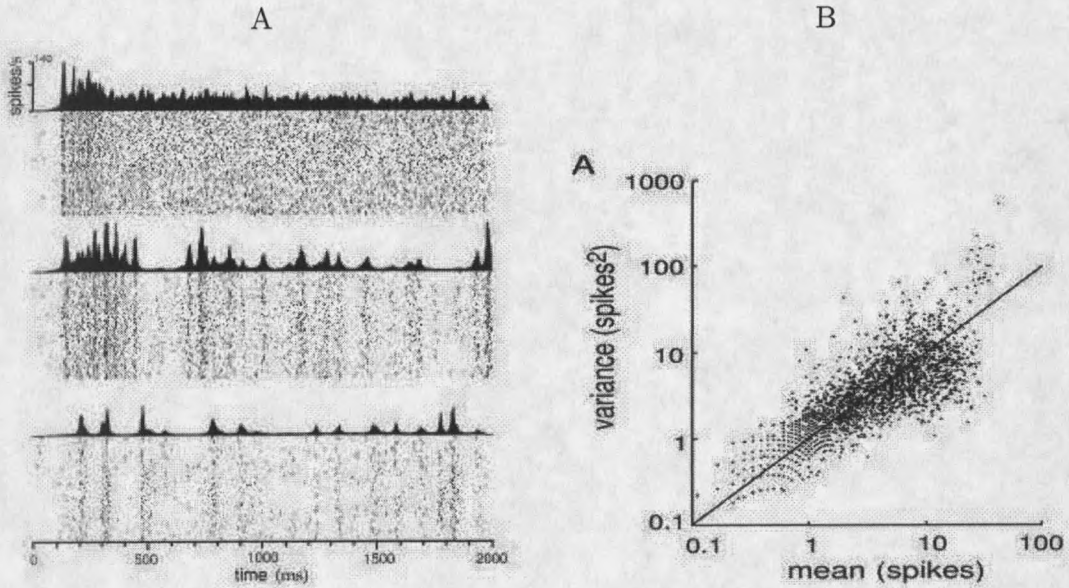


Figure 9. Both panels are from [1]. A: Examples of a peristimulus time histogram for three different stimuli x_1, x_2, x_3 , not shown. Below each PSTH is the raster plot of associated neural responses $Y|x_i$ over many repetitions of the stimulus $X = x_i$. The PSTH is the normalized histogram of the 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.

The Poisson rate is

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

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

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

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

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

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 λ 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

$$Y(t) = f_0 + \int f_1(\tau_1)X(t - \tau_1)d\tau_1 + \int \int f_2(\tau_1, \tau_2)X(t - \tau_1)X(t - \tau_2)d\tau_1d\tau_2 + \dots \quad (1.16)$$

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]

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

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 ω 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

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

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 $\text{Cov}(X|Y)$ 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,*

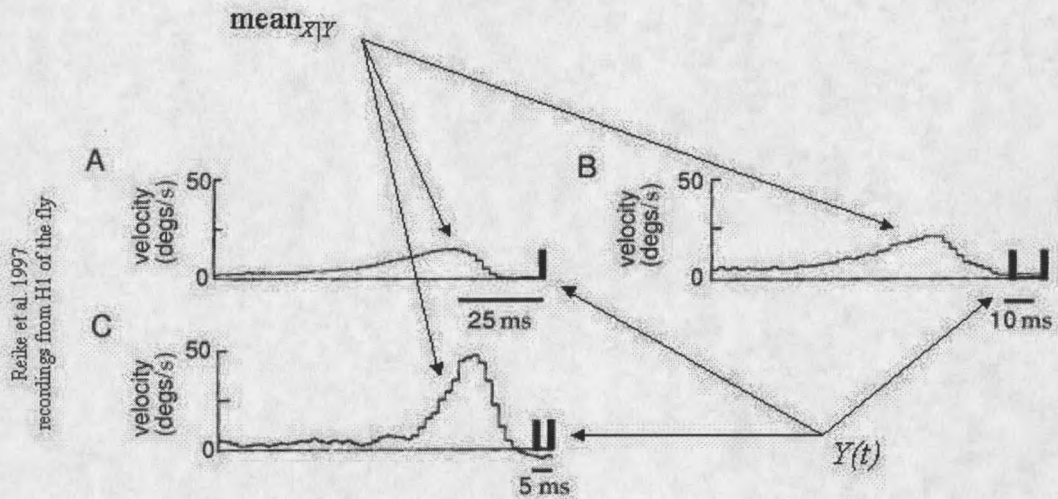


Figure 10. Estimating $p(X|Y)$ with a Gaussian. Examples of three spike trains recorded from the H1 neuron of the blowfly and the corresponding conditional means of the stimuli (velocity of a pattern) which elicited each of these responses. These conditional means, as well as conditional variances, are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

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 $\text{Cov}(X|Y = y)$ 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 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

$$D[r](Y_i, Y_j), \quad (1.19)$$

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 Δt 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

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

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

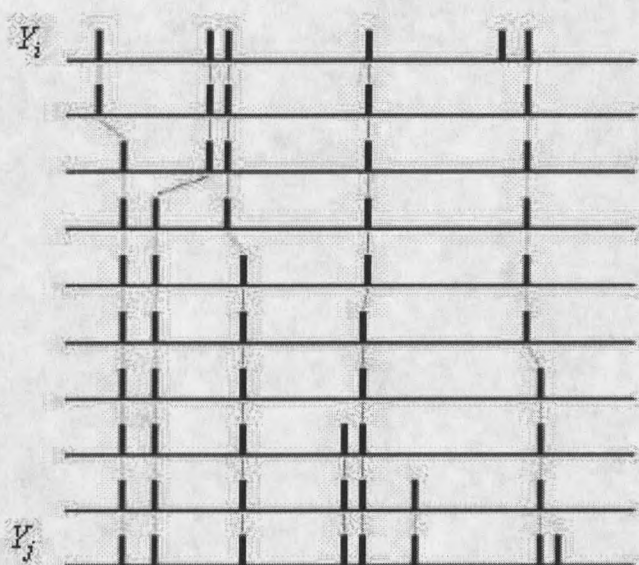


Figure 11. Computing the Spike Train Metric [84]. One path of elementary steps used to transform a spike train Y_i into a spike train Y_j .

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, \dots, 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, \dots, x_C , presented multiple times each, all

A						B					
ν_1	ν_2	ν_3	ν_4	ν_5		ν_1	ν_2	ν_3	ν_4	ν_5	
3	11	3	2	1	x_1	.25	.44	.11	.17	.043	x_1
5	10	3	2	0	x_2	.42	.40	.11	.17	0	x_2
1	1	15	1	2	x_3	.08	.04	.56	.08	.08	x_3
1	0	4	2	13	x_4	.08	0	.15	.17	.54	x_4
2	3	2	5	8	x_5	.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 C gives the decoder $p(X|\nu_i)$. In this example, any of the neural responses which belong to ν_1 are decoded as the stimulus x_2 with certainty .42. Any of the neural responses in class ν_3 are decoded as the stimulus x_3 with certainty .56.

of which elicit a total of K neural responses y_1, y_2, \dots, y_K . Initialize C , the $C \times C$ classification matrix, to zeros, and let $\nu_1, \nu_2, \dots, \nu_C$ be C abstract response classes. Start the algorithm with $i = 1$.

1. Suppose that y_i was elicited by x_α . Assign y_i to response class ν_β if

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

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

2. Increment the component $[C]_{\alpha\beta}$ of the matrix C by 1.
3. Repeat step 1 and 2 for $i = 2, \dots, K$

One normalizes the columns of the matrix 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 ν in the normalized matrix 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

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

The random variables

$$X \rightarrow Y \rightarrow Y_N$$

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* $q(Y_N|Y)$

$$q(Y_N|Y) : \mathcal{Y} \rightarrow \mathcal{Y}_N. \quad (1.20)$$

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 (2.25)). 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

$$\mathcal{L}(q, \lambda, \beta) : \mathfrak{R}^{NK} \times \mathfrak{R}^K \times \mathfrak{R} \rightarrow \mathfrak{R}. \quad (1.21)$$

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

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

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 β . The Jacobian of this system is the Hessian $\Delta_{q,\lambda} \mathcal{L}(q, \lambda, \beta)$. Equilibria, (q^*, λ^*) , 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^*, λ^*, β) 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 $\ker \Delta_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta^*)$ (Theorems 112 and 114). More surprisingly perhaps is that the block diagonal Hessian ΔF

(Claim 72) 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 ΔF 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 114). Such symmetry breaking bifurcations are always pitchfork-like (Theorem 124). The Smoller-Wasserman Theorem is employed to ascertain the existence of bifurcating solutions in higher dimensional subspaces of the kernel (Theorem 112). 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 ΔF

yields bifurcating branches when G is strictly concave (Corollary 111), which is the case for the Information Distortion problem (1.4). We also provide conditions which determine the location (Theorem 81), type (Theorem 124), orientation (i.e. supercritical or subcritical), and stability (Theorems 131 and 132) of bifurcating branches from certain solutions to (1.9). In some instances, unstable branches can not contain solutions to (1.9) (Theorem 133).

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 (2.33). There are two types of bifurcations which we observe numerically: symmetry breaking bifurcations and saddle-node bifurcations. See Figures 17–25 and 26.

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 139 and 145).

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 153) 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 Δ (Algorithm 159). 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 28).

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 161) 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 $\mathcal{Y} := \{y_1, y_2, \dots, y_K\}$.

$K := |\mathcal{Y}| < \infty$, 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* $\mathcal{Y}_N := \{1, 2, \dots, N\}$.

$N := |\mathcal{Y}_N|$, 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

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

$q^\nu := q(\nu|Y)$, 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

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

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

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

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

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

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

$\text{peigenspace}(A) :=$ 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 $[C]_{ij} = A \otimes B = a_{ij}B$.

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

$\langle \mathbf{v}, \mathbf{w} \rangle := \langle \mathbf{v}, \mathbf{w} \rangle_I = \sum_{i,j} [\mathbf{v}]_i [\mathbf{w}]_j$, the Euclidean inner product.

