Symmetry breaking in soft clustering decoding of neural codes

Albert E. Parker, Alexander G. Dimitrov and Tomáš Gedeon

***Abstract*—Information-based distortion methods have successfully been used in the analysis of neural coding problems. These approaches allow the discovery of neural symbols and the corresponding stimulus space of a neuron or neural ensemble quantitatively, while making few assumptions about the nature of either the code or of relevant stimulus features. The neural codebook is derived by quantizing sensory stimuli and neural responses into a small set of clusters, and optimizing the quantization to minimize an information distortion function. The method of annealing has been used to solve the corresponding high dimensional non-linear optimization problem. The annealing solutions undergo a series of bifurcations, which we study using bifurcation theory in the presence of symmetries. In this contribution we describe these symmetry breaking bifurcations in detail, and indicate some of the consequences of the form of the bifurcations. In particular, we show that the annealing solutions break symmetry at pitchfork bifurcations, and that subcritical branches can exist. Thus, at a subcritical bifurcation, there are local information distortion solutions which are not found by the method of annealing. Since the annealing procedure is guaranteed to converge to a local solution eventually, the subcritical branch must turn and become optimal at some later saddle-node bifurcation, which we have shown occur generically for this class of problems. This implies that the rate distortion curve, while convex for non-information based distortion measures, is not convex for information-based distortion methods.**

***Index Terms*—Annealing, bifurcations, clustering, information distortion, neural coding, symmetry breaking**

I. Introduction

A major unresolved problem in neuroscience concerns the manner in which a nervous system represents information. Important questions being studied currently include: What information about the external world is represented in patterns of neural activity? How is this information used by the nervous system to process sensory stimuli? We have yet to reach a generally accepted theory of neural coding and computation. Our difficulty does not stem solely from lack of data. What we lack is a deep understanding of the methods used by interacting populations of neurons to represent and process sensory information.

While we are far from fully answering these deep questions, the theoretical tool we describe here can provide a first step toward discovering general principles of sensory processing in biological systems. It is designed to determine the correspondence between sensory stimuli $\boldsymbol{X}$ and neural activity patterns $\boldsymbol{Y}$. This correspondence is referred to as a *sensory neural code.* Common approaches to this problem often introduce multiple assumptions that affect the obtained solution. For example, the linear stimulus reconstruction method [1] assumes linearity and independence between the neural responses (spikes). The current standard in forward models [2]–[4] places assumptions on either the type of model (for example integrate-and-fire with a stochastic threshold [3]) or the type of point process (essentially, Markov, with specific assumptions about the form of the conditional intensity function [2]) with which the system is characterized.

Any neural code must satisfy several conflicting demands. On one hand the organism must recognize certain natural objects in repeated exposures. Failures on this level may endanger an animal's well-being, for example if a predator is misidentified as a conspecific mate. On this level, the response of the organism needs to be *deterministic*. On the other hand, distinct stimuli need not produce distinguishable neural responses, if such a regime is beneficial to the animal (e.g. a wolf and a fox need not produce distinct responses in a rabbit, just the combined concept of "predator" may suffice.) Thus the representation need not be bijective. Lastly, 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. In these aspects 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 we can view the input-output relationship of a biological sensory system as a *communication system* [5].

We consider the neural encoding process within a probabilistic framework [6], [7]. The *input signal* $\boldsymbol{X}$ to a neuron (or neural ensemble) may be a sensory stimulus or the activity of another set of (pre-synaptic)

All authors are at Montana State University. A. Dimitrov is with the Center for Computational Biology and the Department of Cell Biology and Neuroscience. T. Gedeon is with the Department of Mathematic Sciences. A. Parker is with the Center for Biofilm Engineering as well as the New Zealand Institute of Mathematics at the University of Auckland.

neurons. We consider the input signal to be produced by a stochastic source with probability $p(\boldsymbol{X})$. The *output signal* $\boldsymbol{Y}$ generated by that neuron (or neural ensemble) in response to $\boldsymbol{X}$ is a series of impulses (a spike train or ensemble of spike trains.) Thus the system is completely characterized by its joint distribution, $p(\boldsymbol{X}, \boldsymbol{Y})$. We consider the encoding of $\boldsymbol{X}$ into $\boldsymbol{Y}$ to be a map from one stochastic signal to the other. This stochastic map is the *encoder* $q(\boldsymbol{Y}|\boldsymbol{X})$, which models the operations of this neuronal layer. The output signal $\boldsymbol{Y}$ is induced by the encoder $q(\boldsymbol{Y}|\boldsymbol{X})$ by $p(\boldsymbol{Y}) = \sum_x q(\boldsymbol{Y}|x)p(x)$.

A model of the neural code, which is probabilistic on a fine scale but deterministic on a large scale, emerges naturally in the context of Information Theory [8]. The Noisy Channel Coding Theorem suggests that relations between individual elements of the stimulus and response spaces are not the basic building elements of the system. Rather, the defining objects are relations between *classes* of stimulus-response pairs. Given the mutual information between the two spaces, $\boldsymbol{I}(\boldsymbol{X}; \boldsymbol{Y})$, there are about $2^{\boldsymbol{I}(\boldsymbol{X};\boldsymbol{Y})}$ such *codeword* (or equivalence) classes. When restricted to codeword classes, the stimulus-response relation is almost deterministic. That is, with probability close to 1, elements of $\boldsymbol{Y}$ are associated to elements of $\boldsymbol{X}$ in the same codeword class. This framework naturally deals with lack of bijectivity, by treating it as effective noise. We decode an output y as any of the inputs x that belong to the same codeword class. Similarly, we consider the neural representation of an input x to be any of the outputs y in the same codeword class. Stimuli from the same equivalence class are considered indistinguishable from each other, as are responses from within the same class.

The recently introduced Information Bottleneck [9], [10] and Information Distortion [11], [12] methods approach the neural coding problem in this probabilistic framework by using tools from Rate Distortion theory in order to build simplified models of neural coding and study them in detail. They approximate the joint distribution of interest, $p(\boldsymbol{X}, \boldsymbol{Y})$, by clustering the paired stimulus-response observations $(\boldsymbol{X}; \boldsymbol{Y})$ into smaller stimulus-response spaces $(\boldsymbol{S}; \boldsymbol{T})$. The clustering of the data is called a *soft clustering* since the assignment of the observations to a cluster can be stochastic rather than deterministic. An optimal soft clustering is found by maximizing an information-theoretic cost function subject to both equality and inequality constraints, in hundreds to thousands of dimensions. This analytical approach has several advantages over other current approaches: it yields the most informative approximation of the encoding scheme given the available data (i.e. it gives the lowest distortion, by preserving the most mutual information between stimulus and response classes); the cost function, which is intrinsic to the problem, does not introduce implicit assumptions about the nature or linearity of the encoding scheme; it incorporates an objective, quantitative scheme for refining the *codebook* as more stimulus-response data becomes available; and it does not need repetitions of the stimulus under mild continuity assumptions, so the stimulus space may be investigated more thoroughly.

These types of information theoretic optimization problems also arise in Rate Distortion Theory [8], [13] and the Deterministic Annealing approach to clustering [14]. These methods have been used successfully in neural coding problems [15]–[21] as well as other biological topics [22]–[29] and general data mining problems [14], [30].

One approach to solving this class of optimization problems is through the method of annealing: starting at the uniform (uninformative) soft clustering, one tracks this solution as an annealing parameter varies. The solutions undergo a series of rapid changes (bifurcations or phase transitions) as the annealing parameter increases, ultimately reaching a nearly deterministic clustering of the data. In spite of conjectures about the form of the bifurcations [10], [14], a rigorous treatment of the bifurcations of the annealing solutions and how they relate to bifurcations of solutions to the original information theoretic optimization problem of interest have been lacking. This contribution offers such a description by examining the bifurcations in a dynamical system defined by the gradient flow of the Lagrangian of the optimization problem.

Well established tools are available for exploiting the symmetry of equilibria in a dynamical system. The reason for switching to the gradient flow is to capitalize on these tools. The optimal clustering found by the Information Bottleneck and the Information Distortion methods, which is an equilibrium in the gradient flow, has a symmetry: any clustering of the data gives another equivalent clustering simply by permuting the labels of the N classes. This symmetry is described by S_N, the algebraic group of all permutations on N symbols. The symmetries of the bifurcating equilibria are dictated by the subgroup structure of S_N. We describe these symmetry breaking bifurcations in detail for the gradient flow, relate these back to bifurcations of the annealing solutions, and finally to bifurcations of locally optimal soft clusterings of the information theoretic cost function of interest.

This paper is organized in the following way. In section II we illustrate the application of the method to the analysis of neural coding in the cricket cercal sensory system. In section III we give the Information Bottleneck and Information Distortion optimization problems, and the results of an annealing procedure used to solve the Information Distortion problem on a simple

data set which exhibits the generic bifurcation structure. Section IV presents some relevant constrained optimization theory, and an overview of bifurcation theory with symmetries. Section V is devoted to preparations for applying the theory of bifurcations with symmetries. We introduce the gradient flow of the Lagrangian and the reduced bifurcation problem which, due to the symmetry, determines the directions of all of the emanating equilibria in the much larger space of all soft clusterings. Section VI is the central part of the paper. We present existence theorems for symmetry breaking bifurcating branches, and we derive a condition which determines whether these branches are subcritical (first order phase transitions) or supercritical (second order phase transitions). There are also symmetry preserving bifurcations, which, generically, are saddle-nodes. Numerical illustrations of our results occupy section VII. In section VIII, we discuss some of the insights that the bifurcation structure gives regarding optimal clusterings of the data, and consequences for the *rate distortion curve* from Information Theory.

II. A CASE STUDY

To approach the neural coding problem with the Information Distortion and Information Bottleneck methods [10], [11], [31], one clusters sensory stimuli and neural responses to small reproduction sets in a way which optimizes an information-based distortion function [31]. The essential basis for this approach is to conceptualize a neural coding scheme as a collection of stimulus-response classes akin to a dictionary or *codebook*, with each class corresponding to a neural response *codeword* and its corresponding stimulus feature in the codebook.

A. Finding the codebook

Given the probabilistic model of neural function, we would like to recover the codebook. In our context, this means identifying the joint stimulus-response classes that define the coding relation. We characterize a neural coding scheme by clustering (quantizing or compressing) the joint stimulus-response space $(\boldsymbol{X};\boldsymbol{Y})$ to a smaller joint reproduction space $(\boldsymbol{S};\boldsymbol{T})$. $\boldsymbol{S}$ consists of classes of objects in $\boldsymbol{X}$, and $\boldsymbol{T}$ consists of classes of objects in $\boldsymbol{Y}$. One way to achieve this goal is by clustering the neural responses $\boldsymbol{Y}$ into a coarser representation in a small reproduction space $\boldsymbol{T}$ with $N = |\boldsymbol{T}|$ elements. This quantization induces a quantization of the stimulus space $\boldsymbol{X}$ into a smaller event set $\boldsymbol{S}$ also with N elements. The details of how the clustering is performed are presented in Section III. This method allows us to study coarse (i.e. small N) but highly informative models of a coding scheme, and then to refine them when more data becomes available. The refinement is achieved by simply increasing the sizes of the reproductions, N. We aim to find the best such clustering of the data with fixed N.

Following examples from rate distortion theory [8], [14], the Information Distortion method assumes that the best clustering of the data is the one with maximal entropy [11], [32]. The reason is that, among all clusterings that satisfy a given set of constraints, the maximum entropy clustering of the data does not implicitly introduce additional constraints in the problem. Similarly, the Information Bottleneck method follows the standard settings of Rate-Distortion Theory [8], formulating the problem as a minimal rate at a fixed distortion level.

B. Analysis of stimulus-response relations in the cricket cercal sensory system

We applied these tools to characterize the encoding characteristics of single identified sensory interneurons in the cricket cercal sensory system to complex and biologically relevant stimuli. The goal of the experiments and analyzes were to discover (jointly) the dynamic stimulus waveform features encoded by the cells, and the spike train codeword classes that encoded those features. Most of these results have been presented elsewhere [18], [20].

1) Experimental protocols: The preparation we analyze here is the cercal sensory system of the cricket. In the following sections, we briefly introduce this system, describe the experimental methods used to collect the data, and then discuss the application of the Information Distortion approach to analysis of coding by single sensory interneurons in this system.

Functional organization of the cercal system. This system mediates the detection and analysis of low velocity air currents in the cricket's immediate environment. This sensory system is capable of detecting the direction and dynamic properties of air currents with great accuracy and precision [33]–[36], and can be thought of as a near-field, low-frequency extension of the animal's auditory system.

Primary sensory interneurons. The sensory afferents of the cercal system synapse with a group of approximately thirty local interneurons [37] and approximately twenty identified projecting interneurons that send their axons to motor centers in the thorax and integrative centers in the brain [38]. It is a subset of these projecting interneurons that we study here. Like the afferents, these interneurons are also sensitive to the direction and dynamics of air current stimuli [33]–[36]. Stimulus-evoked neural responses have been measured in several projecting and local interneurons, using several different classes of air current stimuli [34]–[36], [39]. The stimuli that have been used range from simple unidirectional air currents to complex multi-directional, multi-frequency waveforms. Each of the interneurons studied so far

has a unique set of directional and dynamic response characteristics. Previous studies have shown that these projecting interneurons encode a significant quantity of information about the direction and velocity of low frequency air current stimuli with a linear rate code [35], [36], [39]. More recent studies demonstrate that there is also substantial amount of information in the spike trains that cannot be accounted for by a simple linear encoding scheme [18], [40]. Evidence suggests the implementation of an ensemble temporal encoding scheme in this system.

Dissection and preparation of specimens All experiments were performed on adult female crickets obtained from commercial suppliers (Bassett's Cricket Ranch, Visalia, CA, and Sunshine Mealworms, Silverton, OR). Specimens were selected that had undergone their final molt within the previous 24 h. The legs, wings and ovipositor were removed from each specimen, and a thin strip of cuticle was removed from the dorsal surface of the abdomen. After removal of the gut, the body cavity was rinsed and subsequently perfused with hypotonic saline. Hypotonicity facilitated microelectrode penetration of the ganglionic sheath.

The preparation was pinned to the center of a thin disc of silicone elastomer approximately 7 cm in diameter, located within the central arena of an air-current stimulation device, described below. Once the preparation was sealed and perfused with saline, the ganglion was placed on a small platform and gently raised from the ventral surface of the abdomen. This increased the accessibility of the ganglion to electrodes while at the same time improving the stability of electrode penetration by increasing surface tension on the ganglion.

Electrophysiological recording Sharp intracellular electrodes were pulled from glass capillary tubes by a model P*97/PC electrode puller (Sutter Instrument Co.) The electrodes were filled with a mixture of 2% neurobiotin and 3 M KCl, and had resistances in the range of 10 to 30 megohms. During recordings the neurobiotin would diffuse into the nerve cell, allowing for subsequent staining and identification. Data were recorded using an NPI SEC-05L Intracellular amplifier and sampled at 10 kHz rate with a digital data acquisition system running on a Windows 2000 platform.

Stimulus generation The cricket cercal sensory system is specialized to monitor air currents in the horizontal plane. All stimuli for these experiments were produced with a specially-designed and fabricated device that generated laminar air currents across the specimens' bodies. Air currents were generated by the controlled, coordinated movement of loudspeakers. The loudspeakers were mounted facing inward into an enclosed chamber that resembled a miniature multi-directional wind tunnel. The set of speakers were sent appropriate voltage signals to drive them in a "push-pull" manner to drive controlled, laminar air-current stimuli through an enclosed arena in the center of the chamber, where the cricket specimens were placed after dissection.

Stimulus waveforms were constructed prior to the experiment using MATLAB®. During experiments, the stimulus waveforms were sent out through a DAC to audio amplifiers and then to the set of loudspeakers. Stimuli consisted of uninterrupted waveform, for which the air current velocity was drawn from a Gaussian White Noise process, band-passed between 5 and 150 Hz. Two independent waveforms were presented along two orthogonal axes, thus covering all possible planar stimulus directions around the cricket.

2) Results: Stimuli and responses were preprocessed to a form suitable for the algorithm. The response of a single cell is represented as a sequence of inter-spike intervals (ISIs), the times between impulses that the cell emits in response to sensory stimuli [41]. The sequence analyzed here is broken into sets of pairs of ISIs, and embedded in two dimensional space [20], [42]. As described in [18], to be considered a pattern and further processed, a sequence of spikes must start with a spike preceded by a quiet period of at least $D\ ms$. Each ISI is also limited to no more than $T\ ms$. The parameters of the initial processing, D and T, may be varied to verify their effects on the final results. They depend on the cell and system being considered. Typically we use $D \in [10\ 25]ms$ and $T \in [15\ 50]ms$. The stimulus associated with each response is an airflow waveform extracted in a range of $[T^-\ \ T^+]$ around the beginning of each response sequence of ISIs. The stimuli presented to the system consist of two independent time series of air velocities ("along" and "across" the cricket's body), each of length L, and so are embedded in $2L$ dimensional Euclidean space. The number of observations, n, depends on the recording rate and overall cell responsiveness to a given stimulus. The choice of specific parameters is evident in the figures where they are discussed. The complete data set to be processed by the algorithm consists of n pairs $(x, y) \in \Re^{2L} \times \Re^2$, where L is large.

