



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

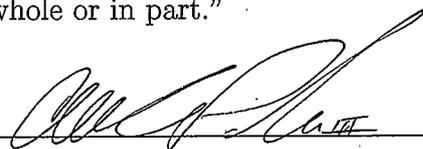
MONTANA STATE UNIVERSITY
Bozeman, Montana

April 2003

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

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

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

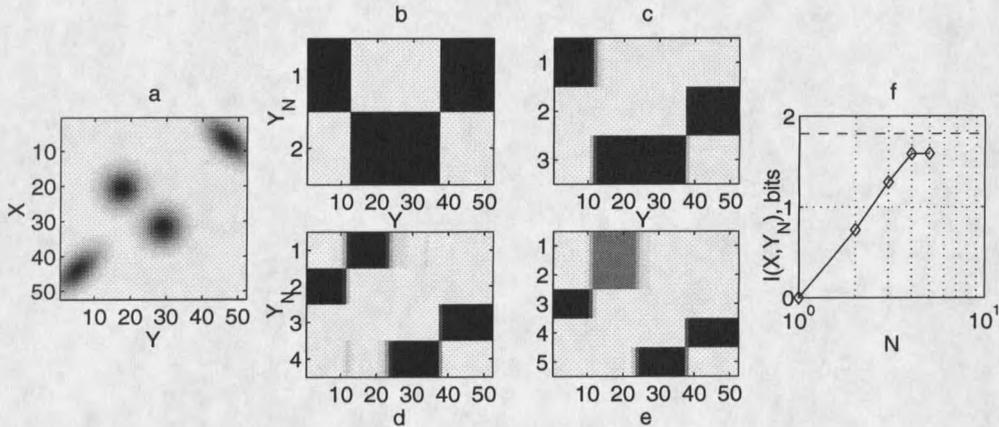


Figure 1. *The Four Blob Problem* from [22, 29]. (a) A joint probability for the relation $p(X, Y)$ between a stimulus set X and a response set Y , each with 52 elements. (b–d) The optimal clusterings $q^*(Y_N|Y)$ for $N = 2, 3$, and 4 classes respectively. These panels represent the conditional probability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