$\|\mathbf{v}\| := \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$, the Euclidean norm.

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

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

$e_i := i^{\text{th}}$ column of the identity I .

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

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

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

$\partial_{\mathbf{x}}^2 \psi :=$ three dimensional array of second derivatives of the vector valued function ψ with respect to the vector \mathbf{x}

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

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

$$\sum_{i,j,k} \frac{\partial^3 \psi}{\partial [\mathbf{x}]_i \partial [\mathbf{x}]_j \partial [\mathbf{x}]_k}(\mathbf{x}_0)[\mathbf{u}]_i [\mathbf{v}]_j [\mathbf{w}]_k,$$

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

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

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

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

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

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

$\mathcal{O}(\mathbf{x}^m)$:= "big oh" of $\|\mathbf{x}\|^m$. By definition, if $f(\mathbf{x}) = \mathcal{O}(\mathbf{x}^m)$, then there exists $n > 0$ such that $\|f(\mathbf{x})\| \leq n\|\mathbf{x}\|^m$ if $\|\mathbf{x}\|$ 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 $\mathbf{x}^T A \mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and is *negative definite* if $\mathbf{x}^T A \mathbf{x} < 0$ for all $\mathbf{x} \in \mathbb{R}^n$. The symmetric matrix A is *non-positive definite* if $\mathbf{x}^T A \mathbf{x} \leq 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and is *non-negative definite* if $\mathbf{x}^T A \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.

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

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

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

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

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 Ω . \mathcal{O} and \mathcal{B} are the respective σ -algebras. A source

X is a mathematical model for a physical system that produces a succession of symbols $\{X_1, X_2, \dots, X_n\}$ 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, \dots, X_m) and (X_k, \dots, 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 $\varphi : \Omega \rightarrow \Omega$ is *measure preserving* if $p(\varphi^{-1}A) = p(A)$ for all $A \in \mathcal{O}$. A set $A \in \mathcal{O}$ is φ -*invariant* if $\varphi^{-1}A = A$. Let $\mathcal{I} = \{A | A \text{ is } \varphi\text{-invariant}\}$. The measurable transformation φ 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 φ is ergodic.

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

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

and an output

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

where the spaces and σ -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

$$H(Y|X) = -E_{X,Y} \log p(Y|X)$$

$$H(X, Y) = -E_{X,Y} \log p(X, Y).$$

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 $\{Y_1, Y_2, \dots, Y_n\}$ 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*

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

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

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

almost surely. □

In many instances, as in the case of physiological recordings from a biological sensory system, the data $\{Y_1, Y_2, \dots, Y_n\}$ 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 φ is a measure preserving transformation on (Ω, \mathcal{O}) and Y is a source with $E(Y) < \infty$.

Then

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

REMARK 7. If φ 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

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

where $H \equiv \lim_{n \rightarrow \infty} E(-\log p(Y_n | Y_{n-1}, \dots, Y_0))$ 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 rigorously in the following way. Each element of the output space \mathcal{Y} can be modelled as a sequence of symbols of a random variable

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

where Ω_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_ϵ^k with respect to probability density $p(Z)$ on Z is the set of sequences $(z_1, z_2, \dots, z_k) \in Z^k$ for which

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

$(z_1, z_2, \dots, z_n) \in A_\epsilon^k$ 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 $(z_1, z_2, \dots, z_k) \in A_\epsilon^k$ then $H(Z) - \epsilon \leq -\frac{1}{k} \log p(z_1, z_2, \dots, z_k) \leq H(Z) + \epsilon$*
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_ϵ^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

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

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

DEFINITION 12. *The set A_ϵ^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*

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

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 ,

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

Thus, a jointly typical set A_ϵ^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

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

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,

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

We will use Greek letters such as ν, δ, ω, μ and η 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

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

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 ν and k . The *uniform quantizer*, which we denote by $q_{\frac{1}{N}}$, is the special case when

$$q_{\frac{1}{N}}(\nu|y) = \frac{1}{N} \quad (2.7)$$

for every ν and k . The constraint space Δ 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 \mathfrak{R}^{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

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

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

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

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 $\mathcal{Y}, \mathcal{Y}_N \subset \mathfrak{R}^{NK}$. 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, ν)

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

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

$$\begin{aligned}
 D_I &= \sum_{y \in \mathcal{Y}, \nu \in \mathcal{Y}_N} p(y, \nu) KL(p(X|y) \| p(X|\nu)) \\
 &= \sum_{y, \nu} p(y, \nu) \sum_x p(x|y) \log \frac{p(x|y)}{p(x|\nu)} \\
 &= \sum_{x, y, \nu} p(x, y, \nu) (\log p(x|y) - \log p(x|\nu)) \tag{2.12}
 \end{aligned}$$

$$\begin{aligned}
 &= \sum_{x, y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} - \sum_{x, \nu} p(x, \nu) \log \frac{p(x, \nu)}{p(x)p(\nu)} \tag{2.13} \\
 &= I(X; Y) - I(X; Y_N)
 \end{aligned}$$

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 Δ [61, 22]. This is not surprising since D_{eff} is convex and Δ 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],

$$\max_{q \in \Delta} H(Y_N|Y) \quad \text{constrained by} \quad (2.16)$$

$$D_{eff}(q) \geq I_0$$

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

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

and

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

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]

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

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

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

The the three dimensional array of third derivatives is

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

The four dimensional array of fourth derivatives is

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

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]

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

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

$$\frac{\delta_{\nu\eta\delta}}{\ln 2} \left(\frac{p(y_k)p(y_l)p(y_m)}{(\sum_k q_{\nu k} p(y_k))^2} - \sum_i \frac{p(x_i, y_k) p(x_i, y_l) p(x_i, y_m)}{(\sum_k q_{\nu k} p(x_i, y_k))^2} \right). \quad (2.23)$$

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

$$\frac{2\delta_{\nu\eta\delta\mu}}{\ln 2} \left(\sum_i \frac{p(x_i, y_k) p(x_i, y_l) p(x_i, y_m) p(x_i, y_p)}{(\sum_k q_{\nu k} p(x_i, y_k))^3} - \frac{p(y_k)p(y_l)p(y_m)p(y_p)}{(\sum_k q_{\nu k} p(y_k))^3} \right). \quad (2.24)$$

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:

$$p(X|Y_N = \nu) = N(x_\nu, C_{X|\nu}). \quad (2.25)$$

The class conditioned stimulus mean x_ν 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 $H(X|Y_N)$:

$$\tilde{H}(X|Y_N = \nu) = \sum_{\nu} p(\nu) \frac{1}{2} \log(2\pi e)^{|X|} \det \left[\sum_y p(y|\nu) (C_{X|y} + x_y^2) - \left(\sum_y p(y|\nu) x_y \right)^2 \right]. \quad (2.26)$$

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

$$\tilde{D}_{eff}(q(Y_N|Y)) := H(X) - \tilde{H}(X|Y_N) \quad (2.27)$$

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

$$\begin{aligned} \max_{q \in \Delta} H(Y_N|Y) & \quad \text{constrained by} & (2.28) \\ \tilde{D}_{eff}(q(\nu|y)) & \geq I_0 & \text{and} \\ \sum_{\nu \in \mathcal{Y}_N} q(\nu|y) & = 1 \quad \text{and} \quad q(\nu|y) \geq 0 \quad \forall y \in Y. \end{aligned}$$

It is not immediately obvious that solutions to (2.28) 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 (2.28), the optimal quantizer $q^*(Y_N|Y)$ will be generically deterministic (Theorems 157 and 158). This means that \tilde{D}_{eff} can be used in place of D_{eff} in the problem (2.33).

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)):

$$\begin{aligned} \min_{q \in \Delta} I(Y; Y_N) \\ D(Y; Y_N) \leq D_0 \end{aligned} \quad (2.29)$$

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))

$$\begin{aligned} \max_{q \in \Delta} H(Y_N|Y) \\ D(Y; Y_N) \leq D_0 \end{aligned} \quad (2.30)$$

The justification for using (2.30) is Jayne's maximum entropy principle [36] (see Remark 3). The formulations (2.29) and (2.30) 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 (2.30)

$$\begin{aligned} \max_{q \in \Delta} H(Y_N|Y) \\ D_{eff}(q) \geq I_0 \end{aligned} \quad (2.31)$$

which uses the nonlinear effective information distortion measure D_{eff} . Tishby et. al. [78, 70] pose an information rate distortion problem as in (2.29)

$$\begin{aligned} \min_{q \in \Delta} I(Y; Y_N) \\ D_{eff}(q) \geq I_0 \end{aligned} \quad (2.32)$$

Using the method of Lagrange multipliers, the rate distortion problems (2.29), (2.30), (2.31), (2.32) can be reformulated as finding the maxima of

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

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 (2.31),

$$F(q, \beta) = H(Y_N|Y) + \beta D_{eff}(q) \quad (2.33)$$

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 (2.32),

$$F(q, \beta) = -I(Y; Y_N) + \beta D_{eff}(q) \quad (2.34)$$

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

We now compare the two formulations (2.31) and (2.33). In [22, 29, 61], one explicitly considers (2.33) for $\beta = \infty$. This involves taking

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

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 Δ , has a maximal value I^* . Therefore, for values of the parameter $I_0 > I^*$ problem (2.31) has no solution. On the other hand, problem (2.33) has a solution for all values of β , since F is a continuous function on a compact set Δ . We have the following result

LEMMA 14. [29] *Let q^* be a solution of (2.31) with $I_0 = I^*$. Let $q(\beta)$ be a solution of problem (2.33) as a function of the annealing parameter β . Then*

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

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

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

The maximum of D_{eff} on Δ is I^* . □

In the Information Bottleneck method, one may only be interested in solutions to (2.34) 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

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

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 Δ , which will help us to understand the bifurcation structure of solutions to (1.9) (see Remarks 19 and 28).

Consider the optimization problem

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

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

$$F(q, \beta) = G(q) + \beta D(q) \quad (3.2)$$

as in (1.9) and (1.10), and

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

(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 ν 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 β on the interior of Δ .