Using the Information Distortion method discussed in Section III, we found optimal soft clusterings that identified synonymous classes of stimulus-response pairs. Stimulus features are represented as waveforms of the mean airflow velocity immediately preceding the elicited spike pattern codewords. The response space was taken to be all pairs of ISIs with $T < 30ms$, preceded by at least $D = 30ms$ of silence. This was done with the intent of analyzing only well-isolated codewords, which are assumed to be independent by this selection process.

Figure 1 illustrates the application of the algorithm to uncovering the stimulus-response relation in an identified cell in the cricket cercal sensory system (cell 10-2, nomenclature as in [38]). The stimulus classes are rep-

resented by their class-conditioned means. We suppress showing confidence intervals for the class conditioned means for reasons of visualization clarity. Each conditional mean has two channels (Panels A and B).

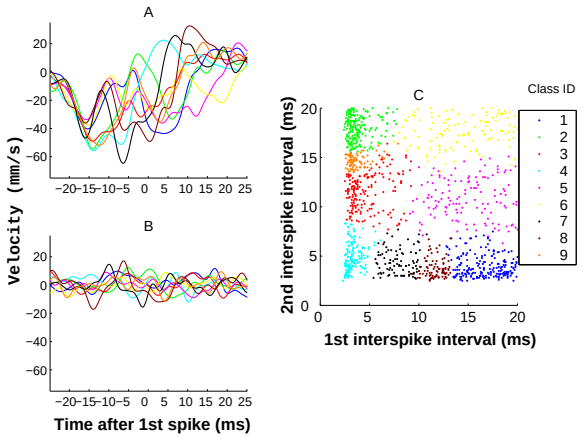

Fig. 1. A quantization to nine classes of the stimulus-response pairs of cell 10-2 in the cricket cercal sensory system. Panels A and B show the two channels of the conditional means of the air flow stimulus for each class. Panel C depicts the two dimensional response space of all pairs of ISIs in the range $[0, 30]ms \times [0, 30]ms$ color-coded by their membership in particular classes. The color labels are consistent among the panels.

The optimal information-based soft clustering produced response classes that were physiologically consistent, in the sense that responses that had similar ISIs were clustered together. Since there was not an explicit similarity criterion for either the stimuli, or the response, this structure is an important emergent property of the algorithm that reflects the underlying structure of the biological system. The stimulus classes are clearly discriminable (Panel A), and associated with features of the clustered responses. For example, the mean of class 2 (green) has two prominent downward excursions separated by about 15 ms, which is the average ISI separation of responses combined in this class. The second trough of the stimulus is consistently related to the second ISI in the response. In panel C, the classes starting with a short first ISI (horizontal axis) are 4, 3, 9 and 2 in order of increasing second ISI (vertical axis). These four classes effectively break the stimulus into a set of discriminable events (Panel A). This sequence also demonstrates the main topic of symmetry in this article: the labels of the clusters are arbitrary. Permuting the labels of the clusters of responses does not effect the discovered relationship between the stimuli and these clusters of responses (this symmetry does not refer to properties of neurons or of the stimulus space).

The information theoretic clustering approach was also used to directly address questions about the consistency of the neural code between individuals of the same species. This extends the approach taken in [21] to select a limited set of neural activity classes and test for similarity across individuals. The quantization was performed on 36 identified 10-2 cells, and 40 identified 10-3 cells (nomenclature as in [38]). 10-3 cells have functionality similar to that of 10-2 cells with directional selectivity offset by 90^o. In Figure 2 we investigate the position of the boundary between class 4 of the neural responses and the neighboring class 7 across a set of individual crickets. This boundary, indicated by the vertical black line near $5.75ms$ for cell 10-2 in Figure 2, can be seen between the light blue and black points in panel C of Figure 1. The standard deviation of the boundary is less than 1 ms *across the set of individuals*! That is, this particular class is very well preserved in the cricket population we study. This directly addresses universal coding behavior at the level of individual response codewords.

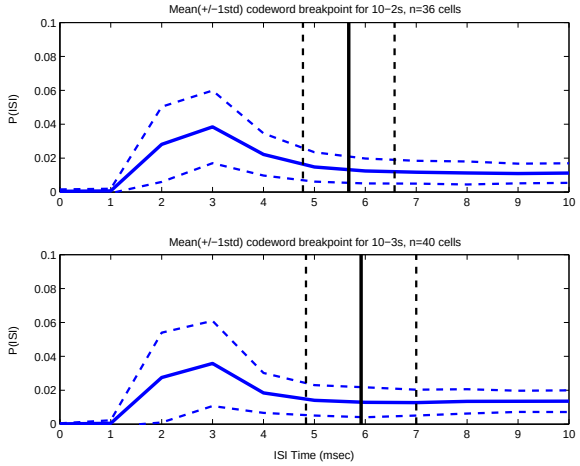

Fig. 2. The top panel shows the summary for cell 10-2 from 36 different individuals; the bottom panel shows cell 10-3 from 40 different individuals. For each animal, the normalized histogram of the first ISI for the neural responses in classes 4 and 7 was calculated. The mean of these distributions is given (solid blue line), as well as 2 standard deviations around the mean (dashed blue line). The solid black vertical line represents the mean time coordinate of the boundary between classes 4 and 7, the dashed black vertical lines indicate 1 standard deviation around the mean. In both cells, class 4 (shortest first doublets) is consistently preserved with a precision of about 1 ms **between different individuals!**

C. Conclusions

The general goal of this section was to demonstrate the application of the Information Distortion method to resolving the neural coding problem. The essential basis for this approach was to conceptualize a neural coding scheme as a collection of stimulus-response classes akin to a dictionary or codebook, with each class corresponding to a neural response codeword and its corresponding stimulus feature in the codebook. The analysis outlined

here enabled the derivation of such a neural codebook, by quantizing stimuli and neural responses into small reproduction sets and optimizing the quantization to minimize the Information Distortion function.

The major advantage of this analytical approach over other current approaches is that it yields the most informative approximation of the encoding scheme given the available data. That is, it gives a representation that preserves the most mutual information between stimulus and response classes. Moreover, the cost function (which is intrinsic to the problem) does not introduce implicit assumptions about the nature or linearity of the encoding scheme, nor does the maximum entropy soft clustering introduce additional implicit constraints to the problem.

A major thrust in this area is to find algorithms through which the relevant stimulus space and the corresponding neural symbols of a neuron or neural ensemble can be discovered simultaneously and quantitatively, making few assumptions about the nature of the code or relevant features. The analysis presented in the following sections of this manuscript enables this derivation of a neural codebook by optimizing the Information Distortion function.

III. Analytic Formulation

How can we characterize a relationship between inputs and outputs $\boldsymbol{X} \leftrightarrow \boldsymbol{Y}$, defined by the joint distribution $p(\boldsymbol{X}, \boldsymbol{Y})$, in which both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are large spaces? We approach this problem by clustering (quantizing) the stimulus and response spaces to smaller reproduction spaces $\boldsymbol{S}$ and $\boldsymbol{T}$ [20], [43]. The joint probability $p(\boldsymbol{S}, \boldsymbol{T})$ between the reproduction stimulus and response spaces, $\boldsymbol{S} \leftrightarrow \boldsymbol{T}$, induces an approximation of the original relationship by

$$p(s,t) = \sum_{x,y} q(s|x)q(t|y)p(x,y).$$

In this section we introduce the Information Bottleneck and Information Distortion methods, which determine an optimal soft clustering $q^*(\boldsymbol{T}|\boldsymbol{Y})$ of the response space $\boldsymbol{Y}$ to a small reproduction space $\boldsymbol{T}$ by optimizing an information-based distortion function [10], [11]. In general the stimulus clustering $q(\boldsymbol{S}|\boldsymbol{X})$ can be optimized independently [20]. In this manuscript we do not explicitly cluster the stimulus space, but set $\boldsymbol{S} \equiv \boldsymbol{X}$ ($q(\boldsymbol{S}|\boldsymbol{X})$ is the identity), and consider only the one-sided quantization of $\boldsymbol{Y}$, so that $p(\boldsymbol{X}, \boldsymbol{Y})$ is approximated by

$$p(s,t) = \sum_{y} q(t|y)p(x,y).$$

The *soft clustering* $q(\boldsymbol{T}|\boldsymbol{Y})$ is a conditional probability which assigns each of the K elements in the large space $\boldsymbol{Y}$ to each of the N classes in the small space $\boldsymbol{T}$ ($N << K$) with some level of uncertainty. The space of valid conditional probabilities $\Delta \subset \Re^{NK}$ is

$$\Delta := \left\{ q(\boldsymbol{T}|\boldsymbol{Y}) \mid \sum_{t=1}^{N} q(t|y) = 1 \text{ and } q(t|y) \geq 0 \ \forall \ t, y \right\}.$$

The Information Bottleneck method finds an optimal soft clustering $q^*(\boldsymbol{T}|\boldsymbol{Y})$ by solving a rate distortion problem of the form

$$R_I(I_0) := \begin{array}{c} \min_{q\in\Delta} \boldsymbol{I}(\boldsymbol{Y};\boldsymbol{T}) \\ \boldsymbol{I}(\boldsymbol{X};\boldsymbol{T}) \geq I_0 \end{array} \tag{1}$$

where $I_0 > 0$ is some *information rate*. The function $R_I(I_0)$ is referred to as the *relevance-compression function* in [44]. The *mutual information*, $I(\boldsymbol{X};\boldsymbol{T})$, is a convex function of $q(\boldsymbol{T}|\boldsymbol{Y})$

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{T}) &= E_{\boldsymbol{X},\boldsymbol{T}} \log \frac{p(\boldsymbol{X},\boldsymbol{T})}{p(\boldsymbol{X})p(\boldsymbol{T})} \\ &= \sum_{x,y,t} q^t p(x,y) \log \left(\frac{\sum_y q^t p(x,y)}{p(x)\sum_y p(y) q^t} \right). \end{aligned}$$

Here, so that the action of the group of symmetries is clear, the soft clustering $q := q(\boldsymbol{T}|\boldsymbol{Y})$ has been decomposed into sub-vectors $q^t = q(t|\boldsymbol{Y}) \in \Re^K$ so that $q = ((q^1)^T \ \ (q^2)^T \ \ldots \ (q^N)^T)^T \in \Re^{NK}$.

The Information Distortion method determines an optimal soft clustering by solving the maximum entropy problem

$$R_H(I_0) := \begin{array}{c} \max_{q\in\Delta} \boldsymbol{H}(\boldsymbol{T}|\boldsymbol{Y}) \\ \boldsymbol{I}(\boldsymbol{X};\boldsymbol{T}) \geq I_0 \end{array} . \tag{2}$$

The *conditional entropy* $\boldsymbol{H}(\boldsymbol{T}|\boldsymbol{Y})$ of the classes $\boldsymbol{T}$ given the neural responses, is a concave function of $q(\boldsymbol{T}|\boldsymbol{Y})$

$$\begin{aligned} H(\boldsymbol{T} \mid \boldsymbol{Y}) &= -E_{\boldsymbol{Y},\boldsymbol{T}} \log q(\boldsymbol{T}|\boldsymbol{Y}) \\ &= -\sum_{y,t} p(y) q^t \log \left(q^t\right) \end{aligned}$$

Both problems (1) and (2) are of the form

$$R(I_0) := \begin{array}{c} \max_{q\in\Delta} G(q) \\ D(q) \geq I_0 \end{array} \tag{3}$$

where

$$G(q) = \sum_{t=1}^{N} g(q^t) \ \text{ and } \ D(q) = \sum_{t=1}^{N} d(q^t), \tag{4}$$

and the real valued functions g and d are sufficiently smooth. This type of problem also arises in Rate Distortion Theory [8], [13] and the Deterministic Annealing approach to clustering [14].

The form of G and D indicates that permuting the sub-vectors q^t does not change the value of G and D. In other words, G and D are invariant to the action of S_N,

$$G(q(\boldsymbol{T}|\boldsymbol{Y})) = G(q(\gamma(\boldsymbol{T})|\boldsymbol{Y})) = G(\gamma(q(\boldsymbol{T}|\boldsymbol{Y})))$$

(and similarly for $D(q)$) where γ acts on $\boldsymbol{T}$ by relabeling the classes $t \in \{1, ..., N\}$. In the language of equivariant bifurcation theory [45], G and D are said to be S_N-invariant, where S_N is the algebraic group of all permutations on N symbols [46], [47].

The method of annealing has been used to find solutions to optimization problems of the form (3) [9]–[12], [14], [17]. The annealing problem is

$$\max_{q \in \Delta} \left(G(q) + \beta D(q) \right) \tag{5}$$

where the non-negative annealing parameter β, a function of I_0 for $\beta > 0$, is the Lagrange multiplier for the constraint $D(q) \geq I_0$ in the optimization problem (3). The reciprocal of the annealing parameter is usually referred to as temperature, in analogy to physical annealing. After starting at q_0 at $\beta_0 = 0$, for which $G(q)$ is maximal, one continues this solution as β increases (temperature decreases) to $\beta_{\max}$, creating a sequence (q_k, β_k) that converges to $(q^*, \beta_{\max})$. We will show that a solution q^* of the annealing problem (5) is always a solution of the optimization problem (3) for $I_0 = D(q^*)$. However, a solution of (3) is not necessarily a solution of (5), although the *stationary points* (critical points or the set of possible solutions) of (3) and (5) are the same when $\beta > 0$ (see section IV-B).

The annealing problem corresponding to (1) is [9], [10], [44]

$$\max_{q \in \Delta} \left(-\boldsymbol{I}(\boldsymbol{Y}; \boldsymbol{T}) + \beta \boldsymbol{I}(\boldsymbol{X}; \boldsymbol{T}) \right), \tag{6}$$

and the annealing problem for (2), in analogy with Deterministic Annealing [14], is

$$\max_{q \in \Delta} \left(\boldsymbol{H}(\boldsymbol{T}|\boldsymbol{Y}) + \beta \boldsymbol{I}(\boldsymbol{X}; \boldsymbol{T}) \right) \tag{7}$$

[11], [12], [17], [48].

The following basic annealing algorithm produces a solution, q^*, of the annealing problem (5) (and of the optimization problem (3) for some I_0) by starting at a maximum q_0 of $G(q)$ (at $\beta_0 = 0$), and then continuing this solution as β increases from 0 to $\beta_{\max}$, creating a sequence (q_k, β_k) that converges to $(q^*, \beta_{\max})$.

Algorithm 3.1 (Annealing): Let

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

and let $\beta_0 = 0$, $\beta_{\max} > 0$. For $k \geq 0$, let (q_k, β_k) be a solution to the annealing problem (5). Iterate the following steps until $\beta_{\mathcal{K}} = \beta_{\max}$ for some $\mathcal{K}$.

1) Perform β-step: Let $\beta_{k+1} = \beta_k + s_k$ where $s_k > 0$.
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) Solve $\max_{q \in \Delta} \left(G(q) + \beta_{k+1} D(q) \right)$ 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 so that in step 3, we find $q_{k+1} \neq q_k$.

A. An Example: The Four Blob Problem

To illustrate the behavior of the annealing solutions, consider the method of annealing applied to (7), for $\beta \in [0, 2]$, where $p(\boldsymbol{X}, \boldsymbol{Y})$ is a discretization of a mixture of four well separated Gaussians, presented by the authors in [11], [12] (Figure 3). In this model, we assume that $\boldsymbol{X} \in \{x_i\}_{i=1}^{52}$ represents a range of possible stimulus properties and that $\boldsymbol{Y} \in \{y_i\}_{i=1}^{52}$ represents a range of possible neural responses. There are four *modes* in $p(\boldsymbol{X}, \boldsymbol{Y})$, where each mode corresponds to a range of responses elicited by a range of stimuli. For example, the stimuli $\{x_i\}_{i=1}^{11}$ elicit the responses $\{y_i\}_{i=39}^{52}$ with high probability, and the stimuli $\{x_i\}_{i=24}^{37}$ elicit the responses $\{y_i\}_{i=22}^{38}$ with high probability. One would expect that the maximizer q^* of (7) will cluster the neural responses $\{y_i\}_{i=1}^{52}$ into four classes, each of which corresponds to a mode of $p(\boldsymbol{X}, \boldsymbol{Y})$. This intuition is justified by the Asymptotic Equipartition Property for jointly typical sequences [8].

The optimal clustering q^* for $N = 2$, 3, and 4 is shown in panels (b)–(d) of Figure 3. The clusters can be labeled by $\boldsymbol{T} \in \{1, ..., N\}$. When $N = 2$ as in panel (b), the optimal clustering q^* yields an incomplete description of the relationship between stimulus and response, in the sense that responses $\{y_i\}_{i=1}^{37}$ are in class 2 and the responses $\{y_i\}_{i=38}^{52}$ are in class 1. The representation is improved for the $N = 3$ case shown in panel (c) since now $\{y_i\}_{i=1}^{11}$ are in class 3, while the responses $\{y_i\}_{i=12}^{37}$ are still clustered together in the same class 2. When $N = 4$ as in panel (d), the elements of $\boldsymbol{Y}$ are separated into the classes correctly. The mutual information in (e) 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.

The results from annealing the Information Distortion problem (7) for $N = 4$ are given in Figure 4. The behavior of $D(q) = \boldsymbol{I}(\boldsymbol{X}; \boldsymbol{T})$ as a function of β can be seen in the top panel. Some of the optimal clusterings q_k for different values of β_k are presented on the bottom row (panels 1 – 6). Panel 1 shows the uniform clustering, denoted by $q_{\frac{1}{N}}$, which is defined componentwise by $q_{\frac{1}{N}}(t|y) := \frac{1}{N}$ for every t and y. The abrupt symmetry breaking transitions as β increases (depicted in panels $1 \rightarrow 2$, $2 \rightarrow 3$ and $5 \rightarrow 6$) are typical for annealing problems of the type (5) [9]–[12], [14].

The action of S_N (where $N = 4$) on the clusterings q can be seen in Figure 4 in any of the bottom panels.

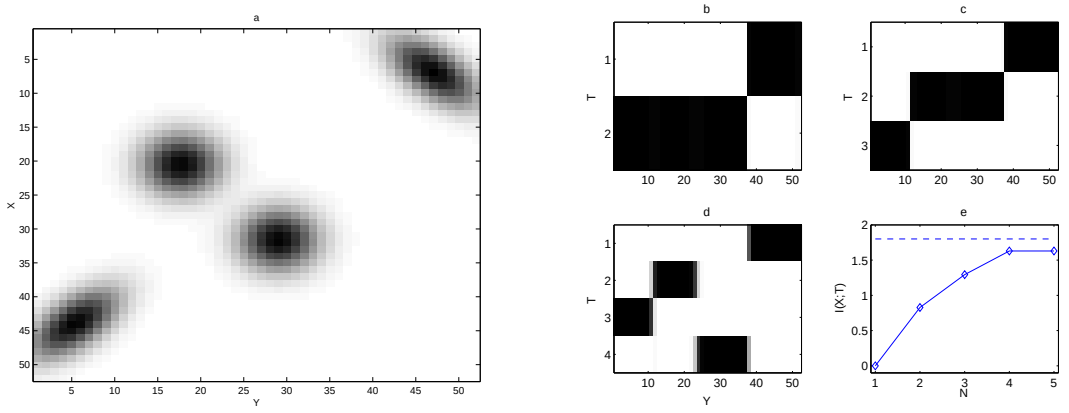

Fig. 3. *The Four Blob Problem* from [11], [12]. (a) A joint probability $p(x, y)$ between a stimulus set $\boldsymbol{X}$ and a response set $\boldsymbol{Y}$, each with 52 elements. (b–d) The optimal clusterings $q^*(t|y)$ for $N = 2$, 3, and 4 classes respectively. These panels represent the conditional probability $q(t|y)$ of a response being classified to a class $t = \nu$. White represents $q(\nu|y) = 0$, black represents $q(\nu|y) = 1$, and intermediate values are represented by levels of gray. 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 $\boldsymbol{I}(\boldsymbol{X}; \boldsymbol{T})$ with increasing N can be seen in (e). The dashed line is $\boldsymbol{I}(\boldsymbol{X}; \boldsymbol{Y})$, which is the least upper bound of $\boldsymbol{I}(\boldsymbol{X}; \boldsymbol{T})$.

The action of S_N permutes the numbers on the vertical axis which merely changes the labels of the classes $\{1, ..., N\}$. Due to the form of G and D given in (4), the value of the annealing cost function (5) is invariant to these permutations.

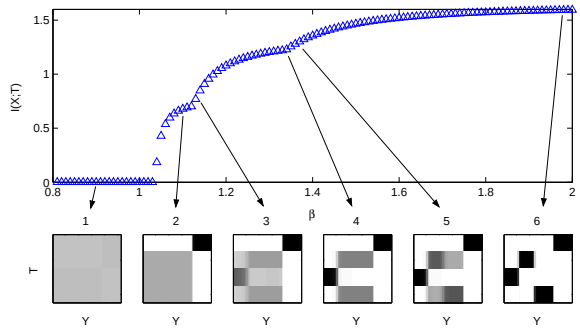

Fig. 4. The bifurcations of the solutions (q^*, β) to the Information Distortion problem (7) initially observed by Dimitrov and Miller in [11]. For a mixture of 4 well-separated Gaussians, the behavior of $D(q) = \boldsymbol{I}(\boldsymbol{X}; \boldsymbol{T})$ as a function of β is shown in the top panel, and some of the solutions $q^*(\boldsymbol{T}|\boldsymbol{Y})$ are shown in the bottom panels.

The bifurcation diagram in Figure 4 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 4, observe that $q \in \Re^{4K} = \Re^{208}$. Why should we observe only 3 bifurcations to local solutions of the annealing problem (5) in such a large dimensional space? What types of bifurcations should we expect: pitchfork, transcritical, saddle-node, or some other type? At a bifurcation, how many bifurcating branches are there? Are the bifurcating branches subcritical ("turn back") or supercritical? When does a bifurcating branch contain solutions of the optimization problem (3) and the corresponding annealing problem (5)? Do bifurcations of solutions to the annealing problem (5) reveal properties (such as convexity) of the original cost function $R(I_0)$ in (3)? How do the bifurcations of solutions to the annealing problem (5) relate to bifurcations of solutions to the optimization problem (3), which has no explicit dependence on the Lagrange multiplier β?

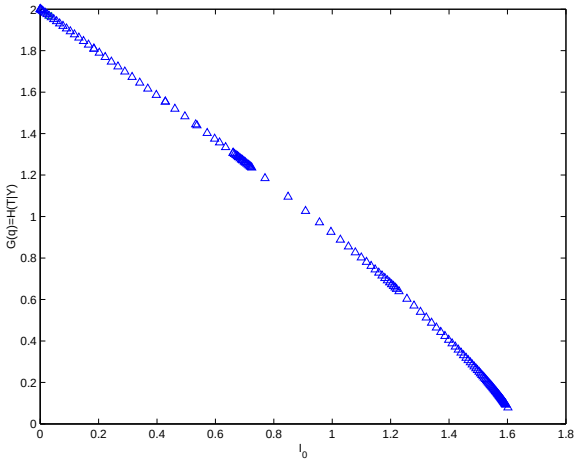

Fig. 5. A graph of $R(I_0)$ (3) for the Information Distortion problem (2).

To help answer this last question, one can solve the optimization problem (3) directly by annealing in I_0. As in Algorithm 3.1, in step 1, one can initially set $I_0 = 0$ and then increment by $I_{k+1} = I_k + s_k$; use the same initial guess in step 2; and now solve (3) in step 3. Using this method, we found solutions of (3) for a sequence of I_0. We plot $R(I_0)$ over this sequence in Figure 5.

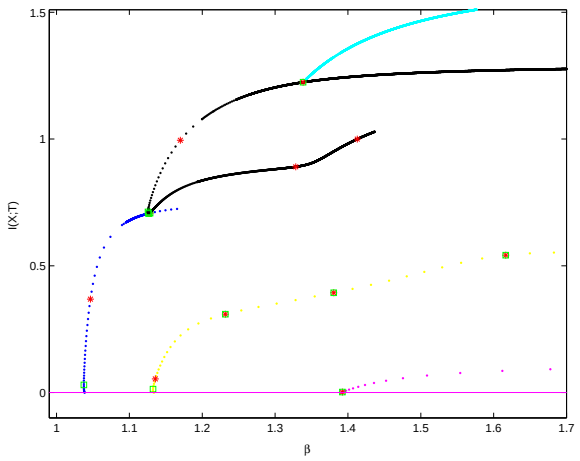

Fig. 6. The bifurcations of stationary points to the Information Distortion problem (7) which exhibit symmetry breaking from $S_4 \rightarrow S_3 \rightarrow S_2 \rightarrow S_1$ (color scheme is purple $\rightarrow$ blue $\rightarrow$ black $\rightarrow$ cyan), for which Figure 4 only shows solutions.

B. Results in this contribution

For any annealing problem of the form (5) that satisfies some regularity conditions, this paper answers many of the questions just posed about the bifurcations.

1) There are $N-1$ symmetry breaking bifurcations observed when continuing from the initial solution $q_{\frac{1}{N}}$ because there are only $N-1$ subgroups in the symmetry breaking chain from $S_N \rightarrow S_1$ (Theorem 6.2), for example $S_N \rightarrow S_{N-1} \rightarrow \ldots \rightarrow S_2 \rightarrow S_1$.
2) The annealing solutions in Figure 4 all have symmetry S_M for some $M \leq N$. There exist other branches with symmetry $S_m \times S_n$ when $m+n=N$ (Figure 6 and Theorem 6.2). In the Four Blob problem, these solutions are suboptimal since they yield mutual information values below the envelope curve depicted in the figure.
3) Symmetry breaking bifurcations are generically pitchforks (Theorem 6.3) and derivative calculations predict whether the bifurcating branches are subcritical or supercritical (Theorem 6.5), as well as determine optimality (Theorem 6.7). Symmetry preserving bifurcations are generically saddle-nodes (Theorem 6.9).
4) The relationship between the bifurcations of solutions to the optimization problem (3) and the annealing problem (5) is given in Figures 4 and 5. The Lagrange multiplier β is a function of I_0 for $\beta > 0$: turning Figure 4 sideways shows this functional relationship. In fact, the bifurcations of all stationary points to (3) is much more complicated (see Figure 17). The curve $R(I_0)$ in Figure 5 is non-increasing and continuous (Lemma 4.2) and envelopes $R(I_0|q)$ over all stationary points q of (3). Any curve below the envelope corresponds to clusterings of the data which are not solutions of the optimization problem (3).
5) A local solution to the annealing problem (5) does not always continue through a symmetry breaking bifurcation (Theorem 8.1). This would explain why, in practice, solving (5) after bifurcation incurs significant computational cost [12], [14]. A solution of the annealing problem (5) is always a solution of the original optimization problem (3). The converse is not true.
6) Bifurcations of solutions to the annealing problem (5) dictate the convexity of the curve (3) (Lemma 8.2). In particular, a subcritical bifurcation of the annealing solutions to (5) at $\beta(I_0^*)$ implies that the curve $R(I_0)$ changes convexity in a neighborhood of I_0^* (Corollary 8.3). This can be compared to the rate distortion curve in information theory,

$$R_{RD}(I_0) := \begin{array}{c} \min_{q \in \Delta} \boldsymbol{I}(\boldsymbol{Y};\boldsymbol{T}) \\ D(q) \geq I_0 \end{array} .$$

When $D(q)$ is linear in q, then the rate distortion curve is non-increasing, convex, and continuous [8], [13]. This convexity result does not generalize to either the Information Bottleneck (1) or the Information Distortion (2) since $D(q)$, in both these cases, is not linear, although both of these curves, under mild regularity conditions, are non-increasing and continuous (Lemma 4.2).

IV. Mathematical Preliminaries

This section is divided into four parts. First, we define notations used throughout the rest of this paper. Secondly, we present some key results from the theory of constrained optimization. In the third part we apply the theory to the optimization problem (3) and the corresponding annealing problem (5). And finally, we give a primer on bifurcation theory in the presence of symmetries.

A. Notation

Let $|\boldsymbol{Y}| = K < \infty$ and $\boldsymbol{T} \in \{1, ..., N\}$ so that $|\boldsymbol{T}| = N < \infty$. There is no further restriction placed on N (i.e. N can be larger than K). Recall that the $K \times N$ matrix defining the conditional probability mass function of the

random variable $\boldsymbol{T}|\boldsymbol{Y}$, is

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

where $(q^\nu)^T := q(\boldsymbol{T} = \nu|\boldsymbol{Y})$ is the $1 \times K$ row of $q(\boldsymbol{T}|\boldsymbol{Y})$. The following notation will also be used throughout the rest of this contribution:

- $\boldsymbol{x}^\nu$:= the ν^{th} $K \times 1$ vector component of $\boldsymbol{x} \in \Re^{NK}$, so that
$$\boldsymbol{x} = ((\boldsymbol{x}^1)^T \ \ (\boldsymbol{x}^2)^T \ \ ... \ \ (\boldsymbol{x}^N)^T)^T.$$
- q := the vector form of $q(\boldsymbol{T}|\boldsymbol{Y})^T$,
$$q = ((q^1)^T \ \ (q^2)^T \ \ ... \ \ (q^N)^T)^T.$$
- $q_{\nu k}$:= $q(\boldsymbol{T} = \nu|\boldsymbol{Y} = y_k)$.
- $q_{\frac{1}{N}}$:= the uniform conditional probability on $\boldsymbol{T}|\boldsymbol{Y}$ such that $q_{\frac{1}{N}}(\boldsymbol{T}|\boldsymbol{Y}) = \frac{1}{N}$ for every $\boldsymbol{T}$ and $\boldsymbol{Y}$.
- I_n := $n \times n$ identity matrix when $n > 0$.
- $\nabla_x f$:= the gradient of a differentiable scalar function $f(x)$ with respect to the vector argument x.
- $d_x^n f$:= the multilinear form of the n dimensional array of the n^{th} derivatives of the scalar function f. For example, $d^3 f[y_1, y_2, y_3] = \sum_{i,j,k} \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}[y_1]_i[y_2]_j[y_3]_k$

B. The Two Optimization Problems

In section III, we considered two different constrained optimization problems, a problem with a nonlinear constraint (3)

$$\max_{q \in \Delta} G(q) \quad D(q) \geq I_0$$

and the annealing problem (5)

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

Let us compare the respective Lagrangians, and the necessary and sufficient conditions for optimality for each of these problems.

The equality constraints from the optimization problem (3) and the annealing problem (5) are the same:

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

Assigning Lagrange multipliers $\{\lambda_k\}_{k=1}^K$ to the K equality constraints (β is an annealing parameter), the Lagrangian $\mathcal{L}(q, \lambda, \beta)$ for the annealing problem (5) with respect to the equality constraints is

$$G(q) + \beta D(q) + \sum_{k=1}^K \lambda_k \left(\sum_{\nu=1}^N q_{\nu k} - 1 \right). \tag{8}$$

Thus, $\lambda \in \Re^K$ is the vector of Lagrange multipliers $\lambda = (\lambda_1, \lambda_2, ..., \lambda_K)^T$. The gradient of the Lagrangian is

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

where $\nabla_q \mathcal{L} = \nabla_q G + \beta \nabla_q D(q) + \Lambda$ and $\Lambda = \begin{pmatrix} \lambda^T & \lambda^T & ... & \lambda^T \end{pmatrix}^T \in \Re^{NK}$. The gradient $\nabla_\lambda \mathcal{L}$ is a vector of the K equality constraints

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

Since we only consider equality constraints, the first order necessary conditions for optimality, the Karush-Kuhn-Tucker (KKT) conditions [49], are satisfied at $(q^*, \lambda^*, \beta^*)$ if and only if $\nabla_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta^*) = \mathbf{0}$. A soft clustering $q^* \in \Delta$ is a stationary point of the annealing problem (5) for some β^* if there exists a vector λ^* such that $\nabla_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta^*) = \mathbf{0}$ for the Lagrangian $\mathcal{L}$ defined in (8).

The Jacobian of the constraints for the annealing problem is

$$J_1 = d_q \nabla_\lambda \mathcal{L} = d_q \begin{pmatrix} \sum_\nu q_{\nu 1} - 1 \\ \sum_\nu q_{\nu 2} - 1 \\ \vdots \\ \sum_\nu q_{\nu K} - 1 \end{pmatrix} = \underbrace{\begin{pmatrix} I_K & I_K & ... & I_K \end{pmatrix}}_{N \text{ blocks}},$$

which has full row rank. Since the constraints are linear, then a stationary point is a solution of the annealing problem (5) if $d^2(G(q^*) + \beta D(q^*))$ is negative definite on $\ker J_1$ [49].