Assumption 15 holds for the Information Distortion and the Information Bottleneck cost functions (2.33) and (2.34). We prove this claim in the former case in Theorem

74.

We rewrite (3.1) using its Lagrangian

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

where the scalar λ_k is the Lagrange multiplier for the constraint $\sum_{\nu=1}^N q_{\nu k} - 1 = 0$, and λ 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

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

where

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

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

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

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

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

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

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

where $\mathbf{0}$ is $K \times K$. The $NK \times NK$ matrix ΔF is the block diagonal Hessian of F ,

$$\Delta F = \begin{pmatrix} B_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & B_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & B_N \end{pmatrix}, \quad (3.9)$$

where $\mathbf{0}$ and B_i are $K \times K$ matrices for $i = 1, \dots, N$. The fact that ΔF is block diagonal follows from the symmetry of F (Claim 72). The proof of this claim is delayed until we can define the explicit group which acts on F in chapter 6.

There are optimization schemes, such as the implicit solution (see (9.20)) and projected Augmented Lagrangian methods (Algorithm 153), 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 Ω 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, λ^* , with components $\lambda_i, i \in \mathcal{E} \cup \mathcal{I}$ such that

$$\nabla_x f(x^*) = - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i \nabla_x c_i(x^*)$$

$$c_i(x^*) = 0, \text{ for all } i \in \mathcal{E}$$

$$c_i(x^*) \geq 0, \text{ for all } i \in \mathcal{I}$$

$$\lambda^* \geq 0, \text{ for all } i \in \mathcal{I}$$

$$\lambda^* c_i(x^*) = 0, \text{ for all } i \in \mathcal{E} \cup \mathcal{I} \quad (3.10)$$

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

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

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

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

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 β . Then there exists a vector of Lagrange multipliers, $\lambda^* \in \mathfrak{R}^K$, 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 $f(\mathbf{x})$ is a point where $\nabla_{\mathbf{x}} f(\mathbf{x}^*) = \mathbf{0}$. A stationary point of a constrained system such as (3.1) is a point where $\nabla_{q, \lambda} \mathcal{L} = \mathbf{0}$. 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

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

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

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

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

If (q, β) is a stationary point of (1.9) in the interior of Δ , 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 $\{c_i\}_{i \in \mathcal{I}} = \{q_{\nu k}\}_{\nu \in \mathcal{N}, y_k \in \mathcal{Y}}$, then for the vector of Lagrange multipliers ξ from (3.13), $\xi_{\nu k} = 0$ for every ν and k . Thus,

$$\nabla_{q, \lambda} \hat{\mathcal{L}} = \nabla_{q, \lambda} \mathcal{L} = \mathbf{0} \quad (3.14)$$

by Theorem 16, which shows that a stationary point to (1.9) in the interior of Δ 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 $q \in \mathbb{R}^{NK}$ 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 $q^* \in \Delta_\varepsilon$ there is a vector of Lagrange multipliers λ^* such that the KKT conditions (Theorem 16) are satisfied. If*

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

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

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

for all $\mathbf{w} \in \ker J$.

Hence, to find a local solution of (3.1) for some β , 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 $\mathbf{w} \in \ker J$ can be written as $Z\mathbf{u}$ for some $\mathbf{u} \in \mathbb{R}^d$.*

The condition

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

can be restated as

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

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

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

where \mathbf{x} is in some Banach space B_2 and $\beta \in \mathfrak{R}$, so that

$$\psi : B_2 \times \mathfrak{R} \rightarrow B_0 \tag{3.16}$$

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

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

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 ψ is $\partial_{\mathbf{x}} \psi$. Let $n(\beta)$ be the number of \mathbf{x} 's for which (\mathbf{x}, β) is a solution of (3.17).

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

REMARK 23. This definition of bifurcation, as used in [33], may seem too restrictive. However, the class of systems we study are gradient systems, $\psi = \nabla_{\mathbf{x}} f$ (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, $\partial_{\mathbf{x}}\psi = \Delta_{\mathbf{x}}f$, 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 (\mathbf{x}^*, β^*) is a bifurcation of (3.17) then $\partial_{\mathbf{x}}\psi(\mathbf{x}^*, \beta^*)$ is singular.

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

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

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

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^*, λ^*) of (3.18) is stable if $\Delta_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta)$ is negative definite. The equilibrium (q^*, λ^*) 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^*, β) of (3.18) where $\Delta F(q^*, \beta)$ is negative definite on $\ker J$ are local solutions of (3.1). Conversely local solutions (q^*, β) 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 260×260 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)

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

There are no equilibria of this system for any β 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

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

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

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

if (3.10) holds, which shows that (q^*, λ^*, β) 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 ΔF . We describe the three types of generic singularities of $\Delta_{q,\lambda}\mathcal{L}$ and ΔF which can occur, and we also provide an overview of how the singularities of $\Delta_{q,\lambda}\mathcal{L}$ and ΔF 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 ΔF for the Information Bottleneck problem (2.34).

General Form of a Vector in the Kernel

Consider an element $\mathbf{k} \in \ker \Delta_{q,\lambda}\mathcal{L}(q^*, \lambda^*, \beta^*)$. 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 \mathbf{k} as

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

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

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

$$\implies \Delta F(q^*, \beta)\mathbf{k}_F = -J^T\mathbf{k}_J \quad (4.3)$$

$$J\mathbf{k}_F = \mathbf{0} \quad (4.4)$$

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

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

We set

$$\mathbf{k}_F = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{pmatrix}, \quad (4.6)$$

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

$$\begin{pmatrix} B_1\mathbf{x}_1 \\ B_2\mathbf{x}_2 \\ \vdots \\ B_N\mathbf{x}_N \end{pmatrix} = - \begin{pmatrix} \mathbf{k}_J \\ \mathbf{k}_J \\ \vdots \\ \mathbf{k}_J \end{pmatrix}. \quad (4.7)$$

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

$$\sum_{\nu} \mathbf{x}_{\nu} = \mathbf{0}. \quad (4.8)$$

THEOREM 29. *Let (q^*, β^*) be a local solution to (3.1) such that $\Delta F(q^*, \beta^*)$ is negative definite on $\ker J$, and let λ^* 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 $\mathbf{w} \in \ker J$ which implies $Z\mathbf{u} = \mathbf{w}$ for some $\mathbf{u} \in \mathfrak{R}^d$. Thus

$$\mathbf{w}^T \Delta F \mathbf{w} = \mathbf{u}^T Z^T \Delta F Z \mathbf{u} < 0 \text{ for every nontrivial } \mathbf{u} \in \mathfrak{R}^d \quad (4.9)$$

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

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

By (4.9), $\mathbf{k}_F = \mathbf{0}$. Substituting this into (4.3) shows that $J^T \mathbf{k}_J = \mathbf{0}$, and so $\mathbf{k}_J = \mathbf{0}$ since J^T has full column rank (by (3.7)). Therefore $\ker \Delta_{q,\lambda} \mathcal{L} = \{\mathbf{0}\}$ 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 ΔF is negative definite, then (4.7) holds if and only if*

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

It follows from (4.8) that $(\sum_{\nu} B_{\nu}^{-1})\mathbf{k}_J = \mathbf{0}$, which has $\mathbf{k}_J = \mathbf{0}$ 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 ΔF . 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 ΔF 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 ΔF is nonsingular with blocks $\{B_i\}_{i=1}^N$, then*

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

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

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

This and the fact that $J = (I_K \ I_K \ \dots \ I_K)$ (see (3.7)) prove the theorem. \square

The following theorem is more general since it does not require the condition that ΔF 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 $(-1)^{(2N-1)K^2} = (-1)^K$. \square

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 73). 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
$$\begin{pmatrix} \frac{m}{m+1} & \frac{-1}{m+1} & \dots & \frac{-1}{m+1} \\ \frac{-1}{m+1} & \frac{m}{m+1} & \dots & \frac{-1}{m+1} \\ \frac{m+1}{m+1} & \frac{m+1}{m+1} & \dots & \frac{m+1}{m+1} \\ \vdots & \vdots & & \vdots \\ \frac{-1}{m+1} & \frac{-1}{m+1} & \dots & \frac{m}{m+1} \end{pmatrix}.$$

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 ΔF 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

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

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

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

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. \square

When ΔF 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 ΔF has M identical blocks, B , which are nonsingular, and $N - M$ other blocks, $\{R_i\}_{i=1}^{N-M}$, then $|\Delta_{g,\lambda} \mathcal{L}|$ is equal to*

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

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

$$(-1)^K \det \left(\begin{array}{c} \left(\begin{array}{cccc} (R_1 + B) & B & \dots & B \\ B & (R_2 + B) & \dots & B \\ B & B & \dots & B \\ \vdots & \vdots & \dots & \vdots \\ B & B & \dots & (R_{N-M} + B) \end{array} \right) & \mathbf{1} \otimes B \\ \mathbf{1}^T \otimes B & T \otimes B \end{array} \right), \quad (4.11)$$

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

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

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

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

which gives the desired result. \square

If ΔF is nonsingular, then its identical blocks must be nonsingular. Thus, Theorem 36 shows that if ΔF 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 139). We now prove a slightly different version of Theorem 36.

COROLLARY 37. *Let (q^*, β^*) 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, β) . Suppose that there exists an $m > 0$ such that $|\det(\mathcal{M}(q, \beta))| < m$ for all (q, β) in some neighborhood about (q^*, β^*) . Then*

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

for all (q, β) about (q^*, β^*) .

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

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

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, β) in a neighborhood of (q^*, β^*) , and we can dispense with the assumption in Theorem 36 that B is nonsingular. \square

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 82 in chapter 6.

LEMMA 38. *Suppose that there exists $1 < M < N$ such that ΔF 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 $\mathbf{u} \in \ker \mathcal{M}$ and decompose it as

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

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

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

Thus,

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

from which it follows that

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

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

$$\left(\sum_j R_j^{-1} B + MI_K \right) \mathbf{v} = \mathbf{0}.$$

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. \square

The converse of this lemma holds as well, which we will prove in chapter 8 (Theorem 139). For now, we state the result.