Only the optimization problem (3) is encumbered with the non-linear constraint $D(q) - I_0 \geq 0$. Assigning the Lagrange multiplier β to this constraint, we see that the Lagrangian in this case is

$$\hat{\mathcal{L}} = G(q)) + \beta(D(q) - I_0) + \sum_{k=1}^K \lambda_k \left(\sum_{\nu=1}^N q_{\nu k} - 1 \right).$$

This shows that the gradient of the Lagrangian is the same for the optimization problem (3) and the annealing problem (5), $\nabla_{q,\lambda} \mathcal{L} = \nabla_{q,\lambda} \hat{\mathcal{L}}$.

The Jacobian of the constraints for the optimization problem (3) is

$$J_2(q) := d_q \nabla_{\lambda,\beta} \hat{\mathcal{L}} = d_q \begin{pmatrix} \sum_\nu q_{\nu 1} - 1 \\ \sum_\nu q_{\nu 2} - 1 \\ \vdots \\ \sum_\nu q_{\nu K} - 1 \\ D(q) - I_0 \end{pmatrix} = \begin{pmatrix} J_1 \\ \nabla D(q)^T \end{pmatrix}.$$

which is a function of q, and, for generic $D(q)$, of full row rank. By the theory of constrained optimization, a stationary point q^* of the annealing problem (5) is a local solution of (3) for some I_0 if $d^2(G(q^*)+\beta D(q^*))$ is negative definite on $\ker J_2$ [49].

If $(q^*, \lambda^*, \beta^*)$ is a solution of the optimization problem (3) for some I_0, then by the KKT conditions, β^* is unique and non-negative. This shows that the optimal β can be written as a function of I_0, $\beta(I_0)$. For $\beta^* > 0$, the KKT conditions are satisfied at $(q^*, \lambda^*, \beta^*)$ if and only if $\nabla_{q,\lambda,\beta} \hat{\mathcal{L}}(q^*, \lambda^*, \beta^*) = \mathbf{0}$. That is, the constraint $D(q^*) - I_0$ is active and equal to zero. Thus, if $(q^*, \lambda^*, \beta^*)$ is a stationary point of the annealing problem (5) for $\beta^* > 0$, then for $I_0 = D(q^*)$, $(q^*, \lambda^*, \beta^*)$ satisfies the KKT conditions for the optimization problem (3).

We have just proved the following theorem.

Theorem 4.1: Suppose that $q^* \in \Delta$ is a stationary point of the annealing problem (5) for some $\beta > 0$ such that $J_2(q^*)$ has full row rank.

1) If $d^2(G(q^*) + \beta D(q^*))$ is negative definite on $\ker J_1$ then q^* is a solution of (3) (for $I_0 = D(q^*)$) and (5).
2) If $d^2(G(q^*) + \beta D(q^*))$ is negative definite on $\ker J_2(q^*)$, then q^* is a solution of (3) for $I_0 = D(q^*)$.
3) Conversely, if q^* is a local solution of (5) for some β^*, then there exists a vector of Lagrange multipliers λ^* so that $\nabla_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta^*) = \mathbf{0}$ and $d^2(G(q^*) + \beta^* D(q^*))$ is non-positive definite on $\ker J_1$.
4) If q^* is a solution of (3) for some I_0, then there exists a vector of Lagrange multipliers (λ^*, β^*) so that $\nabla_{q,\lambda} \mathcal{L}(q^*, \lambda^*, \beta^*) = \mathbf{0}$ and $d^2(G(q^*) + \beta^* D(q^*))$ is non-positive definite on $\ker J_2(q^*)$.

The fact that every solution of the annealing problem (5) is also a solution of the optimization (3) follows from the observation that $\ker J_1$ contains $\ker J_2(q^*)$: if $\boldsymbol{w}$ satisfies $J_2(q^*)\boldsymbol{w} = \mathbf{0}$, then $J_1 \boldsymbol{w} = \mathbf{0}$. However, there may be solutions of (3) which are not annealing solutions of (5). This is illustrated numerically for the Information Distortion problem (7) in section VIII-A.

Now let us consider for what values of I_0 the optimization problem (3) has a solution. Clearly, one necessary condition is that $I_0 \leq \max_{q \in \Delta} D(q) = I_{\max}$. In fact, $R(I_0)$ is a non-increasing curve, and, when defined as in (1) or (2), continuous. This is what we prove next.

Lemma 4.2: The curve $R(I_0)$ is non-increasing on $I_0 \in [0, I_{\max}]$, and is continuous if the stationary points q^* of $G(q)$ (i.e. $\nabla_q G(q^*) = \mathbf{0}$) are not in $\mathcal{Q}_{I_0}$ for $I_0 > 0$, where

$$\mathcal{Q}_{I_0} := \{q \in \Delta \mid D(q) \geq I_0\}$$

.

Proof: If $I_1 \geq I_2$, then $\mathcal{Q}_{I_1} \subset \mathcal{Q}_{I_2}$, which shows that $R(I_1) \leq R(I_2)$. To prove continuity, take an arbitrary $I_0 \in (0, I_{\max})$. Let

$$M_{I_0} := \{y \mid y = G(q) \text{ where } q \in \mathcal{Q}_{I_0}\}$$

be in the range (in $\Re$) of the function $G(q)$ with the domain $\mathcal{Q}_{I_0}$. Given an arbitrary $\epsilon > 0$, let $M^\epsilon_{I_0}$ be an ϵ neighborhood of M_{I_0} in $\Re$. By assumption (4), $G(q)$ is continuous on Δ, and so the set $G^{-1}(M^\epsilon_{I_0})$ is a relatively open set in Δ. Because by definition $G(\mathcal{Q}_{I_0}) = M_{I_0}$, we see that

$$\mathcal{Q}_{I_0} \subset G^{-1}(M^\epsilon_{I_0}). \tag{9}$$

Furthermore, since $\nabla G(q) \neq 0$ for $q \in \mathcal{Q}_{I_0}$, then, by the Inverse Mapping Theorem, $G^{-1}(M^\epsilon_{I_0})$ is an open neighborhood of $\mathcal{Q}_{I_0}$.

The function $D(q)$ is also continuous in the interior of Δ. Observe that $\mathcal{Q}_{I_0} = D^{-1}([I_0, I_{\max}])$ is closed, and thus $\mathcal{Q}_{I_0}$ is closed and hence compact. Thus, by (9) $G^{-1}(M^\epsilon_{I_0})$ is a relatively open neighborhood of a compact set $\mathcal{Q}_{I_0}$. Therefore, since $D(q)$ is continuous, there exists a $\delta > 0$ such that the set

$$\mathrm{Int}\mathcal{Q}_{I_0+\delta} = D^{-1}((I_0 + \delta, I_{\max}]) \cap \Delta$$

is a relatively open set in Δ such that

$$\mathcal{Q}_{I_0} \subset \mathrm{Int}\mathcal{Q}_{I_0+\delta} \subset G^{-1}(M^\epsilon_{I_0}).$$

It then follows that

$$|\max_{\mathcal{Q}_{I_0+\delta}} G(q) - \max_{\mathcal{Q}_{I_0}} G(q)| < \epsilon.$$

which means that

$$|R(D(q)) - R(I_0)| < \epsilon \text{ whenever } D(q) - I_0 < \delta.$$

□

C. An Overview of Bifurcation Theory with Symmetries

In this section, the general terminology and concepts related to studying bifurcations of dynamical systems with symmetries is reviewed. The dynamical system we will study, whose equilibria are stationary points of the optimization problem (3)), in the sequel is the gradient flow of the Lagrangian. For a detailed treatment, see Golubitsky et al. in [45].

Consider the system of ordinary differential equations

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

where $f : V \times \Re \rightarrow V$ is sufficiently smooth for some Banach space V, $\boldsymbol{x} \in V$, and $\beta \in \Re$ is a bifurcation parameter. An *equilibrium* or *steady state* of the differential equation is a zero of f. An equilibrium (x, β) is *linearly stable* if all of the eigenvalues of the Jacobian, $d_{\boldsymbol{x}} f(x, \beta)$, have a negative real part. If some eigenvalue has a positive real part, then the equilibrium is *unstable*. A *bifurcation point* is an equilibrium $(\boldsymbol{x}^*, \beta^*)$ where the number of equilibria changes as β varies in a neighborhood of β^*. At a bifurcation, the Jacobian $d_{\boldsymbol{x}} f(\boldsymbol{x}^*, \beta^*)$ is singular, (i.e. $d_{\boldsymbol{x}} f(\boldsymbol{x}^*, \beta^*)$ has a zero eigenvalue). Otherwise, the Implicit Function Theorem could be used to find a unique solution $\boldsymbol{x}(\beta)$ in a neighborhood of $(\boldsymbol{x}^*, \beta^*)$. The bifurcating directions are in the kernel of the Jacobian, defined as

$$\ker d_{\boldsymbol{x}} f(\boldsymbol{x}^*, \beta^*) := \{\boldsymbol{w} \in V : d_{\boldsymbol{x}} f(\boldsymbol{x}^*, \beta^*)\boldsymbol{w} = \boldsymbol{0}\}.$$

An equilibrium $(\boldsymbol{x}^*, \beta^*)$ is a *singularity* of f if $d_{\boldsymbol{x}} f(\boldsymbol{x}^*, \beta^*)$ is singular. A singularity is a possible bifurcation point, since it satisfies the necessary condition for a bifurcation.

Let Γ be a compact Lie group which acts on V (S_N is a specific case of such a group). The vector function f is Γ-invariant if

$$f(\boldsymbol{x}, \beta) = f(\gamma \boldsymbol{x}, \beta)$$

for every $\gamma \in \Gamma$. f is Γ-equivariant if

$$\gamma f(\boldsymbol{x}, \beta) = f(\gamma \boldsymbol{x}, \beta)$$

for every $\gamma \in \Gamma$. The *isotropy subgroup* $\Sigma \subseteq \Gamma$ of $\boldsymbol{x} \in V$ is defined as

$$\Sigma := \{\gamma \in \Gamma : \gamma \boldsymbol{x} = \boldsymbol{x}\}.$$

In other words, $\boldsymbol{x}$ has symmetry Σ. The *fixed point space* of a subgroup $\Sigma \subseteq \Gamma$ is

$$\mathrm{Fix}(\Sigma) := \{\boldsymbol{x} \in V : \gamma \boldsymbol{x} = \boldsymbol{x} \quad \text{for every} \quad \gamma \in \Sigma\}.$$

A *symmetry breaking* bifurcation is a bifurcation for which the isotropy group of the bifurcating equilibria is a proper subgroup of the group which fixes the bifurcation point. A *symmetry preserving* bifurcation is one for which the symmetry of the bifurcating equilibria is the same as the group which fixes the bifurcation point.

The Equivariant Branching Lemma, attributed to Vanderbauwhede [50] and Cicogna [51], [52], relates the subgroup structure of Γ with the existence of symmetry breaking bifurcating branches of equilibria of $\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \beta)$. For a proof see [45] p.83.

Theorem 4.3 (Equivariant Branching Lemma): Let f be a smooth function, $f : V \times \Re \rightarrow V$ which is Γ-equivariant for a compact Lie group Γ, and a Banach space V. Let Σ be an isotropy subgroup of Γ with $\dim(\mathrm{Fix}(\Sigma)) = 1$. Suppose that $\mathrm{Fix}(\Gamma) = \{\boldsymbol{0}\}$, the Jacobian $d_{\boldsymbol{x}} f(\boldsymbol{0}, 0) = \boldsymbol{0}$, and the crossing condition $d^2_{\beta \boldsymbol{x}} f(\boldsymbol{0}, 0)\boldsymbol{x}_0 \neq 0$ is satisfied for $\boldsymbol{x}_0 \in \mathrm{Fix}(\Sigma)$. Then there exists a unique smooth solution branch $(t\boldsymbol{x}_0, \beta(t))$ to $f = 0$ with isotropy subgroup Σ.

Remark 4.4: For an arbitrary Γ-equivariant system where bifurcation occurs at $(\boldsymbol{x}^*, \beta^*)$, the requirement in Theorem 4.3 that the bifurcation occurs at the origin is accomplished by a translation. Assuring that the Jacobian vanishes, $d_{\boldsymbol{x}} f(\boldsymbol{0}, 0) = \boldsymbol{0}$, can be effected by restricting and projecting the system onto the kernel of the Jacobian. This transform is called the Liapunov-Schmidt reduction (see [53]).

Definition 4.5: The branch $(t\boldsymbol{x}_0, \beta(t))$ is *transcritical* if $\beta'(0) \neq 0$. If $\beta'(0) = 0$ then the branch is *degenerate*. If $\beta'(0) = 0$ and $\beta''(0) \neq 0$ then the branch is a *pitchfork*. The branch is *subcritical* if for all nonzero t such that $|t| < \epsilon$ for some $\epsilon > 0$, $t\beta'(t) < 0$. The branch is *supercritical* if $t\beta'(t) > 0$.

Subcritical bifurcations are sometimes called *first order phase transitions* or *jump* bifurcations. Supercritical bifurcations are also called *second order phase transitions*.

An Example: Pitchforks and Saddle-nodes: To illustrate some of the concepts just introduced, let us consider the following Z_2-equivariant differential equation

$$\dot{x} = f(x, \beta) = \beta x + x^3 - x^5$$

whose equilibria are shown as a function of β in Figure 7 (see also [54]). This simple problem illustrates both types of bifurcations which we expect to see for any S_N-equivariant annealing problem of the form (5) such that (4) holds.

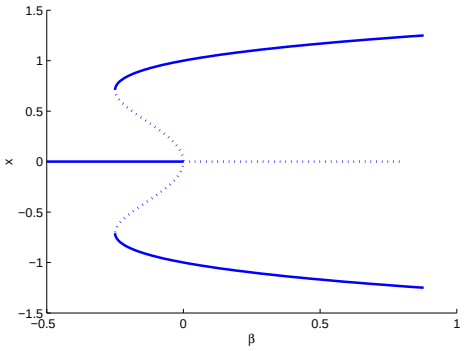

Fig. 7. The bifurcation diagram of equilibria of $\dot{x} = f(x, \beta) = \beta x + x^3 - x^5$. A subcritical pitchfork bifurcation occurs at $(x = 0, \beta = 0)$, and saddle-node bifurcations occur at $(\pm\sqrt{\frac{1}{2}}, -\frac{1}{4})$. The branches drawn with dots are composed of unstable equilibria, and the branches drawn with a solid line are composed of stable equilibria.

The group $Z_2 = \{-1, 1\}$ acts on a scalar by multipli-

cation by either -1 or 1. Equivariance is established since $-f = f(-x, \beta)$. For all β, $x = 0$ is an equilibrium. Since $d_x f = \beta + 3x^2 - 5x^4$, then $(x = 0, \beta = 0)$ is a singularity. Observe that $x = 0$ is the only scalar invariant to the action of Z_2 (i.e. $\mathrm{Fix}(Z_2) = \{0\}$) and $\Sigma = \{1\} \subset Z_2$ is an isotropy subgroup with a one dimensional fixed point space, $\mathrm{Fix}(\Sigma) = \Re$. Since the crossing condition $d^2_{\beta x} f(0,0) = 1 \neq 0$ is satisfied, then the Equivariant Branching Lemma gives the existence of a bifurcating solution emanating from $(x = 0, \beta = 0)$, with direction $\boldsymbol{x}_0 = 1$. Parameterizing the bifurcating branch as $(t, \beta(t))$, we have that

$$\beta(t) = t^4 - t^2$$

for $t \neq 0$. As a consequence of the Z_2 symmetry, we actually have two bifurcating branches, one for positive t, and one for negative t. Since $\beta'(0) = 0$, then the bifurcation at the origin is degenerate, and $\beta''(0) = -2$ implies that the bifurcation is in fact a subcritical pitchfork bifurcation.

The bifurcating branches emanating from the origin are unstable since the Jacobian $d_x f(x, \beta) < 0$ for all $|x| < \sqrt{\frac{1}{2}}$ and $\beta < 0$. As $|x|$ increases, the higher order quintic term of f eventually dominates and causes the branches to turn around and become stable at the *saddle-node* bifurcations at $(x = \pm\sqrt{\frac{1}{2}}, \beta = -\frac{1}{4})$.

The methodology we have applied in this simple example is how we will proceed to analyze bifurcations of stationary points to arbitrary annealing problems of the form (5) when (4) holds.

V. Symmetries

Why do the optimization problem (3) and the annealing problem (5) have symmetry? How can we capitalize on this symmetry to solve these problems? These are the questions which are addressed in this section.

The symmetries of the optimization problems (3) and (5) arise from the structure of $q \in \Delta$ and from the form of the functions $G(q)$ and $D(q)$ given in (4): permuting the sub-vectors q^ν does not change the value of G and D: this is the symmetry, S_N-invariance.