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

Generic Singularities

In this chapter, we have considered the case where ΔF 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 ΔF *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} be the set of M unresolved classes

and let

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

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

Let B_ν be the block of ΔF corresponding to class ν . 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 ν^{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 $(q, \beta) \in \Delta \times \mathfrak{R}$. 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 $(q, \beta) \in \Delta \times \mathfrak{R}$.

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 ΔF are singular. A generic assumption in this instance is that the $N - M$ resolved blocks, $\{R_\nu\}$, are nonsingular at (q^*, β) . By Corollary 90, $\Delta_{q, \lambda} \mathcal{L}$ must be singular. Conversely, suppose that $\Delta_{q, \lambda} \mathcal{L}$ is singular. Generically, the resolved blocks of ΔF are nonsingular, and

$B \sum_{\nu} R_{\nu}^{-1} + MI_K$ is nonsingular. Then Corollary 90 shows that ΔF is singular. We will see in chapter 6 that this is the type of singularity that exhibits symmetry breaking bifurcation (Theorems 114 and 112).

The second type of singularity is a special case in which no bifurcation occurs. If only a single block, R_{ν} , of ΔF 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 118) 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 ΔF 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 142), we prove that ΔF is generically nonsingular at any bifurcation that is not a symmetry breaking bifurcation, which includes saddle-node bifurcations. Observe that if ΔF 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 ΔF 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 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 ΔF 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

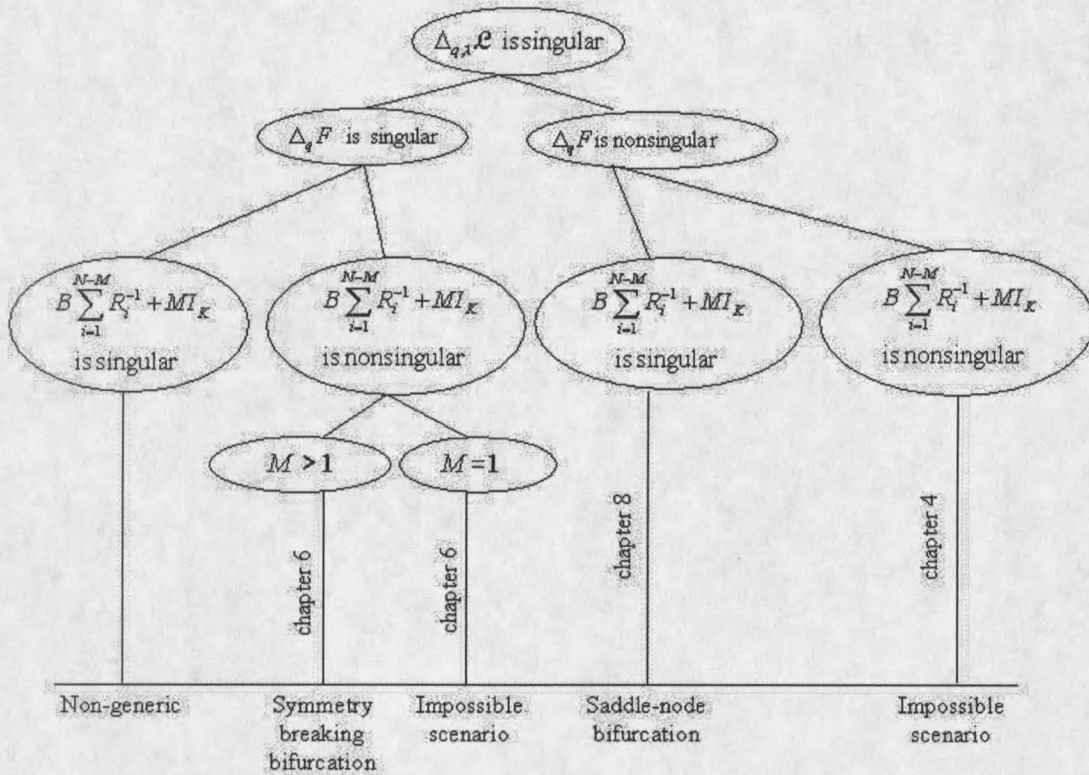


Figure 12. A hierarchical diagram showing how the singularities of $\Delta_{q,\lambda}\mathcal{L}$ and ΔF affect the bifurcation structure of equilibria of (3.18).

diagram, we have indicated the type of bifurcation possible given the assumptions on $\Delta_{q,\lambda}\mathcal{L}$ and ΔF above. We have indicated the chapter which justifies the different conclusions. In particular, see Theorem 36 and Lemma 38 in chapter 4; see Corollary 115 and Theorems 112, 114 and 118 in chapter 6; and see Theorems 139 and 145 in chapter 8.

Singularities of the Information Bottleneck

For the Information Bottleneck problem (2.34),

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

the $NK \times 1$ vector q is always in the kernel of $\Delta F_B(q, \beta)$ for every value of β (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 β . We prove this observation in this section, which shows that ΔF_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]

$$-I(Y; Y_N) = H(Y_N|Y) - H(Y_N). \quad (4.13)$$

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 $\frac{\partial p(\mu)}{\partial q_{\nu k}} = \delta_{\nu\mu} p(y_k)$, 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

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

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

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

This implies that ΔF_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 ΔF_B is singular.

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

Proof. To show that the vector q^ν is in the kernel of $\Delta F_B^\nu(q, \beta)$, the ν^{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

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

This shows that q^ν is in the kernel of ν^{th} block ΔF_B .

THEOREM 43. *For an arbitrary pair (q, β) , the dimension of the kernel of matrix ΔF_B is at least N .*

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

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

By Lemma 42, $\{\mathbf{v}_i\}_{i=1}^N$ 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)

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

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, $\phi(\mathbf{w}, \beta)$, 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

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

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

$$\phi : V \times \mathfrak{R} \rightarrow V.$$

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

$$\phi(gw) = \phi(w)$$

for every $w \in V$ and every $g \in G$. The function ϕ is *G-equivariant* if

$$\phi(gw) = g\phi(w)$$

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

$$H = H_w = \{h \in G | hw = w\}.$$

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

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

The subspace W is *G-invariant* if $gw \in W$ for all $w \in W$. The subspace W is *H-irreducible* if the only H -invariant subspaces of W are $\{0\}$ 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 $\phi : V \times \mathfrak{R} \rightarrow V$ be a G -equivariant function for some Banach space V and let $H \leq G$. Then

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

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

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

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

1. ([33] p.304) The Jacobian $\partial_{\mathbf{w}}\phi(\mathbf{0}, \beta)$ commutes with every $g \in G$.
2. ([34] p.82 or [33] p. 304) The spaces $\ker \partial_{\mathbf{w}}\phi(\mathbf{0}, 0)$ and $\text{range } \partial_{\mathbf{w}}\phi(\mathbf{0}, 0)$ are each G -invariant.
3. ([34] p.69) Let $g \in G$. The vector $\mathbf{w} \in V$ has isotropy subgroup $H \leq G$ if and only if $g\mathbf{w}$ has isotropy subgroup $gHg^{-1} \leq G$.
4. (Trace Formula) ([34] p.76) Let $H \leq G$ where $|H| < \infty$. Then

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

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 $\text{Fix}(G) = \{\mathbf{0}\}$.

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

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

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

$$\partial_{\mathbf{w}}\phi(\mathbf{0}, 0)g = g\partial_{\mathbf{w}}\phi(\mathbf{0}, 0)$$

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

This proves 1.

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

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

Existence Theorems for Bifurcating Branches

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

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

where $\phi : V \times \mathfrak{R} \rightarrow V$ for some Banach space V . If ϕ 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 $\phi : V \times \mathfrak{R} \rightarrow V$ from (5.1) is G equivariant for a compact Lie group G , and a Banach space V .*
2. *The Jacobian $\partial_{\mathbf{w}}\phi(\mathbf{0}, 0) = \mathbf{0}$.*

3. The group G acts absolutely irreducibly on $\ker \partial_{\mathbf{w}}\phi(\mathbf{0}, 0)$ so that $\partial_{\mathbf{w}}\phi(\mathbf{0}, \beta) = c(\beta)I$ 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 $\dim \text{Fix}(H) = 1$.

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

Proof. Let $\hat{\phi} := \phi|_{\text{Fix}(H) \times \mathfrak{R}}$ and let $\mathbf{w}_0 \in \text{Fix}(H)$. By Lemma 44

$$\hat{\phi} : \text{Fix}(H) \times \mathfrak{R} \rightarrow \text{Fix}(H) \quad (5.2)$$

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

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

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

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

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

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

where

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

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

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

Differentiating this equation yields

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

and so

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

from which it follows that

$$k(0, 0) = 0 \quad (5.8)$$

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

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

Thus

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

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

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

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

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

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

By assumption, $\mathbf{w}_0 \in \text{Fix}(H)$, from which it follows that the isotropy group of the branch $(t\mathbf{w}_0, \beta(t))$ 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

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

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 $B_2 \times \mathfrak{R}$,

$$\nabla_w f : B_2 \times \mathfrak{R} \rightarrow B_0,$$

such that $\nabla_w f$ is G -equivariant.

3. The equation $\nabla_w f(\mathbf{0}, \beta) = \mathbf{0}$ holds for every $\beta \in I$ where I is some interval in \mathfrak{R} :

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

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

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

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

$$\dim(\text{peigenspace}(\Delta_w f(\mathbf{0}, \beta_1)) \cap \text{Fix}(H)) \neq \dim(\text{peigenspace}(\Delta_w f(\mathbf{0}, \beta_2)) \cap \text{Fix}(H)),$$

and that

$$\dim(\text{peigenspace}(\Delta_w f(\mathbf{0}, \beta_1)) \cap \text{Fix}(L)) \neq \dim(\text{peigenspace}(\Delta_w f(\mathbf{0}, \beta_2)) \cap \text{Fix}(L)).$$

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 $(w = \mathbf{0}, \beta_H)$ and $(w = \mathbf{0}, \beta_L)$ 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 $\partial_w \phi(\mathbf{0}, \beta)$ changes sign for some β^* in the interval $I \subset \mathfrak{R}$, 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 $\partial_w \phi(\mathbf{0}, \beta)$ 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 $B_2 \times \mathfrak{R}$,*

$$\nabla_w f : B_2 \times \mathfrak{R} \rightarrow B_0,$$

such that $\nabla_w f$ is G -equivariant.

3. The equation $\nabla_w f(\mathbf{0}, \beta) = \mathbf{0}$ holds for every $\beta \in I$ where I is some interval in \mathfrak{R} .
4. Bifurcation of solutions to $\nabla_w f(\mathbf{0}, \beta) = \mathbf{0}$ occurs at $\beta = \beta^*$.
5. The fixed point space $\text{Fix}(G) = \{\mathbf{0}\}$.
6. The kernel $\ker \Delta_w f(\mathbf{0}, \beta)$ is G -irreducible.
7. Let H be a maximal isotropy subgroup of G .

Then there exists bifurcating solutions to

$$\nabla_w f(\mathbf{0}, \beta) = \mathbf{0}$$

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 (w^*, β^*) to (5.1),

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

whose existence is guaranteed by the Equivariant Branching Lemma, is considered.

The independent variable w is in some Banach space V and $\beta \in \mathfrak{R}$, so that

$$\phi : V \times \mathfrak{R} \rightarrow V. \quad (5.13)$$

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 $(tw_0, \beta(t))$ from $(0, 0)$ of the flow (5.1) where $w_0 \in \text{Fix}(H)$ for an isotropy group $H \leq G$. The assumptions we make throughout this section are that

1. ϕ is G -equivariant and infinitely differentiable in w and β , with $\partial_w \phi(0, 0) = 0$.

2. G acts absolutely irreducibly on $\ker \partial_w \phi(\mathbf{0}, 0)$ so that $\partial_w \phi(\mathbf{0}, \beta) = c(\beta)I$ for some scalar function $c(\beta)$.
3. $c(0) = 0$ and $c'(0) > 0$.
4. $H \leq G$ with $\dim \text{Fix}(H) = 1$.

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 $(t\mathbf{w}_0, \beta(t))$ 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 $(t\mathbf{w}_0, \beta(t))$ is transcritical if $\beta'(0) \neq 0$. If $\beta'(0) = 0$, then the branch is called pitchfork-like.*

Golubitsky ([34] p.90) shows that

$$\text{sgn}\beta'(0) = -\text{sgn}c'(0)\text{sgn} \langle \mathbf{w}_0, \partial_{ww}^2 \phi(\mathbf{0}, 0)[\mathbf{w}_0, \mathbf{w}_0] \rangle .$$

We now prove the following generalization.

LEMMA 53. *If Assumption 50 holds, then*

$$\beta'(0) = \frac{-\langle \mathbf{w}_0, \partial_{\mathbf{w}\mathbf{w}}^2 \phi(\mathbf{0}, 0)[\mathbf{w}_0, \mathbf{w}_0] \rangle}{2\|\mathbf{w}_0\|^2 c'(0)}.$$

Proof. As in (5.5), we write

$$\phi(t\mathbf{w}_0, \beta(t)) = tk(t, \beta)\mathbf{w}_0$$

where $k(t, \beta)$ is defined in (5.4). Differentiating (5.12) shows that

$$\partial_t k(t, \beta(t)) + \partial_\beta k(t, \beta(t))\beta'(t) = 0 \quad (5.14)$$

$$\implies \beta'(t) = -\frac{\partial_t k(t, \beta(t))}{\partial_\beta k(t, \beta(t))}. \quad (5.15)$$

By (5.10), $\partial_\beta k(0, 0) = c'(0)$. Differentiating (5.6) yields

$$\partial_{\mathbf{w}\mathbf{w}}^2 \phi(t\mathbf{w}_0, \beta)[\mathbf{w}_0, \mathbf{w}_0] = (2\partial_t k(t, \beta) + t\partial_{tt}^2 k(t, \beta))\mathbf{w}_0 \quad (5.16)$$

showing that

$$\partial_t k(0, 0) = \frac{\langle \mathbf{w}_0, \partial_{\mathbf{w}\mathbf{w}}^2 \phi(\mathbf{0}, 0)[\mathbf{w}_0, \mathbf{w}_0] \rangle}{2\|\mathbf{w}_0\|^2}.$$

Substituting this and $\partial_\beta k(0, 0) = c'(0)$ into (5.15) gives the desired result. \square

REMARK 54.

1. ([34] p.90) *By Assumption 50.3, $\text{sgn}\beta'(0) = -\text{sgn} \langle \mathbf{w}_0, \partial_{\mathbf{w}\mathbf{w}}^2 \phi(\mathbf{0}, 0)[\mathbf{w}_0, \mathbf{w}_0] \rangle$.*

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 β as a function of t about $t = 0$, then equations (5.15) and (5.16) show that

$$\beta'(t) = \frac{\langle \mathbf{w}_0, \partial_{\mathbf{w}\mathbf{w}}^2 \phi(t\mathbf{w}_0, \beta)[\mathbf{w}_0, \mathbf{w}_0] \rangle \|\mathbf{w}_0\|^{-2} - t \partial_{tt}^2 k(t, \beta)}{2\partial_{\beta} k(t, \beta)}. \quad (5.17)$$

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 $(t\mathbf{w}_0, \beta(t))$ is subcritical. If $\beta'(0) > 0$, then the branch is supercritical for $t > 0$.

4. To determine whether a branch $(t\mathbf{w}_0, \beta(t))$ 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 $(t\mathbf{w}_0, \beta(t))$ to $\phi(\mathbf{w}, \beta)$, as guaranteed by Theorem 47, is subcritical, then it consists of unstable solutions.

Proof. Write ϕ as in (5.5),

$$\phi(t\mathbf{w}_0, \beta(t)) = tk(t, \beta).$$

Note that (5.6) shows that \mathbf{w}_0 is an eigenvector of $\partial_{\mathbf{w}} \phi(t\mathbf{w}_0, \beta)$, with eigenvalue

$$\zeta(t, \beta) = k(t, \beta) + t\partial_t k(t, \beta). \quad (5.18)$$

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

$$\zeta(t, \beta) = -t\partial_\beta k(t, \beta)\beta'(t). \quad (5.19)$$

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

$$\zeta(t, \beta) > 0. \quad (5.20)$$

for small t and β . Thus, this branch is unstable for sufficiently small t . \square

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*

$$\text{trace}(\partial_w \phi(tw_0, \beta)) = \dim(V)c'(0)\beta'(0)t + \mathcal{O}(t^2)$$

where V is the Banach space defined in (5.13).

Proof. The Taylor series for $\phi(\mathbf{w}, \beta)$ about $\mathbf{w} = \mathbf{0}$ is

$$\phi(\mathbf{w}, \beta) = \phi(\mathbf{0}, \beta) + \partial_{\mathbf{w}}\phi(\mathbf{0}, \beta)\mathbf{w} + \frac{1}{2}\partial_{\mathbf{w}\mathbf{w}}^2\phi(\mathbf{0}, \beta)[\mathbf{w}, \mathbf{w}] + \mathcal{O}(\mathbf{w}^3). \quad (5.21)$$

Equation (5.3) shows that $\phi(\mathbf{0}, \beta) = \mathbf{0}$, and by Assumption 50.2, $\partial_{\mathbf{w}}\phi(\mathbf{0}, \beta) = c(\beta)I$.

Letting

$$Q(\mathbf{w}, \beta) = \frac{1}{2}\partial_{\mathbf{w}\mathbf{w}}^2\phi(\mathbf{0}, \beta)[\mathbf{w}, \mathbf{w}] \quad (5.22)$$

gives

$$\phi(\mathbf{w}, \beta) = c(\beta)\mathbf{w} + Q(\mathbf{w}, \beta) + \mathcal{O}(\mathbf{w}^3). \quad (5.23)$$

Hence,

$$\partial_{\mathbf{w}}\phi(\mathbf{w}, \beta) = c(\beta)I + \partial_{\mathbf{w}}Q(\mathbf{w}, \beta) + \mathcal{O}(\mathbf{w}^2)$$

from which it follows that

$$\text{trace}(\partial_{\mathbf{w}}\phi(\mathbf{w}, \beta)) = \dim(V)c(\beta) + \text{trace}(\partial_{\mathbf{w}}Q(\mathbf{w}, \beta)) + \mathcal{O}(\mathbf{w}^2).$$

Observe that Q is G -equivariant by the equivariance of ϕ , from which we get $Q(g\mathbf{w}, \beta) = gQ(\mathbf{w}, \beta)$ and so

$$\partial_{\mathbf{w}}Q(g\mathbf{w}, \beta) = g\partial_{\mathbf{w}}Q(\mathbf{w}, \beta)g^{-1}$$

giving

$$\text{trace}(\partial_{\mathbf{w}}Q(g\mathbf{w}, \beta)) = \text{trace}(\partial_{\mathbf{w}}Q(\mathbf{w}, \beta)).$$

Thus, $\text{trace}(\partial_{\mathbf{w}}Q(\mathbf{w}, \beta))$ is a G -invariant function. Furthermore, $\text{trace}(\partial_{\mathbf{w}}Q(\mathbf{w}, \beta))$ is linear in \mathbf{w} since Q is quadratic. Therefore, Propositions 45 and 46.6 assure that

$$\text{trace}(\partial_{\mathbf{w}}Q(\mathbf{w}, \beta)) = 0.$$

Finally, we see that

$$\text{trace}(\partial_{\mathbf{w}}\phi(t\mathbf{w}_0, \beta(t))) = \dim(V)c(\beta(t)) + \mathcal{O}(t^2),$$

which can be rewritten using the Taylor expansion of $c(\beta(t))$ about $t = 0$, showing that

$$\begin{aligned} \text{trace}(\partial_{\mathbf{w}}\phi(t\mathbf{w}_0, \beta(t))) &= \dim(V) (c(0) + c'(0)\beta'(0)t + \mathcal{O}(t^2)) \\ &= \dim(V)c'(0)\beta'(0)t + \mathcal{O}(t^2), \end{aligned} \quad (5.24)$$

where the last equality follows from Assumption 50.3. \square

PROPOSITION 58. ([34] p.93) *Suppose that Assumption 50 holds. If $\beta'(0) > 0$, then for $t > 0$, the unique branch of bifurcating solutions $(t\mathbf{w}_0, \beta(t))$ to $\phi(\mathbf{w}, \beta)$, as guaranteed by Theorem 47, is supercritical and consists of unstable solutions.*