We will capitalize upon the symmetry of S_N by using the Equivariant Branching Lemma to determine the bifurcations of stationary points, which includes local annealing solutions, to (5)

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

As we pointed out in section IV-B, this also yields the bifurcation structure of stationary points of the optimization problem (3) with respect to I_0.

In this section we lay the groundwork necessary to give the bifurcation structure for a larger class of constrained optimization problems of the form

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

as long as F satisfies the following:

Assumption 5.1: The function $F(q, \beta)$ is of the form

$$F(q, \beta) = \sum_{\nu=1}^{N} f(q^\nu, \beta)$$

for some smooth scalar function f, where the vector $q \in \Delta \subset \Re^{NK}$ is decomposed into N sub-vectors $q^\nu \in \Re^K$.

The annealing problem (5) satisfies Assumption 5.1 when

$$F(q, \beta) = G(q) + \beta D(q), \tag{10}$$

and G and D are of the form given in (4). This includes the Information Bottleneck problem (6), and the Information Distortion problem (7).

It is straightforward to verify that any F satisfying Assumption 5.1 has the following properties.

1) F is S_N-invariant, where the action of S_N on q permutes the sub-vectors q^ν of q.
2) The $NK \times NK$ Hessian $d^2 F$ is block diagonal, with N $K \times K$ blocks.

The rest of this section is divided into three parts. In the first part, we define the gradient flow of the Lagrangian, whose equilibria are stationary points to the annealing problem (5), and show how the symmetries manipulate the form of its Jacobian (i.e. the Hessian of the Lagrangian). Secondly, we classify the equilibria of the gradient flow according to their symmetries. Thirdly, we give a detailed description of the kernel of the Hessian at a bifurcation. This space is determined by considering the reduced problem: one only needs to compute the one dimensional kernel of a single block of $d^2 F(q^*)$. The form of the larger kernel, as well as the many bifurcating directions, follows from applying the symmetries.

A. The Gradient Flow

We now formulate a dynamical system whose equilibria correspond to the stationary points of the annealing problem (5). This system is the gradient flow of the Lagrangian.

With $F(q, \beta) = G(q) + \beta D(q)$ as in (10) such that G and D satisfy (4), the Lagrangian of the annealing problem (5), which we derived in (8), can be written as

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

The gradient of the Lagrangian is

$$\nabla \mathcal{L} := \nabla_{q,\lambda} \mathcal{L}(q, \lambda, \beta) = \begin{pmatrix} \nabla_q \mathcal{L} \\ \nabla_\lambda \mathcal{L} \end{pmatrix} = \begin{pmatrix} \nabla_q F + \Lambda \\ \nabla_\lambda \mathcal{L} \end{pmatrix},$$

where $\Lambda = (\lambda^T \ \lambda^T \ ... \ \lambda^T)^T$. The $(NK+K)\times(NK+K)$ Hessian of the Lagrangian is

$$d^2\mathcal{L}(q) := d^2_{q,\lambda}\mathcal{L}(q,\lambda,\beta) = \begin{pmatrix} d^2F(q,\beta) & J_1^T \\ J_1 & \mathbf{0} \end{pmatrix}, \quad (11)$$

where $\mathbf{0}$ is $K \times K$. The $NK \times NK$ matrix d^2F is the block diagonal Hessian of F,

$$d^2F(q) := d^2_qF(q,\beta) = \begin{pmatrix} B_1 & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & B_2 & ... & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & ... & B_N \end{pmatrix},$$

where $B_\nu = d^2f(q^\nu,\beta)$ (see Assumption 5.1) are $K\times K$ matrices for $\nu = 1,...,N$.

The dynamical system whose equilibria are stationary points of the optimization problem (3) and the annealing problem (5) can now be posed as the gradient flow of the Lagrangian

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla\mathcal{L}(q,\lambda,\beta) \quad (12)$$

for $\beta \in [0,\infty)$. Recall that equilibria of (12) are points $\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} \in \Re^{NK+K}$ where

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

The Jacobian of this system is the Hessian $d^2\mathcal{L}(q,\lambda,\beta)$ from (11).

The methodology we applied to the simple example in IV-C is how we will proceed to analyze bifurcations of equilibria of the gradient flow (12). The Equivariant Branching Lemma gives the existence of branches of equilibria at symmetry breaking bifurcations. At such a bifurcation, we will show that $\beta'(0) = 0$, so that the bifurcations are degenerate. When $\beta''(0) \neq 0$, then the bifurcations are pitchforks, and the sign of $\beta''(0)$ determines whether the bifurcating branches are subcritical or supercritical. We will determine the stability of these equilibria by considering the eigenvalues of the Hessian $d^2\mathcal{L}(q,\lambda,\beta)$.

Yet, by Theorem 4.1, it is the Hessian $d^2F(q^*,\beta)$ which determines whether a given equilibrium is a solution of the optimization problem (3) or of the annealing problem (5). We will show how stability relates to optimality in the optimization problems (3) and (5) in section VIII-A.

B. Equilibria with Symmetry

Next, we categorize the equilibria of the gradient flow (12) according to their symmetries, which allows us to determine when we expect symmetry breaking bifurcations versus symmetry preserving bifurcations.

Recall that $q = ((q^1)^T,...,(q^N)^T)^T$ is the vector form of the soft clustering $q(\boldsymbol{T}|\boldsymbol{Y})$ of the responses $\boldsymbol{Y}$ into the classes $\boldsymbol{T} = \{1,\dots,N\}$. Let $\{\mathcal{U}_i\}_{i=1}^I$ be a partition of the classes of $\boldsymbol{T}$ such that $q^\nu = q^\eta$ if and only if $\nu,\eta \in \mathcal{U}_i$. That is, $\mathcal{U}_j \cap \mathcal{U}_k = \emptyset$ for $j \neq k$ and $\bigcup_{i=1}^I \mathcal{U}_i = \{1,...,N\}$. If $M_i := |\mathcal{U}_i|$ is the order of $\mathcal{U}_i$ (so that $\sum_{i=1}^I M_i = N$), then we have that q has isotropy group

$$S_{M_1} \times S_{M_2} \times ... \times S_{M_I},$$

where S_{M_i} acts on q by permuting the vector subcomponents q^ν for every $\nu \in \mathcal{U}_i$. For example, in bottom panel 2 of Figure 4, $N = 4$, $\mathcal{U}_1 = \{1\}$, $\mathcal{U}_2 = \{2,3,4\}$, and $q^2 = q^3 = q^4$. So q has isotropy subgroup $S_1 \times S_3$, or, more simply, S_3. In panels 3, 4 and 5, $\mathcal{U}_1 = \{1\}$, $\mathcal{U}_2 = \{2,4\}$ and $\mathcal{U}_3 = \{3\}$, and the associated clustering q has isotropy group S_2. It is clear from Assumption 5.1 that if $q^\nu = q^\eta$, then $d^2f(q^\nu) = d^2f(q^\eta)$: the ν^{th} and η^{th} blocks of $d^2F(q)$ are equal. So, $d^2F(q)$ has M_i blocks, B_ν for $\nu \in \mathcal{U}_i$, that are equal for each i.

Suppose that (q^*,λ^*,β^*) is a singularity such that q^* has isotropy group $S_{M_1} \times S_{M_2} \times ... \times S_{M_I}$. By definition, $d^2\mathcal{L}(q^*)$ is singular. Additionally, only one of the following is also true:

1) $d^2F(q^*)$ is singular.
2) $d^2F(q^*)$ is nonsingular.

In the first case we expect to get a symmetry breaking bifurcation (Theorem 6.2). In the second case we get a symmetry preserving bifurcation (Theorem 6.9).

Let us investigate case 1 and assume that $d^2\mathcal{L}(q^*)$ is singular, and that $d^2F(q^*)$ is singular, with only M_i singular blocks B_ν for $\nu \in \mathcal{U}_i$. To ease the notation we set

$$\mathcal{U} := \mathcal{U}_i, \quad M := M_i \ \text{ and } \ \mathcal{R} := \bigcup_{j\neq i} \mathcal{U}_j.$$

To distinguish between singular blocks B_ν, $\nu \in \mathcal{U}$ and non-singular blocks B_ν, $\nu \in \mathcal{R}$. We will write

$$B := B_\nu \text{ for } \nu \in \mathcal{U}. \quad (13)$$

The type of symmetry breaking bifurcation we get from a singular equilibrium (q^*,λ^*,β^*) only depends on M, the number of blocks B which are singular. This motivates the following definition.

Definition 5.2: An equilibrium (q^*,λ^*,β^*) of the gradient flow (12) is *M-singular* (or, equivalently, q^* is M-singular) if:

1) $|\mathcal{U}| = M$.
2) $q^\nu = q^\eta$ for every $\nu,\eta \in \mathcal{U}$ (i.e. $q \in \text{Fix}(S_M)$).
3) For B, the M block(s) of the Hessian defined in (13),

$$\ker B \text{ has dimension 1 with basis vector } \boldsymbol{v} \in \Re^K \quad (14)$$

4) The $N-M$ block(s) of the Hessian $\{B_\nu\}_{\nu\in\mathcal{R}}$ are nonsingular.
5) The matrix

$$A := B\sum_{\nu\in\mathcal{R}} B_\nu^{-1} + MI_K \tag{15}$$

is nonsingular. When $M = N$, $\mathcal{R}$ is empty, and in this case we define $A = NI_K$.

We wish to emphasize that when $d^2F(q^*)$ is singular, that the requirements 3-5 in Definition 5.2 hold generically [31]. The technical requirement 5 is crucial for a symmetry breaking bifurcation to occur. We will see later that the matrix A becomes singular at symmetry preserving bifurcations.

From Assumption 5.1, it is clear that $\mathcal{L}(q,\lambda,\beta)$ and F are S_N-invariant, and that $\nabla\mathcal{L}(q,\lambda,\beta)$ and ∇F are S_N-equivariant.

It is straightforward to show that every block of the Hessian of the Information Bottleneck cost function (6) is always singular. At a bifurcation point (q^*,λ^*,β^*) which is in $\mathrm{Fix}(S_M)$, the M blocks of $d^2F(q^*) = d^2\left(-\boldsymbol{I}(\boldsymbol{Y};\boldsymbol{T}) + \beta\boldsymbol{I}(\boldsymbol{X};\boldsymbol{T})\right)$ referred to in requirement 3 of Definition 5.2 have a two dimensional kernel, requirement 4 is not met, and the matrix A in requirement 5 is not even defined. A similar theory to that presented here, which projects out the "perpetual kernel," explains the bifurcation structure of solutions for the Information Bottleneck problem (6). Some details will be discussed in section VIII-B.

C. The Kernel of the Hessian $d^2\mathcal{L}(q^*)$

Here, we see how the symmetry of q^* and F eases the computation of multiple equilibria (q^*,λ^*,β^*) of the gradient system (12) at a bifurcation. As reviewed in section IV-C, the Jacobian $d^2\mathcal{L}(q^*)$ from (11) is singular, and the bifurcating branches are tangent to $\ker d^2\mathcal{L}(q^*) \subset \Re^{NK+K}$. To describe these bifurcating branches when q^* is M-singular, we need only work with a reduced space, the kernel of B from (14), which is a one dimensional subspace of $\Re^K$ with basis vector $\boldsymbol{v}$. By the symmetry, this one vector explicitly determines the larger spaces $\ker d^2F(q^*)$ and $\ker d^2\mathcal{L}(q^*)$ (Theorem 5.3), and yields the bifurcating branches (Lemma 5.5).

Intuitively, it is the vector $\boldsymbol{v}\in\ker(B)$ which specifies how each of the K responses of $\boldsymbol{Y}$ ought to split at a bifurcation in order to increase the value of F on Δ. It is the symmetry which specifies how the responses are explicitly assigned to the classes, and these assignments are the bifurcating directions.

We first determine a basis for $\ker d^2F(q^*)$ at an M-singular q^*. Recall that in the preliminaries, when $\boldsymbol{x}\in\Re^{NK}$, we defined $\boldsymbol{x}^\nu\in\Re^K$ to be the ν^{th} vector component of $\boldsymbol{x}$. Using this notation, the linearly independent vectors $\{\boldsymbol{v}_i\}_{i=1}^M$ in $\Re^{NK}$ can be defined by

$$\boldsymbol{v}_i^\nu := \begin{cases} \boldsymbol{v} \text{ if } \nu \text{ is the } i^{th} \text{ uniform class of } \mathcal{U} \\ \mathbf{0} \text{ otherwise} \end{cases} \tag{16}$$

where $\mathbf{0}\in\Re^K$. Since q^* is M-singular, then $\dim(\ker d^2F(q^*)) = M$, which implies that $\{\boldsymbol{v}_i\}_{i=1}^M$ is a basis for $\ker d^2F(q^*)$. For example, consider the bifurcation where symmetry breaks from S_3 to S_2 in Figure 4 (see panels 2 and 3 in the bottom row). At this bifurcation, $M = N-1 = 3$, $\mathcal{R} = \{1\}$, $\mathcal{U} = \{2,3,4\}$, and $d^2F(q^*)$ is three dimensional with basis vectors

$$\boldsymbol{v}_1 := (\mathbf{0},\boldsymbol{v}^T,\mathbf{0},\mathbf{0})^T,\quad \boldsymbol{v}_2 := (\mathbf{0},\mathbf{0},\boldsymbol{v}^T,\mathbf{0})^T,$$
$$\boldsymbol{v}_3 := (\mathbf{0},\mathbf{0},\mathbf{0},\boldsymbol{v}^T)^T$$

where $\mathbf{0}$ is $1\times K$.

The basis vectors of $\ker d^2F(q^*)$ can be used to construct a basis for $\ker d^2\mathcal{L}(q^*)$ when $M>1$. Let

$$\boldsymbol{w}_i = \begin{pmatrix} \boldsymbol{v}_i \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{v}_M \\ \mathbf{0} \end{pmatrix} \tag{17}$$

for $i = 1,...,M-1$ where $\mathbf{0}\in\Re^K$. Using (11), it is easy to see that $d^2\mathcal{L}(q^*)\boldsymbol{w}_i = \mathbf{0}$, which shows that $\{\boldsymbol{w}_i\}$ are in $\ker d^2\mathcal{L}(q^*)$. Thus, if $d^2F(q^*)$ is singular and q^* is M-singular for $1<M\le N$, then $d^2\mathcal{L}(q^*)$ is singular.

The fact that the vectors $\{\boldsymbol{w}_i\}$ are linearly independent is straightforward to establish. To show that they actually span $\ker d^2\mathcal{L}(q^*)$ (and so are a basis) relies on the assumption that q^* is M-singular, which assures that the matrix $A = B\sum_\nu B_\nu^{-1} + MI_K$, introduced in Definition 5.2, is nonsingular.

We have the following Theorem. The proof of the first two parts is above, and a proof of the third part can be found in [31].

Theorem 5.3: If q^* is M-singular for some $1<M\le N$, then

1) The vectors $\{\boldsymbol{v}_i\}_{i=1}^M$ defined in (16) are a basis for $\ker d^2F(q^*)$.
2) If $d^2F(q^*)$ is singular then $d^2\mathcal{L}(q^*)$ is singular.
3) The vectors $\{\boldsymbol{w}_i\}_{i=1}^{M-1}$ defined in (17) are a basis for $\ker d^2\mathcal{L}(q^*)$.

Observe that the dimensionality of $\ker d^2\mathcal{L}(q^*)$ is one less than $\ker d^2F(q^*)$. This insight suggests that when $\dim\ker d^2F(q^*) = 1$, then $d^2\mathcal{L}(q^*)$ is nonsingular. This is indeed the case.

Corollary 5.4: If q^* is 1-singular, then $d^2\mathcal{L}(q^*)$ is nonsingular.

D. Isotropy Groups

The isotropy group of an equilibrium $q = ((q^1)^T,...,(q^N)^T)^T$ of the gradient system (12) is a subgroup of S_N which fixes q. If $q^\nu = q^\eta$ for all of the M classes $\nu,\eta\in\mathcal{U}$, then $S_M\subseteq S_N$ is the isotropy

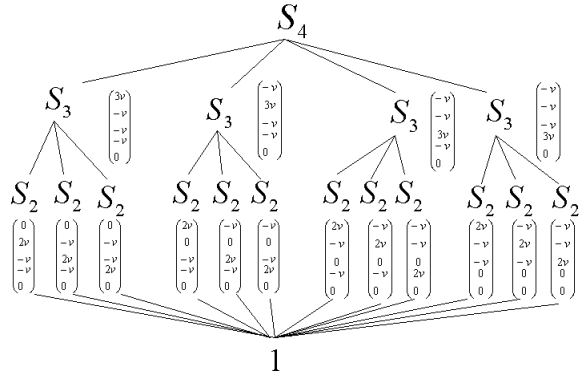

Fig. 8. The lattice of the isotropy subgroups $S_M < S_N$ for $N = 4$ and the corresponding basis vectors of the fixed point spaces of the corresponding groups.

group of q, where S_M freely permutes the sub-vectors q^ν if $\nu \in \mathcal{U}$, but holds fixed the sub-vectors q^ν if $\nu \in \mathcal{R}$.