Proof. Remark 54.3 implies that $(t\mathbf{w}_0, \beta(t))$ is supercritical. Claim 57 shows that

$$\text{trace}(\partial_{\mathbf{w}}\phi(t\mathbf{w}_0, \beta)) = \dim(V)c'(0)\beta'(0)t + \mathcal{O}(t^2).$$

from which it follows that $\text{trace}(\partial_{\mathbf{w}}\phi(t\mathbf{w}_0, \beta))$ is positive for sufficiently small t . Thus, some eigenvalue of $\partial_{\mathbf{w}}\phi(t\mathbf{w}_0, \beta)$ has positive real part. \square

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 $\phi(\mathbf{w}, \beta) = \mathbf{0}$, as guaranteed by Theorem 47, consists of unstable solutions.*

Proof. The theorem follows from Propositions 55 and 58, and Remarks 56 and 59. \square

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 $\partial_{\mathbf{w}}Q(\mathbf{w}_0, \beta)$ has an eigenvalue with nonzero real part, where $Q(\mathbf{w}, \beta)$ is the quadratic part of ϕ as in (5.22). Then the unique branch of bifurcating solutions $(t\mathbf{w}_0, \beta(t))$ to $\phi(\mathbf{w}, \beta)$, 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 $\partial_{\mathbf{w}}Q(\mathbf{w}_0, \beta)$ 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

$$\beta''(0) = \frac{-\langle \mathbf{w}_0, \partial_{\mathbf{w}\mathbf{w}\mathbf{w}}^3 \phi(\mathbf{0}, 0)[\mathbf{w}_0, \mathbf{w}_0, \mathbf{w}_0] \rangle}{3\|\mathbf{w}_0\|^2 c'(0)}.$$

Proof. As in (5.5), we write

$$\phi(t\mathbf{w}_0, \beta(t)) = tk(t, \beta)\mathbf{w}_0$$

where $k(t, \beta)$ is defined in (5.4). Twice differentiating (5.12) (or, equivalently, once differentiating (5.14)) shows that

$$\partial_{tt}^2 k + \partial_\beta \partial_t k \beta'(t) + (\partial_t \partial_\beta k + \partial_{\beta\beta}^2 k \beta'(t)) \beta'(t) + \partial_\beta k \beta''(t) = 0.$$

Thus

$$\beta''(t) = \frac{-\partial_{tt}^2 k - 2\partial_\beta \partial_t k \beta'(t) - \partial_{\beta\beta}^2 k \beta'(t)^2}{\partial_\beta k}$$

and so

$$\beta''(0) = \frac{-\partial_{tt}^2 k(\mathbf{0}, 0)}{\partial_\beta k(\mathbf{0}, 0)}. \quad (5.25)$$

By (5.10), $\partial_\beta k(0, 0) = c'(0)$. Differentiating (5.16) with respect to t gives

$$\partial_{\mathbf{w}\mathbf{w}\mathbf{w}}^3 \phi(t\mathbf{w}_0, \beta)[\mathbf{w}_0, \mathbf{w}_0, \mathbf{w}_0] = (3\partial_{tt}^2 k(t, \beta) + t\partial_{ttt}^3 k(t, \beta))\mathbf{w}_0$$

from which it follows that

$$\partial_{tt}^2 k(0, 0) = \frac{\langle \mathbf{w}_0, \partial_{\mathbf{w}\mathbf{w}\mathbf{w}}^3 \phi(\mathbf{0}, 0)[\mathbf{w}_0, \mathbf{w}_0, \mathbf{w}_0] \rangle}{3\|\mathbf{w}_0\|^2}.$$

Substituting this and $\partial_\beta k(0, 0) = c'(0)$ into (5.25) gives the desired result. \square

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,*

$$\operatorname{sgn}(\beta''(0)) = -\operatorname{sgn}(\langle \mathbf{w}_0, \partial_{\mathbf{w}\mathbf{w}\mathbf{w}}^3 \phi(\mathbf{0}, 0)[\mathbf{w}_0, \mathbf{w}_0, \mathbf{w}_0] \rangle).$$

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 $(t\mathbf{w}_0, \beta(t))$, as guaranteed by Theorem 47, is pitchfork-like with $\beta''(0) > 0$, and if*

$$\sum_{i,j,m} \frac{\partial^3 \phi_m(\mathbf{0}, 0)}{\partial x_i \partial x_j \partial x_m} [\mathbf{w}_0]_i [\mathbf{w}_0]_j > 0,$$

then the branch is supercritical and consists of unstable solutions.

Proof. The branch is supercritical by Remark 54.4. To show instability, we determine $\operatorname{trace}(\partial_{\mathbf{w}} \phi(t\mathbf{w}_0, \beta))$ 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

$$T(\mathbf{w}, \beta) = \frac{1}{6} \partial_{\mathbf{w}\mathbf{w}\mathbf{w}}^3 \phi(\mathbf{0}, \beta)[\mathbf{w}, \mathbf{w}, \mathbf{w}], \quad (5.26)$$

then (5.23) can be rewritten as

$$\phi(\mathbf{w}, \beta) = c(\beta)\mathbf{w} + Q(\mathbf{w}, \beta) + T(\mathbf{w}, \beta) + \mathcal{O}(\mathbf{w}^4)$$

and from the proof to Proposition 58 it follows that

$$\operatorname{trace}(\partial_{\mathbf{w}} \phi(t\mathbf{w}_0, \beta(t))) = \dim(V)c(\beta(t)) + \operatorname{trace}(\partial_{\mathbf{w}} T(t\mathbf{w}_0, \beta)) + \mathcal{O}(t^3).$$

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, $\text{trace}(\partial_w \phi(t\mathbf{w}_0, \beta(t)))$ is equal to

$$\dim(V) \left(c'(0)\beta'(0)t + (c''(0)\beta'(0)^2 + c'(0)\beta''(0))\frac{t^2}{2} \right) + \text{trace}(\partial_w T(t\mathbf{w}_0, \beta)) + \mathcal{O}(t^3).$$

This and Assumption 50.3 show that when $\beta'(0) = 0$ and $\beta''(0) > 0$,

$$\text{trace}(\partial_w \phi(t\mathbf{w}_0, \beta(t))) > 0$$

if

$$\text{trace}(\partial_w T(t\mathbf{w}_0, \beta)) > 0$$

for sufficiently small t . Thus, if $\text{trace}(\partial_w T(t\mathbf{w}_0, \beta)) > 0$ for sufficiently small t , then some eigenvalue of $\partial_w \phi(t\mathbf{w}_0, \beta)$ is positive, which implies that the supercritical branch $(t\mathbf{w}_0, \beta(t))$ is unstable.

We now show that $\text{sgn}(\text{trace}(\partial_w T(t\mathbf{w}_0, \beta)))$ for small t is determined by

$$\text{sgn} \left(\sum_{i,j,k} \frac{\partial^3 \phi_k(\mathbf{0}, 0)}{\partial x_i \partial x_j \partial x_k} [\mathbf{w}_0]_i [\mathbf{w}_0]_j \right).$$

The function $[T(\mathbf{w}, \beta)]_l$ from (5.26) can be written as

$$\begin{aligned} & \frac{1}{6} \left(\sum_{i \neq m, j \neq m, k \neq m} \frac{\partial^3 \phi_l(\mathbf{0}, \beta)}{\partial x_i \partial x_j \partial x_k} x_i x_j x_k + 3 \sum_{i \neq m, j \neq m} \frac{\partial^3 \phi_l(\mathbf{0}, \beta)}{\partial x_i \partial x_j \partial x_m} x_i x_j x_m \right. \\ & \left. + 3 \sum_{i \neq m} \frac{\partial^3 \phi_l(\mathbf{0}, \beta)}{\partial x_i \partial x_m \partial x_m} x_i x_m^2 + \frac{\partial^3 \phi_l(\mathbf{0}, \beta)}{\partial x_m^3} x_m^3 \right). \end{aligned} \quad (5.27)$$

Thus, $\partial_{x_m} [T(t\mathbf{w}_0, \beta)]_l$ is

$$\frac{1}{6} t^2 \left(3 \sum_{i \neq m, j \neq m} \frac{\partial^3 \phi_l(\mathbf{0}, \beta)}{\partial x_i \partial x_j \partial x_m} [\mathbf{w}_0]_i [\mathbf{w}_0]_j + 6 \sum_{i \neq m} \frac{\partial^3 \phi_l(\mathbf{0}, \beta)}{\partial x_i \partial x_m \partial x_m} [\mathbf{w}_0]_i [\mathbf{w}_0]_m + 3 \frac{\partial^3 \phi_l(\mathbf{0}, \beta)}{\partial x_m^3} [\mathbf{w}_0]_m^2 \right)$$

which shows that

$$[\partial_{\mathbf{w}} T(t\mathbf{w}_0, \beta)]_{lm} = \frac{1}{2} t^2 \sum_{i,j} \frac{\partial^3 \phi_l(\mathbf{0}, \beta)}{\partial x_i \partial x_j \partial x_m} [\mathbf{w}_0]_i [\mathbf{w}_0]_j.$$

It follows that

$$\text{trace}(\partial_{\mathbf{w}} T(t\mathbf{w}_0, \beta)) = \frac{1}{2} t^2 \sum_{i,j,m} \frac{\partial^3 \phi_m(\mathbf{0}, \beta)}{\partial x_i \partial x_j \partial x_m} [\mathbf{w}_0]_i [\mathbf{w}_0]_j$$

which is positive for sufficiently small t if

$$\sum_{i,j,m} \frac{\partial^3 \phi_m(\mathbf{0}, 0)}{\partial x_i \partial x_j \partial x_m} [\mathbf{w}_0]_i [\mathbf{w}_0]_j > 0.$$

□

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)

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

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 $(\mathbf{w} = \mathbf{0}, \beta = 0)$ is an equilibrium of (5.1). The second assumption is that $\partial_{\mathbf{w}} \phi(\mathbf{0}, 0) = \mathbf{0}$. In other words, we assumed that bifurcation occurs at $(\mathbf{0}, 0)$, and that at the bifurcation, the Jacobian of ϕ vanishes.