The isotropy groups of (q, λ) for the soft clusterings q pictured in Figure 4 are clear. In panel 1 of the bottom row, $\mathcal{U} = \{1, 2, 3, 4\}$, and the isotropy group is S_4. In panel 2, $\mathcal{U} = \{2, 3, 4\}$ and the isotropy group is S_3. In panels 3 and 4, $\mathcal{U} = \{2, 4\}$ and the isotropy group is S_2.

Restricted to $\ker d^2\mathcal{L}(q^*)$, the fixed point space of the subgroup $S_{M-1} \subset S_M$ is one dimensional (see Corollary 5.6 and Figure 8). Golubitsky and Stewart [55] show that all of the isotropy subgroups in S_M with one dimensional fixed point spaces are of the form $S_m \times S_n$, where $m+n = M$. The following Lemma which follows from this result will allow us to use the Equivariant Branching Lemma (Theorem 4.3 and Remark 4.4) to ascertain the existence of explicit bifurcating solutions.

Lemma 5.5: Let $M = m + n$ such that $M > 1$ and $m, n > 0$. Let $\mathcal{U}_m$ be a set of m classes, and let $\mathcal{U}_n$ be a set of n classes such that $\mathcal{U}_m \cap \mathcal{U}_n = \emptyset$ and $\mathcal{U}_m \cup \mathcal{U}_n = \mathcal{U}$. Now define $\hat{\boldsymbol{u}}_{(m,n)} \in \Re^{NK}$ such that

$$\hat{\boldsymbol{u}}^\nu_{(m,n)} = \begin{cases} \frac{n}{m}\boldsymbol{v} & \text{if } \nu \in \mathcal{U}_m \\ -\boldsymbol{v} & \text{if } \nu \in \mathcal{U}_n \\ \mathbf{0} & \text{otherwise} \end{cases}$$

where $\boldsymbol{v}$ is defined as in (14), and let

$$\boldsymbol{u}_{(m,n)} = \begin{pmatrix} \hat{\boldsymbol{u}}_{(m,n)} \\ \mathbf{0} \end{pmatrix} \tag{18}$$

where $\mathbf{0} \in \Re^K$. Then the isotropy subgroup of $\boldsymbol{u}_{(m,n)}$ is $S_m \times S_n$, where S_m acts on $\boldsymbol{u}^\nu$ when $\nu \in \mathcal{U}_m$ and S_n acts $\boldsymbol{u}^\nu$ when $\nu \in \mathcal{U}_n$. The fixed point space of $S_m \times S_n$ restricted to $\ker d^2\mathcal{L}(q^*)$ is one dimensional.

Without loss of generality, one can assume that $\mathcal{U}_n$ contains the first n classes of $\mathcal{U}$, and that $\mathcal{U}_m$ contains the other m classes. Now it is straightforward to verify that $\boldsymbol{u}_{(m,n)} = -\sum_{i=1}^{n} \boldsymbol{w}_i + \frac{n}{m}\sum_{j=n+1}^{M-1} \boldsymbol{w}_j$, confirming that $\boldsymbol{u}_{(m,n)} \in \ker d^2\mathcal{L}(q^*)$ as claimed.

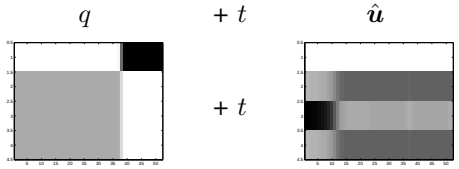

Fig. 9. A bifurcating solution from the soft clustering $q = ((q^1)^T, (q^2)^T, (q^3)^T, (q^4)^T)^T \in \mathrm{Fix}(S_3)$ at $\beta \approx 1.1339$ (panel 3 in the bottom row of Figure 4) where S_3 acts on q by freely permuting the three sub-vectors q^2, q^3, q^4. Note that t is a scalar. The bifurcating direction is $\hat{u} = (\mathbf{0}, -\boldsymbol{v}^T, 2\boldsymbol{v}^T, -\boldsymbol{v}^T)^T \in \mathrm{Fix}(S_2)$, which is invariant under S_2 permuting $\hat{\boldsymbol{u}}^2$ and $\hat{\boldsymbol{u}}^4$. The new soft clustering $q + t\hat{u}$ after the bifurcation has isotropy group S_2.

Letting $m = 1$ and $n = M - 1$ yields the following Corollary.

Corollary 5.6: Let $\hat{\boldsymbol{u}}_k \in \Re^{NK}$ such that

$$\hat{\boldsymbol{u}}^\nu_k = \begin{cases} (M-1)\boldsymbol{v} & \text{if } \nu \text{ is the } k^{th} \text{ class of } \mathcal{U} \\ -\boldsymbol{v} & \text{if } \nu \neq k \text{ is any other class of } \mathcal{U} \\ \mathbf{0} & \text{otherwise} \end{cases}$$

where $\boldsymbol{v}$ is defined as in (14), and let

$$\boldsymbol{u}_k = \begin{pmatrix} \hat{\boldsymbol{u}}_k \\ \mathbf{0} \end{pmatrix} \tag{19}$$

where $\mathbf{0} \in \Re^K$. Then the isotropy subgroup of $\boldsymbol{u}_k$ is S_{M-1}. The fixed point space of S_{M-1} restricted to $\ker d^2\mathcal{L}(q^*)$ is one dimensional.

Figure 8 gives the lattice of isotropy subgroups of S_N when $N = 4$, and the corresponding basis vectors of the fixed point spaces.

Figure 9 depicts a soft clustering $q \in \mathrm{Fix}(S_3)$ where S_3 acts on $q = ((q^1)^T, (q^2)^T, (q^3)^T, (q^4)^T)^T$ by permuting the three sub-vectors q^2, q^3, q^4. Also depicted is a vector $\hat{\boldsymbol{u}} = (\mathbf{0}, -\boldsymbol{v}^T, 2\boldsymbol{v}^T, -\boldsymbol{v}^T)^T \in \mathrm{Fix}(S_2)$ where S_2 permutes $\hat{\boldsymbol{u}}^2$ and $\hat{\boldsymbol{u}}^4$.

VI. Bifurcations

There are two types of bifurcations of equilibria in any dynamical system with symmetry: symmetry breaking bifurcations and symmetry preserving bifurcations. We next address each of these bifurcation types for the flow (12), and conclude with a generic picture of the full bifurcation structure.

Equilibria of the gradient flow of the Lagrangian (12) are stationary points of the optimization problem (3) and of the annealing problem (5). Thus, this section gives the bifurcation structure of these stationary points.

A. Symmetry Breaking Bifurcations

We have laid the groundwork so that we may ascertain the existence of explicit bifurcating branches of equilibria of (12)

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

from an equilibrium $(q,^*,\lambda^*,\beta^*)$ when q^* is M-singular for $M > 1$ (Theorem 6.2). We will show that these symmetry breaking bifurcations are always degenerate (Theorem 6.3), that is, $\beta'(0) = 0$. If $\beta''(0) \neq 0$, which is a generic assumption, then these bifurcations are pitchforks. We will provide a condition, called the bifurcation discriminator, which ascertains whether the bifurcating branches with isotropy group $S_m \times S_n$ are subcritical or supercritical (Theorem 6.5). Lastly, we also provide a condition which determines whether branches are stable or unstable (Theorem 6.7).

Throughout this section we assume that (q^*,λ^*,β^*) is an M-singular point for $M > 1$. The reduced problem, finding the vector $\boldsymbol{v} \in \Re^K$ in the kernel of the M singular blocks B of $d^2F(q^*)$, specifies how the data $\boldsymbol{Y}$ ought to split. Thus $d^2F(q^*)$ and $d^2\mathcal{L}(q^*)$ is both singular. The vectors $\boldsymbol{w}_i \in \Re^{NK+K}$ are constructed from the vector $\boldsymbol{v}$, and they form a basis for $\ker d^2\mathcal{L}$ which has dimension $M-1$. The vectors $u_{(m,n)} \in \Re^{NK+K}$ are particular vectors in $\ker d^2\mathcal{L}$ which have isotropy group $S_m \times S_n$. Since these belong to $\ker d^2\mathcal{L}$, they are in the span of the vectors $\boldsymbol{w}_i$ and hence are also constructed using the vector $\boldsymbol{v}$. The vectors $\boldsymbol{u}_{(m,n)}$ determine which classes of $\boldsymbol{T}$ the data is split into.

1) Crossing Condition: Before presenting the existence theorem for bifurcating branches, it is first necessary to address when the crossing condition ("$d^2_{\beta\boldsymbol{x}}f(\mathbf{0},0) \neq \mathbf{0}$"), required by Theorem 4.3, is satisfied. Observe that when $F(q,\beta) = G(q) + \beta D(q)$ as in (10), then $d_\beta(d^2\mathcal{L}) = \begin{pmatrix} d^2D \\ \mathbf{0} \end{pmatrix}$. For annealing problems of the form (5), we have shown [31] that the crossing condition in Theorem 4.3 is satisfied if and only if

$$\boldsymbol{v}_i^T d^2D(q^*)\boldsymbol{v}_i \neq 0 \tag{20}$$

where $\boldsymbol{v}_i$ is any of the basis vectors of $\ker d^2F(q^*)$ (see (16)). This result is illuminating: if $d^2D(q^*)$ is either positive or negative definite on $\ker d^2F(q^*)$, then the crossing condition is satisfied. We have the following Theorem.

Theorem 6.1: Suppose that q^* is M-singular for $1 < M \leq N$.

1) If $d^2D(q^*)$ is either positive or negative definite on $\ker d^2F(q^*)$, then (q^*,λ^*,β^*) is a singularity of the gradient flow of the Lagrangian (12) if and only if (q^*,λ^*,β^*) is a bifurcation point.
2) If $d^2G(q^*)$ is either positive or negative definite on $\ker d^2F(q^*)$, then (q^*,λ^*,β^*) is a singularity of (12) if and only if (q^*,λ^*,β^*) is a bifurcation point.

Proof: The first part of the Theorem follows from the claim that the crossing condition is equivalent to (20). To prove the second part, observe that if $\boldsymbol{k} \in \ker d^2F(q^*)$, then $d^2F(q^*)\boldsymbol{k} = \mathbf{0}$ implies that $\boldsymbol{k}^T d^2G(q^*)\boldsymbol{k} + \beta^*\boldsymbol{k}^T d^2D(q^*)\boldsymbol{k} = 0$. Since $\boldsymbol{k}^T d^2G(q^*)\boldsymbol{k} < 0$ (or $\boldsymbol{k}^T d^2G(q^*)\boldsymbol{k} > 0$), then $\boldsymbol{k}^T \Delta D(q^*)\boldsymbol{k} > 0$ (or $\boldsymbol{k}^T \Delta D(q^*)\boldsymbol{k} < 0$). Now apply the first part of the Theorem. □

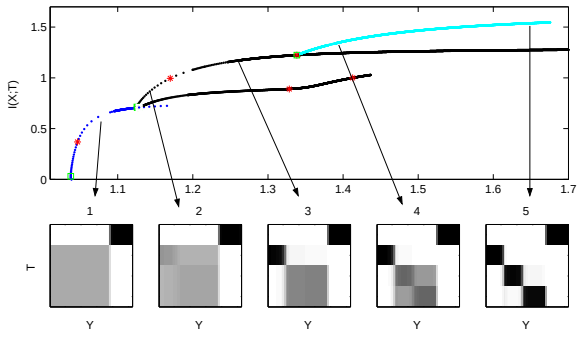

Fig. 10. Bifurcation diagram of stationary points of (7) when $N = 4$. Figure 4 showed an incomplete bifurcation diagram for this same scenario since the algorithm in that case was affected by the stability of the branches. The panels illustrate the sequence of symmetry breaking bifurcations from the branch $(q_{\frac{1}{N}},\lambda,\beta)$ with symmetry S_4, to a branch with symmetry S_3 (blue), then to S_2 (black), and finally, to S_1 (cyan).

By Theorem 6.1, for annealing problems where $G(q)$ is strictly concave, $d^2D(q)$ is positive definite on $\ker d^2F(q^*)$, so every singularity is a bifurcation point. For the Information Distortion problem (7), $G(q) = \boldsymbol{H}(\boldsymbol{T}|\boldsymbol{Y})$ is strictly concave, so every singularity of $d^2\mathcal{L}(q^*)$ is a bifurcation. For the Information Bottleneck problem (6), $G(q) = -\boldsymbol{I}(\boldsymbol{Y};\boldsymbol{T})$ is concave, but not strictly concave, and $D(q) = \boldsymbol{I}(\boldsymbol{X};\boldsymbol{Y})$ is convex, but not strictly convex.

2) Explicit Bifurcating Branches: By Lemma 5.5 and the Equivariant Branching Lemma, we have the following existence theorem.

Theorem 6.2: Let (q^*,λ^*,β^*) be an equilibrium of the gradient flow (12) such that q^* is M-singular for $1 < M \leq N$, and the crossing condition (20) is satisfied. Then there exists $\frac{M!}{m!n!}$ bifurcating solutions, $\left(\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}_{(m,n)}, \beta^* + \beta(t)\right)$, where $\boldsymbol{u}_{(m,n)}$ is defined in (18), for every pair (m,n) such that $M = m+n$ and $m,n > 0$, each with isotropy group isomorphic to $S_m \times S_n$. Of these solutions, there are M of the form $\left(\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}_k, \beta^* + \beta(t)\right)$, where $\boldsymbol{u}_k$ is defined in (19), for $1 \leq k \leq M$, each with isotropy group S_{M-1}.

Figure 8 depicts the lattice of subgroups of S_4 of the form S_M for $1 \leq M < N$, as well as the M bifurcating directions from a bifurcation at (q^*,λ^*,β^*) guaranteed by Theorem 6.2. Observe that $\sum_\nu \boldsymbol{u}^\nu = \mathbf{0}$, which is true for any vector in $\ker d^2\mathcal{L}(q^*)$ by (17). This assures that for small enough t, $q^* + t\hat{\boldsymbol{u}}$ is in Δ.

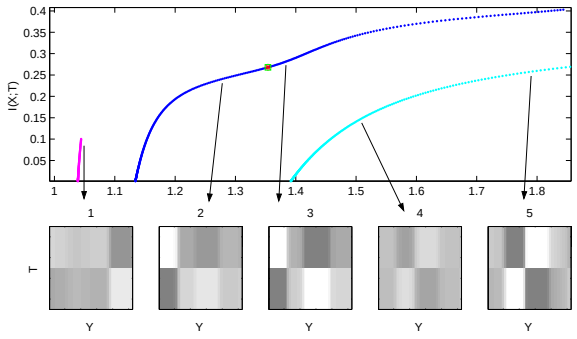

Fig. 11. Symmetry breaking bifurcations from the branch $(q_{\frac{1}{N}}, \lambda, \beta)$ with symmetry S_4 to branches which have symmetry $S_2 \times S_2$.

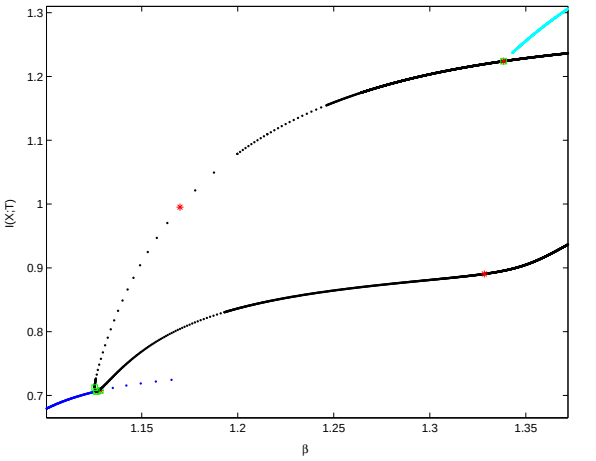

Fig. 12. A close up, from Figure 10, of the branch with S_2 symmetry (in black) which connects the S_3 symmetric branch below (blue branch in the lower left) to the S_1 symmetric branch (cyan branch in the upper right). The soft clusterings on the suboptimal S_2 symmetric branch (lower black branch) are investigated further in Figure 15. By Theorem 6.3, the symmetry breaking bifurcations from $S_3 \to S_2$ and from $S_2 \to S_1$ are degenerate, and, since $\beta''(0) \neq 0$, a pitchfork.

Figure 9 depicts a symmetry breaking bifurcating solution from $q \in \mathrm{Fix}(S_3)$ to $(q + t\hat{\boldsymbol{u}}) \in \mathrm{Fix}(S_2)$ at $\beta^* = 1.1339$.

Figures 6 and 10 show some of the bifurcating branches guaranteed by Theorem 6.2 when $N = 4$ for the Information Distortion problem (7) (see section VII for details). The symmetry of the clusterings shown depict symmetry breaking from $S_4 \to S_3 \to S_2 \to S_1$.

Figure 11 depicts symmetry breaking from S_4 to $S_2 \times S_2$. The first bifurcation occurs at $\beta^* = 1.0387$, as does the symmetry breaking bifurcation from S_4 to S_3 given in Figure 10. The subsequent two bifurcating branches given in Figure 11 correspond to bifurcations at $\beta^* = 1.1339$ and $\beta^* = 1.3910$.

Theorem 6.2 does not exclude the existence of other bifurcating directions with symmetry other than S_{M-1} or $S_m \times S_n$ (consider for example the symmetry $S_a \times S_b \times S_c$ where $a + b + c = M$). To our best knowledge, for the Information Distortion, Information Bottleneck, and Deterministic Annealing methods, such bifurcating solutions have never been observed [10], [11], [14]. However, rigorous results in this direction are still lacking.

3) Pitchfork Bifurcating Branches: Suppose that a bifurcation occurs at $(q^*, \lambda^*, \beta^*)$ where q^* is M-singular. This section examines the structure of the bifurcating branches

$$\left(\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}, \beta^* + \beta(t) \right), \tag{21}$$

whose existence is guaranteed by Theorem 6.2. The proofs to the results which follow rely on the explicit computation of the derivatives of the Liapunov-Schmidt reduction referred to in Remark 4.4. We will cite the Theorems, and the interested reader is referred to [31] for the proofs.

Theorem 6.3: If q^* is M-singular for $1 < M \leq N$, then all of the bifurcating branches (21) guaranteed by Theorem 6.2 are degenerate (i.e. $\beta'(0) = 0$).

From Definition 4.5, the sign of $\beta''(0)$ determines whether a bifurcating branch (21) is a pitchfork and subcritical ($\beta''(0) < 0$) or a pitchfork and supercritical ($\beta''(0) > 0$). Without further restrictions on $\beta(t)$, $\beta''(0) \neq 0$ generically, as in the case study presented in section II-B, and the four blob Gaussian mixture model in section III-A. Thus, symmetry breaking bifurcations are generically pitchforks. Next, a condition is given which determines the sign of $\beta''(0)$ for the bifurcating branches with a given isotropy group.

Definition 6.4: The *bifurcation discriminator* of the bifurcating branches (21) with isotropy group $S_m \times S_n$ is

$$\zeta(q^*, \beta^*, m, n) := 3\Xi - d^4 f[\boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}],$$

where

$$\begin{aligned} \Xi &:= \boldsymbol{b}^T B^- \left(I_K - \frac{mn(m+n)}{m^2 - mn + n^2} A^{-1} \right) \boldsymbol{b} \\ \boldsymbol{b} &:= d^3 f[\boldsymbol{v}, \boldsymbol{v}]. \end{aligned}$$

The matrix B^- is the Moore-Penrose generalized inverse [56] of a block of the Hessian (13), $A = B \sum_{\nu \in \mathcal{R}} B_\nu^{-1} + MI_K$ from (15), and $\boldsymbol{v}$ is the basis vector of $\ker(B)$ from (14).

When $q^* = q_{\frac{1}{N}}$ is N-singular, then $A^{-1} = \frac{1}{N} I_K$, and so in this case the bifurcation discriminator $\zeta(q_{\frac{1}{N}}, \beta^*, m, n)$ is

$$3 \left(1 - \frac{mn}{m^2 - mn + n^2} \right) \boldsymbol{b}^T B^- \boldsymbol{b} - d^4 f[\boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}]. \tag{22}$$

The discriminator $\zeta(q^*, \beta^*, m, n)$ is defined purely in terms of the constitutive function f of $F(q, \beta) = \sum_{\nu=1}^{N} f(q^\nu, \beta)$ (see Assumption 5.1). This follows since the blocks of $d^2F(q^*)$ are written as $B_\nu = d^2 f(q^\nu, \beta)$, A is a function of these blocks, and $B = d^2 f(q^\nu, \beta)$ for $\nu \in \mathcal{U}$. The fourth derivative $d^4 f[\boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}]$ in $\zeta(q^*, \beta^*, m, n)$ can be expressed as

$$\sum_{r,s,t,u \in Y} \frac{\partial^4 F(q^*, \beta^*)}{\partial q_{\nu r} \partial q_{\nu s} \partial q_{\nu t} \partial q_{\nu u}} [\boldsymbol{v}]_r [\boldsymbol{v}]_s [\boldsymbol{v}]_t [\boldsymbol{v}]_u$$

and the vector $\boldsymbol{b}$ has t^{th} component

$$[\boldsymbol{b}]_t = [d^3 f[\boldsymbol{v}, \boldsymbol{v}]]_t = \sum_{r,s \in Y} \frac{\partial^3 F(q^*, \beta^*)}{\partial q_{\nu r} \partial q_{\nu s} \partial q_{\nu t}} [\boldsymbol{v}]_r [\boldsymbol{v}]_s.$$

The next theorem shows that the sign of $\beta''(0)$ is determined by the sign of $\zeta(q^*, \beta^*, m, n)$.

Theorem 6.5: Suppose q^* is M-singular for $1 < M \leq N$ and that $d^2 D(q^*)$ is positive definite on $\ker d^2 F(q^*)$. If $\zeta(q^*, \beta^*, m, n) < 0$, then the bifurcating branches (21) guaranteed by Theorem 6.2, are pitchforks and subcritical. If $\zeta(q^*, \beta^*, m, n) > 0$, then the bifurcating branches are pitchforks and supercritical.

This theorem is in contrast to the behavior of generic S_N invariant functions, such as the model for speciation in [57], [58], where the symmetry breaking bifurcations are transcritical. The difference is due to the constraints imposed by $q \in \Delta$ and the form of $F = G + \beta D$ given in Assumptions 5.1.

A result similar to Theorem 6.5 holds when $d^2 D(q^*)$ is negative definite on $\ker d^2 F(q^*)$, but now $\zeta < 0$ predicts supercritical branches, and $\zeta > 0$ predicts subcritical branches.

In section VI-A1, we showed that for the Information Distortion problem (7), the condition in Theorem 6.5 that $d^2 D(q^*)$ be positive definite on $\ker d^2 F(q^*)$ is always satisfied for every singularity. Thus, for the Information Distortion, Theorem 6.5 can always be applied to determine whether pitchforks are subcritical or supercritical. To calculate $\zeta(q^*, \beta^*, m, n)$ for the Information Distortion problem, we have the following Lemma.

Lemma 6.6: For the Information Distortion problem (7), $[d^3 f]_{rst} = \frac{\partial^3 F}{\partial q_{\nu r} \partial q_{\nu s} \partial q_{\nu t}}$ is equal to

$$\frac{1}{\ln 2} \left(\delta_{rst} \frac{p(y_r)}{q_{\nu r}^2} + \beta \left(\frac{p(y_r) p(y_s) p(y_t)}{(\sum_j p(y_j) q_{\nu j})^2} - A \right) \right)$$

where $A = \sum_i \frac{p(x_i, y_r) p(x_i, y_s) p(x_i, y_t)}{(\sum_j p(x_i, y_j) q_{\nu j})^2}$. The expression $[d^4 f]_{rstu} = \frac{\partial^4 F}{\partial q_{\nu r} \partial q_{\nu s} \partial q_{\nu t} \partial q_{\nu u}}$ is equal to

$$\frac{2}{\ln 2} \left(\beta \left(B - \frac{p(y_r) p(y_s) p(y_t) p(y_u)}{(\sum_j p(y_j) q_{\nu j})^3} \right) - \delta_{rstu} \frac{p(y_r)}{q_{\nu r}^3} \right)$$

where $B = \sum_i \frac{p(x_i, y_r) p(x_i, y_s) p(x_i, y_t) p(x_i, y_u)}{(\sum_j p(x_i, y_j) q_{\nu j})^3}$

Proof: Direct computation of the derivatives of $G(q) = \boldsymbol{H}(\boldsymbol{T}|\boldsymbol{Y})$ and $D(q) = I(\boldsymbol{X}; \boldsymbol{T})$. □

Consider the bifurcation at $(q_{\frac{1}{N}}, \lambda^*, \beta^* = 1.0387)$ in Figure 10 where symmetry breaks from S_4 to S_3. The value of the discriminator at this bifurcation is $\zeta(q_{\frac{1}{N}}, 1.0387, 1, 3) = -.0075$ (see section VII for details), which predicts that this bifurcation is a pitchfork and subcritical. Figure 13, a close up of the bifurcation diagram at this bifurcation, illustrates the subcritical bifurcating branch.

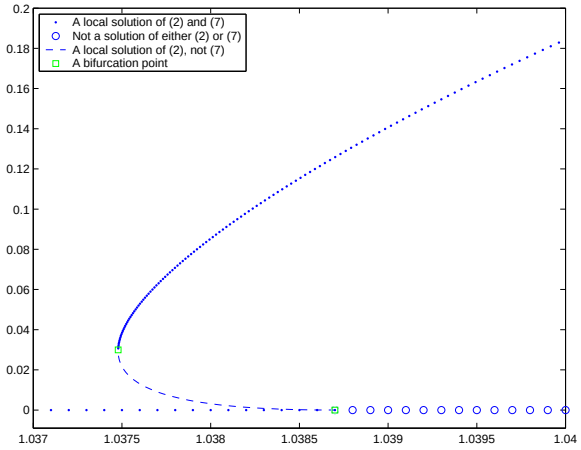

Fig. 13. A close-up of Figures 6 and 10 at $\beta^* \approx 1.0387$. Illustrated here is a subcritical pitchfork bifurcation from the branch $(q_{\frac{1}{4}}, \beta)$, a break in symmetry from S_4 to S_3. This was predicted by the fact that $\zeta(q_{\frac{1}{4}}, 1.0387, 1, 3) < 0$. It is at the symmetry preserving saddle node at $\beta \approx 1.0375$ that this branch changes from being composed of stationary points to local solutions of the problem (7) (see section VIII-A).

4) Stability: We now address the stability of the bifurcating branches. We will relate the stability of equilibria to optimality in the optimization problem (3) and the annealing problem (5) in section VIII-A.

As illustrated in section IV-C, to ascertain stability, one determines whether or not $d^2\mathcal{L}$, evaluated at the equilibria on a bifurcating branch, has positive eigenvalues ($d^2\mathcal{L}$ is a symmetric matrix, so it only has real eigenvalues). The next theorem, whose proof is in [31], provides a condition to determine when this occurs.

Theorem 6.7: Suppose q^* is M-singular for $1 < M \leq N$ and that $d^2 D(q^*)$ is positive definite on $\ker d^2 F(q^*)$. All of the subcritical bifurcating branches (21) guaranteed by Theorem 6.2 are unstable. If the bifurcating branch is supercritical and if

$$\theta(q^*, \beta^*, m, n) := \sum_{k=1}^{M-1} (\theta_1 - 2\theta_2 - \theta_3) > 0,$$

then the branch consists of unstable solutions. The component functions of θ are

$$\begin{aligned}\theta_1 &= d^4\mathcal{L}[\boldsymbol{w}_k, \boldsymbol{w}_k, \boldsymbol{u}, \boldsymbol{u}],\\ \theta_2 &= d^3\mathcal{L}[\boldsymbol{w}_k, \boldsymbol{u}, L^- d^3\mathcal{L}[\boldsymbol{w}_k, \boldsymbol{u}]],\\ \theta_3 &= d^3\mathcal{L}[\boldsymbol{w}_k, \boldsymbol{w}_k, L^- d^3\mathcal{L}[\boldsymbol{u}, \boldsymbol{u}]],\end{aligned}$$

where all of the derivatives are taken with respect to (q, λ), L^- is the Moore-Penrose inverse of L, and $\boldsymbol{w}_k$ is a basis vector from (17).

The expression $\theta(q^*, \beta^*, m, n)$ from Theorem 6.7 can be simplified to a form which only uses derivatives of the constituent functions f of F (as we did in Definition 6.4),

$$\begin{aligned}\theta_1 &= \left(\frac{n^2}{m} + n\right) d^4 f[\boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}, \boldsymbol{v}]\\ \theta_2 &= \boldsymbol{b}^T B^- \left(a_1 I_K + a_2 A^{-1}\right) \boldsymbol{b}\\ \theta_3 &= \boldsymbol{b}^T B^- \left(a_3 I_K + a_4 A^{-1}\right) \boldsymbol{b}\end{aligned}$$

where a_i are scalars which depend only on m and n.

By Theorem 6.7, the subcritical bifurcating branch depicted in Figure 13 is unstable.

B. Symmetry Preserving Bifurcations

We now turn our attention to bifurcations which are not symmetry breaking bifurcations of equilibria of (12),

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

We show that, generically, these bifurcations are saddle-node bifurcations, which we have illustrated numerically in Figure 13 for the Information Distortion problem (7).

In contrast to the conditions which led to a symmetry breaking bifurcation in section VI-A, where $d^2F(q^*)$ had a high dimensional kernel (see Definition 5.2), for a symmetry preserving bifurcation, $d^2F(q^*)$ is (generically) nonsingular.

Lemma 6.8: At a generic symmetry preserving bifurcation $(q^*, \lambda^*, \beta^*)$, the Hessian $d^2F(q^*)$ is nonsingular.

Proof: If $d^2F(q^*)$ is singular, then at least one of the blocks B_ν is singular. If there are multiple blocks equal to B_ν, then Theorem 6.2 implies that q^* undergoes a symmetry breaking bifurcation. Thus B_ν is the only block that is singular, and now Corollary 5.4 shows that $d^2\mathcal{L}$ is nonsingular. This leads to a contradiction since we assume that a bifurcation takes place at q^*. □

If $(q^*, \lambda^*, \beta^*)$ is a singularity of the gradient flow (12) such that $d^2F(q^*)$ is nonsingular, then $\ker d^2\mathcal{L}(q^*)$ looks very different than the form of $\ker d^2\mathcal{L}(q^*)$ when symmetry breaking bifurcation occurs (see section V-C). In fact, when $d^2F(q^*)$ is nonsingular, it can be shown

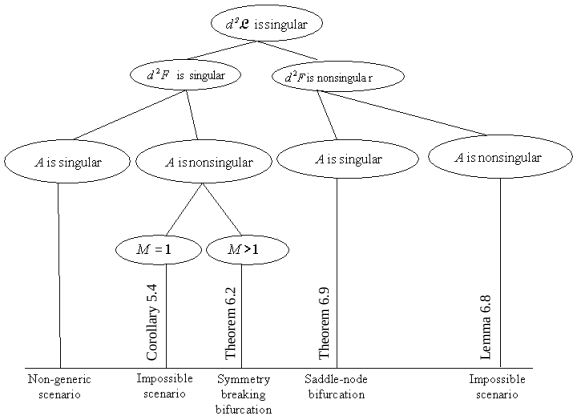

Fig. 14. A hierarchical diagram showing how the singular points of $d^2\mathcal{L}$ and d^2F affect the bifurcating branches of stationary points of the optimization problem (3) and stationary points of the annealing problem (5).

[31] that $\ker d^2\mathcal{L}(q^*)$ is one dimensional, with basis vector

$$\boldsymbol{w} = \left((B_1^{-1}\boldsymbol{v})^T, (B_2^{-1}\boldsymbol{v})^T, \ldots, (B_N^{-1}\boldsymbol{v})^T, -\boldsymbol{v}^T\right)^T,$$

where $\{B_\nu\}_{\nu=1}^N$ are the blocks of $d^2F(q^*)$, and $\boldsymbol{v}$ is in $\ker A$, where $A = B\sum_{\nu\in\mathcal{R}} B_\nu^{-1} + MI_K$ (see (15)). At a symmetry breaking bifurcation, the matrix A is generically nonsingular.

Now we provide a sufficient condition for the existence of saddle-node bifurcations. The first assumption given in the following theorem is satisfied generically at any symmetry preserving bifurcation (Lemma 6.8), the second assumption is a crossing condition, and the third condition assures that $\beta''(0) \neq 0$.

Theorem 6.9: [31] Let $\boldsymbol{w} \in \ker d^2\mathcal{L}(q^*)$. Suppose that $(q^*, \lambda^*, \beta^*)$ is a singularity of the gradient system 12) such that:

1) The Hessian $d^2F(q^*)$ is nonsingular.
2) The dot product $\boldsymbol{w}^T \begin{pmatrix} \nabla D(q^*) \\ \mathbf{0} \end{pmatrix} \neq 0.$
3) $d^3\mathcal{L}[\boldsymbol{w}, \boldsymbol{w}, \boldsymbol{w}] \neq 0.$

Then, generically, $(q^*, \lambda^*, \beta^*)$ is a saddle-node bifurcation.