This section examines in detail how to transform an arbitrary G -equivariant system such as (3.15),

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

where

$$\psi : B_2 \times \mathfrak{R} \rightarrow B_0,$$

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 (\mathbf{x}^*, β^*) , then the translation $\psi(\mathbf{x} + \mathbf{x}^*, \beta + \beta^*)$ 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

$$\Psi := \partial_{\mathbf{x}}\psi(\mathbf{0}, 0) = \mathbf{0},$$

that the Jacobian of ψ 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), ϕ , which is the restriction of ψ onto $\ker \Psi$ about $(\mathbf{0}, 0)$. More precisely, ψ is restricted to $\ker \Psi$, and ϕ 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 $\text{range} \Psi$ 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 $\text{range} \Psi$ 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

ψ in $\text{range}\Psi$ and in \mathcal{N} are zero:

$$\psi(\mathbf{x}, \beta) = \mathbf{0} \Leftrightarrow E\psi(\mathbf{x}, \beta) = \mathbf{0} \text{ and } (I - E)\psi(\mathbf{x}, \beta) = \mathbf{0}. \quad (5.29)$$

Consider the decomposition $\mathbf{x} = \mathbf{w} + U$, where $\mathbf{w} \in \ker \Psi$ and $U \in \mathcal{M}$, so that the problem $E\psi(\mathbf{x}, \beta) = \mathbf{0}$ can be rewritten as

$$E\psi(\mathbf{w}, U, \beta) = E\psi(\mathbf{w} + U, \beta) = \mathbf{0}.$$

We define the matrix L as

$$L := E\Psi|_{\mathcal{M}}, \quad (5.30)$$

the Jacobian $\partial_{\mathbf{x}}\psi(\mathbf{0}, 0)$ projected onto $\text{range}\Psi$, and restricted to \mathcal{M} . Thus, L is invertible, and the Implicit Function Theorem shows that $E\psi(\mathbf{w} + U, \beta) = \mathbf{0}$ can be solved for $U = U(\mathbf{w}, \beta)$ near $(\mathbf{0}, 0)$,

$$E\psi(\mathbf{x}, \beta) = E\psi(\mathbf{w} + U(\mathbf{w}, \beta), \beta) = \mathbf{0}. \quad (5.31)$$

Substituting this expression into (5.29), we see that $\psi(\mathbf{x}, \beta) = \mathbf{0}$ if and only if

$$(I - E)\psi(\mathbf{w} + U(\mathbf{w}, \beta), \beta) = \mathbf{0}.$$

This function is the Liapunov-Schmidt reduction $\phi(\mathbf{w}, \beta)$:

$$\phi : \ker \Psi \times \mathfrak{R} \rightarrow \mathcal{N}$$

$$\phi(\mathbf{w}, \beta) = (I - E)\psi(\mathbf{w} + U(\mathbf{w}, \beta), \beta). \quad (5.32)$$

Using the chain rule, the Jacobian of (5.32) is the matrix

$$\partial_{\mathbf{w}}\phi(\mathbf{w}, \beta) = (I - E) \cdot \partial_{\mathbf{x}}\psi(\mathbf{x}, \beta) \cdot (I + \partial_{\mathbf{w}}U) \quad (5.33)$$

Since $\ker(I - E) = \text{range}\Psi$, then

$$\partial_w \phi(\mathbf{0}, 0) = \mathbf{0} \quad (5.34)$$

and so the Jacobian of ϕ 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 ψ :

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

Ascertaining the bifurcation structure of the equilibria of this system, solutions to $\phi(\mathbf{w}, \beta) = \mathbf{0}$, means determining the bifurcating branches $(t\mathbf{w}, \beta(t))$ from $(\mathbf{0}, 0)$ for $\mathbf{w} \in \ker \Psi$. The associated bifurcating branch of $\psi = \mathbf{0}$ is straightforward to get:

$$(t\mathbf{w}, \beta(t)) \text{ is a bifurcating branch of } \phi = 0$$

$$\text{if and only if} \quad (5.35)$$

$$\begin{pmatrix} \mathbf{x}^* \\ \beta^* \end{pmatrix} + \begin{pmatrix} t\mathbf{w} \\ \beta(t) \end{pmatrix} \text{ is a bifurcating branch of } \psi = 0.$$

It is convenient to use an equivalent representation of the Liapunov-Schmidt reduction (5.32). Let

$$\{\mathbf{w}_i\}_{i=1}^m \text{ be a basis for } \ker \Psi$$

and let W be the $(NK + K) \times m$ matrix whose column space is $\ker \Psi$. So

$$W = \begin{pmatrix} | & | & & | \\ \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_m \\ | & | & & | \end{pmatrix}.$$

Thus, for every $\mathbf{w} \in \ker \Psi$, there is a $\mathbf{z} \in \mathfrak{R}^m$ such that $W\mathbf{z} = \mathbf{w}$. Now define r by

$$\begin{aligned} r &: \mathfrak{R}^m \times \mathfrak{R} \rightarrow \mathfrak{R}^m \\ r(\mathbf{z}, \beta) &= W^T \phi(W\mathbf{z}, \beta) \\ &= W^T (I - E) \psi(W\mathbf{z} + U(W\mathbf{z}, \beta), \beta) \end{aligned} \quad (5.36)$$

where the last equality is from (5.32). We say that r is equivalent to ϕ since

$$r(\mathbf{z}, \beta) = \mathbf{0} \Leftrightarrow \phi(\mathbf{w}, \beta) = \mathbf{0} \Leftrightarrow \psi(\mathbf{x}, \beta) = \mathbf{0},$$

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

$$\partial_{\mathbf{z}} r(\mathbf{z}, \beta) = W^T (I - E) \cdot \partial_{\mathbf{x}} \psi(\mathbf{x}, \beta) \cdot (W + \partial_{\mathbf{w}} U W). \quad (5.37)$$

So we have introduced the necessary ingredients to define a dynamical system defined by r

$$\dot{\mathbf{z}} = r(\mathbf{z}, \beta).$$

Ascertaining the bifurcation structure of the equilibria of this system, solutions to $r(\mathbf{z}, \beta) = \mathbf{0}$, means determining the bifurcating branches $(t\mathbf{z}, \beta(t))$ from $(\mathbf{0}, 0)$ for

$\mathbf{z} \in \mathfrak{R}^m$. The bifurcating branch of $\psi = \mathbf{0}$ is found via the following relationship:

$$(\mathbf{tz}, \beta(t)) \text{ is a bifurcating branch of } r = 0$$

if and only if (5.38)

$$\begin{pmatrix} \mathbf{x}^* \\ \beta^* \end{pmatrix} + \begin{pmatrix} tW\mathbf{z} \\ \beta(t) \end{pmatrix} \text{ is a bifurcating branch of } \psi = 0.$$

We now compute the derivative of r with respect to β , which we will need in chapter 8 when examining saddle-node bifurcations. Beginning with the definition (5.36), we see that

$$\begin{aligned} \partial_\beta r(\mathbf{z}, \beta) &= W^T(I - E) \frac{\partial}{\partial \beta} \psi(\mathbf{x}, \beta) \\ &= W^T(I - E) \left(\partial_\beta \psi(\mathbf{x}, \beta) + \partial_{\mathbf{x}} \psi(\mathbf{x}, \beta) \frac{\partial}{\partial \beta} (W\mathbf{z} + U(W\mathbf{z}, \beta)) \right) \\ &= W^T(I - E) (\partial_\beta \psi(\mathbf{x}, \beta) + \partial_{\mathbf{x}} \psi(\mathbf{x}, \beta) \partial_\beta U). \end{aligned}$$

Since $(I - E)\partial_{\mathbf{x}}\psi(\mathbf{0}, 0) = \mathbf{0}$, then

$$r(\mathbf{0}, 0) = W^T(I - E)\partial_\beta \psi(\mathbf{0}, 0). \quad (5.39)$$

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

$$\frac{\partial r_i}{\partial z_j} = \langle \mathbf{w}_i, (I - E)\partial_{\mathbf{x}}\psi(\mathbf{x}, \beta) \left(\mathbf{w}_j + \frac{\partial U}{\partial z_j} \right) \rangle.$$

Thus, we get that

$$\frac{\partial^2 r_i}{\partial z_j \partial z_k} = \langle \mathbf{w}_i, (I - E) \left(\partial_x \psi(\mathbf{x}, \beta) \frac{\partial^2 U}{\partial z_j \partial z_k} + \partial_x^2 \psi(\mathbf{x}, \beta) \left[\mathbf{w}_j + \frac{\partial U}{\partial z_j}, \mathbf{w}_k + \frac{\partial U}{\partial z_k} \right] \right) \rangle \quad (5.40)$$

It can be shown that ([33] p.31)

$$\partial_{\mathbf{w}} U(\mathbf{0}, 0) = \mathbf{0}, \quad (5.41)$$

from which it follows that $\frac{\partial U}{\partial z_j}(\mathbf{0}, 0) = \partial_{\mathbf{w}} U(\mathbf{0}, 0) \frac{\partial \mathbf{w}}{\partial z_j}(\mathbf{0}) = \mathbf{0}$. Furthermore, since $(I - E) \partial_x \psi(\mathbf{0}, 0) = \mathbf{0}$, then

$$\frac{\partial^2 r_i}{\partial z_j \partial z_k}(\mathbf{0}, 0) = \langle \mathbf{w}_i, (I - E) \partial_x^2 \psi(\mathbf{0}, 0) [\mathbf{w}_j, \mathbf{w}_k] \rangle. \quad (5.42)$$

Applying the chain rule to (5.40), we get the 4 dimensional array of third derivatives

$$\begin{aligned} \frac{\partial^3 r_i}{\partial z_j \partial z_k \partial z_l} &= \langle \mathbf{w}_i, (I - E) \left(\partial_x^2 \psi(\mathbf{x}, \beta) \left[\mathbf{w}_l + \frac{\partial U}{\partial z_l}, \frac{\partial^2 U}{\partial z_j \partial z_k} \right] + \partial_x \psi(\mathbf{x}, \beta) \frac{\partial^3 U}{\partial z_j \partial z_k \partial z_l} \right. \\ &\quad + \partial_x^3 \psi \left[\mathbf{w}_j + \frac{\partial U}{\partial z_j}, \mathbf{w}_k + \frac{\partial U}{\partial z_k}, \mathbf{w}_l + \frac{\partial U}{\partial z_l} \right] \\ &\quad \left. + \partial_x^2 \psi \left[\mathbf{w}_j + \frac{\partial U}{\partial z_j}, \frac{\partial^2 U}{\partial z_k \partial z_l} \right] + \partial_x^2 \psi \left[\mathbf{w}_k + \frac{\partial U}{\partial z_k}, \frac{\partial^2 U}{\partial z_j \partial z_l} \right] \right) \rangle \quad (5.43) \end{aligned}$$

Using the fact that $\partial_z U(\mathbf{0}, 0) = \mathbf{0}$ and $(I - E) \partial_x \psi = \mathbf{0}$, it follows that

$$\begin{aligned} \frac{\partial^3 r_i}{\partial z_j \partial z_k \partial z_l}(\mathbf{0}, 0) &= \langle \mathbf{w}_i, (I - E) \left(\partial_x^2 \psi(\mathbf{0}, 0) \left[\mathbf{w}_l, \frac{\partial^2 U}{\partial z_j \partial z_k}(\mathbf{0}, 0) \right] \right. \\ &\quad + \partial_x^3 \psi(\mathbf{0}, 0) [\mathbf{w}_j, \mathbf{w}_k, \mathbf{w}_l] \\ &\quad + \partial_x^2 \psi(\mathbf{0}, 0) \left[\mathbf{w}_j, \frac{\partial^2 U}{\partial z_k \partial z_l}(\mathbf{0}, 0) \right] \\ &\quad \left. + \partial_x^2 \psi(\mathbf{0}, 0) \left[\mathbf{w}_k, \frac{\partial^2 U}{\partial z_j \partial z_l}(\mathbf{0}, 0) \right] \right) \rangle. \quad (5.44) \end{aligned}$$

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

$$\theta(\mathbf{z}, \beta) := E\psi(W\mathbf{z} + U(W\mathbf{z}, \beta), \beta).$$

Observe that $\psi = 0$ implies that $\theta = 0$. Differentiating $\theta = 0$ yields

$$\frac{\partial \theta}{\partial z_j} = E\partial_{\mathbf{x}}\psi(\mathbf{w}_j + \frac{\partial U}{\partial z_j}) = 0$$

and

$$\frac{\partial^2 \theta}{\partial z_j \partial z_k} = E\left(\partial_{\mathbf{x}}^2\psi[\mathbf{w}_j + \frac{\partial U}{\partial z_j}, \mathbf{w}_k + \frac{\partial U}{\partial z_k}] + \partial_{\mathbf{x}}\psi\frac{\partial^2 U}{\partial z_j \partial z_k}\right) = 0.$$

Since $\partial_{\mathbf{z}}U(\mathbf{0}, 0) = \mathbf{0}$, we get

$$\frac{\partial \theta}{\partial z_j \partial z_k}(\mathbf{0}, 0) = E\left(\partial_{\mathbf{x}}^2\psi(\mathbf{0}, 0)[\mathbf{w}_j, \mathbf{w}_k] + \partial_{\mathbf{x}}\psi(\mathbf{0}, 0)\frac{\partial^2 U}{\partial z_j \partial z_k}\right) = 0,$$

and $E\partial_{\mathbf{x}}\psi(\mathbf{0}, 0) = L$ (from (5.30)) shows that

$$\frac{\partial^2 U}{\partial z_j \partial z_k}(\mathbf{0}, 0) = -L^{-1}E\partial_{\mathbf{x}}^2\psi(\mathbf{0}, 0)[\mathbf{w}_j, \mathbf{w}_k]. \quad (5.45)$$

Finally, substituting (5.45) into (5.44) shows that

$$\begin{aligned} \frac{\partial^3 r_i}{\partial z_j \partial z_k \partial z_l}(\mathbf{0}, 0) &= \langle \mathbf{w}_i, (I - E)(\partial_{\mathbf{x}}^3\psi(\mathbf{0}, 0)[\mathbf{w}_j, \mathbf{w}_k, \mathbf{w}_l] \\ &\quad - \partial_{\mathbf{x}}^2\psi(\mathbf{0}, 0)[\mathbf{w}_j, L^{-1}E\partial_{\mathbf{x}}^2\psi(\mathbf{0}, 0)[\mathbf{w}_k, \mathbf{w}_l]] \\ &\quad - \partial_{\mathbf{x}}^2\psi(\mathbf{0}, 0)[\mathbf{w}_k, L^{-1}E\partial_{\mathbf{x}}^2\psi(\mathbf{0}, 0)[\mathbf{w}_j, \mathbf{w}_l]] \\ &\quad - \partial_{\mathbf{x}}^2\psi(\mathbf{0}, 0)[\mathbf{w}_l, L^{-1}E\partial_{\mathbf{x}}^2\psi(\mathbf{0}, 0)[\mathbf{w}_j, \mathbf{w}_k]]) \rangle. \end{aligned} \quad (5.46)$$

In chapter 6, it proves useful to use Lemma 63 to compute $\beta''(0)$,

$$\beta''(0) = \langle \mathbf{x}_0, \partial_{\mathbf{z}\mathbf{z}\mathbf{z}}^3 r(\mathbf{0}, 0)[\mathbf{z}_0, \mathbf{z}_0, \mathbf{z}_0] \rangle,$$

where r is the Liapunov Schmidt reduction of some function ψ , z_0 is defined as $Wz_0 = \mathbf{u}$, where z_0 is a solution branch of r , and \mathbf{u} is the corresponding solution branch of ψ . The next Lemma writes $\langle z_0, \partial_{zzz}^3 r(\mathbf{0}, 0)[z_0, z_0, z_0] \rangle$ in terms of ψ and \mathbf{u} .

LEMMA 66. Let $Wz_0 = \mathbf{u}$, where the columns of W are $\{w_i\}$, a basis for $\ker \partial_x \psi(\mathbf{0}, 0)$.

Then $\langle z_0, \partial_{zzz}^3 r(\mathbf{0}, 0)[z_0, z_0, z_0] \rangle$ is equal to

$$\langle \mathbf{u}, \partial_x^3 \psi(\mathbf{0}, 0)[\mathbf{u}, \mathbf{u}, \mathbf{u}] - 3\partial_x^2 \psi(\mathbf{0}, 0)[\mathbf{u}, L^{-1}E\partial_x^2 \psi(\mathbf{0}, 0)[\mathbf{u}, \mathbf{u}]] \rangle$$

Proof. The Lemma follows from (5.46). □

Equivariance of the Reduction

By assumption, the vector valued function ψ from (3.15),

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

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 ϕ equivariant?

This is answered by Lemma 67.1: If \mathcal{M} and \mathcal{N} from (5.28) are G -invariant then ϕ 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 $\{\mathbf{w}_i\}_{i=1}^m$ be a basis for $\ker \Psi$. For each $g \in G$ Proposition 46.2 assures that $g\mathbf{w}_j = \sum_i a_{ij}\mathbf{w}_i$ for $a_{ij} \in \mathfrak{R}$. Define the $m \times m$ matrix $A(g)$ by setting

$$[A(g)]_{ij} := a_{ij}. \quad (5.47)$$

The group for which r is equivariant is

$$\mathcal{A} := \{A(g) | g \in G\}. \quad (5.48)$$

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 ψ 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 $\{\mathbf{w}_i\}_{i=1}^m$ are a basis for $\ker \Psi$. Then $A(g) \in \mathcal{A}$ fixes $\mathbf{x} \in \mathfrak{R}^m$ if and only if $g \in G$ fixes $\mathbf{y} = W\mathbf{x} \in \ker \Psi$.

Proof.

$$\begin{aligned}
 & A(g)\mathbf{x} = \mathbf{x} \\
 \Leftrightarrow & \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} \sum_j a_{1j}x_j \\ \vdots \\ \sum_j a_{mj}x_j \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \\
 \Leftrightarrow & \sum_i \left(\sum_j a_{ij}x_j \right) \mathbf{w}_i = \sum_j x_j \mathbf{w}_j \\
 \Leftrightarrow & \sum_j x_j \sum_i a_{ij} \mathbf{w}_i = \sum_j x_j \mathbf{w}_j \\
 \Leftrightarrow & \sum_j x_j g \mathbf{w}_j = \sum_j x_j \mathbf{w}_j \\
 \Leftrightarrow & g \sum_j x_j \mathbf{w}_j = \sum_j x_j \mathbf{w}_j \\
 \Leftrightarrow & gW\mathbf{x} = W\mathbf{x}.
 \end{aligned}$$

□

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)

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

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 (2.33), then a singularity of the Hessian of (3.18) always gives a bifurcation (Corollary 111), and so one can always apply the bifurcation structure results, which we present in this chapter, to problems of this type (Corollary 121).

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 86 and 88), 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 (Theorem 101 and Lemma 103), and, using these, the existence of bifurcating branches is proved (Theorems 112 and 114). 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 Γ 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^*, λ^*) 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^*, λ^*, β) .