C. Generic Bifurcations

We have described the generic bifurcation structure of stationary points to problems of the form

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

as long as $G + \beta D = \sum_{\nu=1}^N f(q^\nu, \beta)$. Symmetry breaking bifurcations are pitchforks, and symmetry preserving bifurcations are saddle-nodes. The type of bifurcation which occurs depends on three types of singular points,

which depend on $d^2\mathcal{L}(q^*)$, $d^2F(q^*)$, and the matrix $A = B\sum_{\nu\in\mathcal{R}} B_\nu^{-1} + MI_K$ (see (15)) which we have depicted in Figure 14.

The first type of singular point is where the $M > 1$ blocks B_ν of d^2F, for $\nu \in \mathcal{U}$, are singular. By Theorem 5.3, $d^2\mathcal{L}$ must be singular. Generically, the blocks, $\{B_\nu\}_{\nu\in\mathcal{R}}$, of d^2F are nonsingular, and $A = B\sum_{\nu\in\mathcal{R}} B_\nu^{-1} + MI_K$ is nonsingular. Theorem 6.2 shows that this is the type of singularity that exhibits symmetry breaking bifurcation.

The second type of singular point is a special case in which no bifurcation occurs. If only a single block, B_ν, of d^2F is singular (i.e. $M = 1$) , and if the generic condition that the corresponding A is nonsingular holds, then we show in Corollary 5.4 that $d^2\mathcal{L}$ is nonsingular. Thus, generically, no bifurcation occurs for this case.

The third type of singular point is when $d^2\mathcal{L}$ is singular, but when d^2F is nonsingular. In this case, the matrix A must be singular [31]. This singular point manifests itself as a saddle-node bifurcation (Theorem 6.9). Figure 14, which summarizes the preceding discussion, indicates how the singular points of $d^2\mathcal{L}$ and d^2F affect the bifurcations of equilibria of the flow (12).

Another way to categorize the bifurcations of the annealing solutions to (5) is to consider the derivatives of $D(q)$. The second condition in Theorem 6.9, which guarantees the existence of a symmetry preserving saddle-node bifurcation, is equivalent to requiring that $\begin{pmatrix} \nabla D \\ \mathbf{0} \end{pmatrix} \notin \text{range}(d^2\mathcal{L}(q^*))$. For symmetry breaking bifurcations, $\begin{pmatrix} \nabla D \\ \mathbf{0} \end{pmatrix} \in \text{range}(d^2\mathcal{L}(q^*))$. In fact, whenever $d^2\mathcal{L}(q^*)$ is nonsingular, by the Implicit Function Theorem, taking the total derivative of $\nabla_{q,\lambda}\mathcal{L}(q,\lambda,\beta) = 0$ shows that $\begin{pmatrix} \nabla D \\ \mathbf{0} \end{pmatrix}$ is always in $\text{range}(d^2\mathcal{L}(q^*))$. Furthermore, equation (20) shows that the crossing condition depends on d^2D, and Theorems 6.5 and 6.7 show that d^2D influences whether bifurcating branches are subcritical or supercritical, as well as stable or unstable.

VII. Numerical Results

We created software in MATLAB® to implement pseudo-arclength continuation to numerically illustrate the bifurcation diagram of stationary points to the optimization problem (3) and the annealing problem (5) as guaranteed by the theory of section VI.

This continuation scheme, due to Keller [59]–[61], uses Newton's method to find the next equilibrium, $(q_{k+1}, \lambda_{k+1}, \beta_{k+1})$, from $(q_k, \lambda_k, \beta_k)$ by allowing both (q, λ) and β to vary. The advantage of this approach over Algorithm 3.1 is twofold. First, the step size in β, $\Delta\beta_{k+1} = \beta_{k+1} - \beta_k$, changes automatically depending on the "steepness" of the curve $\nabla\mathcal{L} = 0$ at $(q_k, \lambda_k, \beta_k)$ and so this method allows for continuation of equilibria around a saddle-node bifurcation. Secondly, this algorithm is able to continue along unstable branches.

N	2	3	4	5	6
$\zeta(q_{\frac{1}{N}}, \beta^*)$	0.0006	-0.0010	-0.0075	- 0.0197	-.0391

TABLE I
THE BIFURCATION DISCRIMINATOR: NUMERICAL EVALUATIONS OF THE BIFURCATION DISCRIMINATOR $\zeta(q_{\frac{1}{N}}, \beta^*) := \zeta(q_{\frac{1}{N}}, \beta^* \approx 1.038706, m = 1, n = N-1)$ AS A FUNCTION OF N FOR THE FOUR BLOB PROBLEM (SEE FIGURE 3A) WHEN F IS DEFINED AS IN (7). A SUPERCRITICAL BIFURCATION IS PREDICTED WHEN $N = 2$, AND SUBCRITICAL BIFURCATIONS FOR $N \in \{3, 4, 5, 6\}$.

All of the results presented here are for the Information Distortion problem (7),

$$\max_{q\in\Delta}(H(q) + \beta I(q))$$

where $p(\boldsymbol{X}, \boldsymbol{Y})$ is the mixture of four Gaussian blobs introduced in Figure 3, and we optimally cluster the responses $\boldsymbol{T}$ into $N = 4$ clusters.

Figures 6 and 10 are analogous to Figure 4, using the same mixture of Gaussians $p(\boldsymbol{X}, \boldsymbol{Y})$ and the same Information Distortion cost function. The difference is that Figure 4 was obtained using the Basic Annealing Algorithm, while we used the continuation algorithm in Figures 6 and 10. The continuation algorithm shows that the bifurcation structure is richer than shown in Figure 4. In Figure 6 we show bifurcating branches which emanate from the uniform S_4 invariant branch $(q_{\frac{1}{N}}, \lambda, \beta)$ at $\beta^* \approx 1.0387$, 1.1339, and 1.3910. In the bottom row of Figure 10, panels 1-5 show that the clusterings along the branches break symmetry from S_4 to S_3 to S_2, and, finally, to S_1. An "*" indicates a point where $d^2F(q^*)$ is singular, and a square indicates a point where $d^2\mathcal{L}(q^*)$ is singular. Notice that there are points denoted by "*" from which no bifurcating branches emanate. At these points a single block of d^2F is singular, and, as explained by Corollary 5.4, $d^2\mathcal{L}(q^*)$ is nonsingular, and so no bifurcation occurs. Notice that there are also points where both $d^2\mathcal{L}(q^*)$ and $d^2F(q^*)$ are singular (at the symmetry breaking bifurcations) and points where just $d^2\mathcal{L}(q^*)$ is singular (at the saddle-node bifurcations). These three types of singular points are depicted in Figure 14.

Figure 11 illustrates symmetry breaking from S_4 to $S_2 \times S_2$. The clusterings depicted in the panels are not found when using an algorithm which is affected by the stability of the equilibria (such as the Basic Annealing Algorithm).

Theorem 6.5 shows that the bifurcation discriminator, $\zeta(q^*, \beta^*, m, n)$, can determine whether the bifurcating

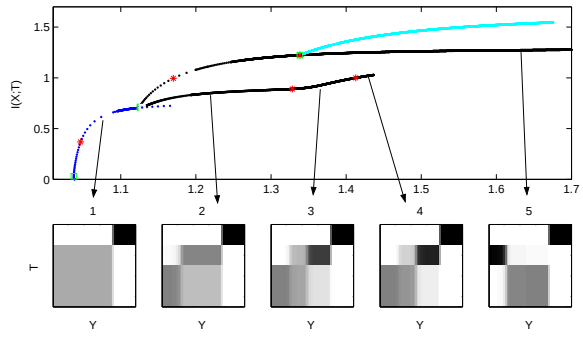

Fig. 15. The symmetry breaking bifurcating branches from the solution branch $(q_{\frac{1}{N}}, \lambda, \beta)$ (which has symmetry S_4) at $\beta^* \approx 1.0387$, as in Figure 10, but now we investigate further the branches which have S_2 symmetry.

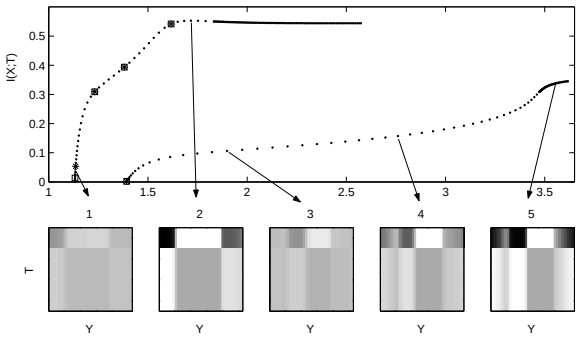

Fig. 16. Depicted here are bifurcating branches with S_3 symmetry from the $q_{\frac{1}{N}}$ branch at the β values 1.133929 and 1.390994 shown in Figure 6. The bottom panels show some of the clusterings along these branches.

branches guaranteed by Theorem 6.2 are subcritical ($\zeta < 0$) or supercritical ($\zeta > 0$). We considered the bifurcating branches from $(q_{\frac{1}{N}}, \lambda^*, \beta^* \approx 1.0387)$ with isotropy group S_3. The numerical results obtained by calculating $\zeta(q_{\frac{1}{N}}, \beta^*, 1, N-1)$ for $N = 2, 3, 4, 5$ and 6 at $\beta^* \approx 1.0387$ are shown in Table I. Supercritical branches are predicted when $N = 2$. Subcritical branches with symmetry S_{N-1} are predicted when $N > 2$. The subcritical bifurcation predicted by the discriminator for the Information Distortion problem (7) for $N = 4$ is shown in Figure 13. This change from supercritical to subcritical branches as N increases is discussed in more detail in section VIII-B.

Figure 15 explores some of the soft clusterings on one of the secondary branches after symmetry breaks from S_3 to S_2.

Figure 16 illustrates clusterings along branches which bifurcate from $q^* = q_{\frac{1}{N}}$ at $\beta > \beta^* = 1.0387$ at the first bifurcation (see Figure 6). By Theorem 4.1, these branches do not give solutions of (7) after a bifurcation. However, we cannot at the moment reject the possibility that these branches continue to a branch that leads to a global maximum of both the optimization problem (3) and the annealing problem (5) as $\beta \to \beta_{\max}$.

Now let us examine how the bifurcations of stationary points to the annealing problem (5), given with respect to the annealing parameter β, yields the bifurcation structure of stationary points of the optimization problem (3) with respect to I_0. Figure 5 depicts a realization of the curve $R(I_0)$ which we produced by solving (3) for G and D as defined for the Information Distortion problem (2),

$$R_H(I_0) = \begin{array}{c} \max_{q \in \Delta} \boldsymbol{H}(\boldsymbol{T}|\boldsymbol{Y}) \\ \boldsymbol{I}(\boldsymbol{X};\boldsymbol{T}) \geq I_0 \end{array},$$

for different values of I_0 using the data set from a mixture of four Gaussians given in Figure 3. Although it appears that the curve is concave, this is not the case,

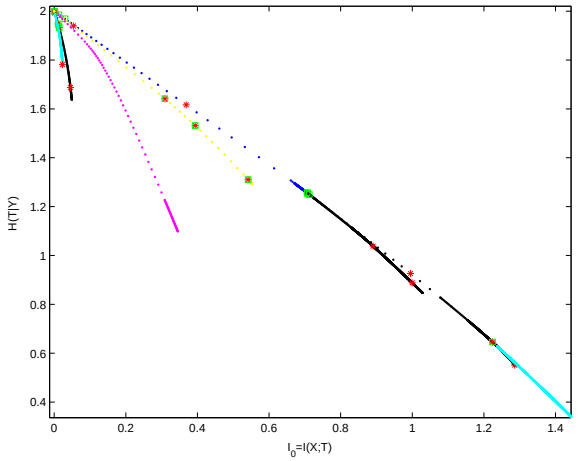

Fig. 17. The bifurcation diagram of stationary points to the problem (2) with respect to I_0.

which we show in section VIII-B. The curve R_H is an envelope for the full structure of all stationary points of (2), which we give in Figure 17. All curves below this envelope correspond to clusterings of the data which are not maxima of the optimization problem (3).

In section IV-B, we showed that at a solution q^* of the optimization problem (3) for some I_0, that the Lagrange multiplier β for the constraint $D(q) \geq I_0$ is unique and non-negative, $\beta := \beta(I_0) \geq 0$. For solutions q^* where $\beta(I_0) > 0$, $D(q^*) = I_0$. When solving (2) for each I_0 (as we did to produce Figure 5), we computed the corresponding Lagrange multiplier $\beta(I_0)$, which is the subcritical curve shown in Figure 13. Turning the Figure sideways shows β as a function of I_0. The existence of the subcritical bifurcation indicates that β is not a one-to-one function of I_0. To produce the bifurcation diagram depicted in Figure 17, we simply plotted $G(q) = \boldsymbol{H}(\boldsymbol{T}|\boldsymbol{Y})$ versus $I_0 = D(q) = \boldsymbol{I}(\boldsymbol{X};\boldsymbol{T})$ for the stationary points (q, λ, β) we found when annealing in

β as in Figure 6.

VIII. Consequences of the Bifurcations

We have provided a theoretical analysis of the bifurcation structure of stationary points for the optimization problem (3) with respect to I_0, and for the corresponding annealing problem (5) with respect to the Lagrange multiplier β. In this section, we turn our attention to consequences of these bifurcations.

First, we relate how the structure and stability of bifurcating branches affects the optimality of stationary points in the problems (3) and the corresponding annealing problem (5). In the second part, we address implications for the convexity of the curve $R(I_0)$ in (3), which includes the rate distortion curve from Information Theory.

A. Stability and Optimality

We now relate the stability of the equilibria (q^*, λ^*, β) in the flow (12) with optimality of the stationary points q^* in each of the optimization problem (3) and the corresponding annealing (5).

First, we give a general theorem which determines when equilibria (q^*, λ^*, β) are not annealing solutions of (5). We will show that, if a bifurcating branch corresponds to an eigenvalue of $d^2\mathcal{L}(q^*)$ changing from negative to positive, then the branch consists of stationary points (q^*, β^*) which are not annealing solutions of (5). By Theorem 4.1, positive eigenvalues of $d^2\mathcal{L}(q^*)$ do not necessarily show that q^* is not an annealing solution of (5), unless the projection of the corresponding eigenvector is in $\ker J_1$. For example, consider the Information Distortion problem (7) applied to the Four Blob problem presented in Figure 3. In this scenario, for the equilibrium (q^*, λ^*, β) of the gradient system (12), $d^2\mathcal{L}(q^*)$ always has at least $K = 52$ positive eigenvalues, even when $d^2F(q^*)$ is negative definite. In fact, for arbitrary annealing problems of the form (5) and for any data set $(\boldsymbol{X}, \boldsymbol{Y})$, $d^2\mathcal{L}(q_{\frac{1}{N}})$ always has at least K positive eigenvalues.

Theorem 8.1: For the bifurcating branch (21) guaranteed by Theorem 6.2, $\boldsymbol{u}$ is an eigenvector of $d^2\mathcal{L}(\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix} + t\boldsymbol{u}, \beta^* + \beta(t))$ for sufficiently small t. Furthermore, if the corresponding eigenvalue is positive, then the branch consists of stationary points which are not annealing solutions to (5).

Proof: We first show that $\boldsymbol{u}$ is an eigenvector of $d^2\mathcal{L}(q^* + t\hat{\boldsymbol{u}}, \lambda^*, \beta + \beta(t))$ for small t. Let $Q = \begin{pmatrix} q \\ \lambda \end{pmatrix}$ so that

$$\mathcal{F}(Q, \beta) := \nabla\mathcal{L}(q^* + q, \lambda^* + \lambda, \beta^* + \beta).$$

Thus, a bifurcation of solutions to $\mathcal{F}(Q, \beta) = \mathbf{0}$ occurs at $(\mathbf{0}, 0)$. For $\gamma \in S_m \times S_n$, $\mathcal{F}(t\boldsymbol{u}, \beta) = \mathcal{F}(t\gamma\boldsymbol{u}, \beta) = \gamma\mathcal{F}(t\boldsymbol{u}, \beta)$, where the first equality follows from Lemma 5.5, and the second equality follows from S_N-equivariance. Hence, $\mathcal{F}(t\boldsymbol{u}, \beta)$ is in $\mathrm{Fix}(S_m \times S_n)$, which is one dimensional with basis vector $\boldsymbol{u}$, showing that $\mathcal{F}(t\boldsymbol{u}, \beta) = h(t, \beta)\boldsymbol{u}$ for some scalar function $h(t, \beta)$. Taking the derivative of this equation with respect to t, we get

$$d_Q\mathcal{F}(t\boldsymbol{u}, \beta)\boldsymbol{u} = d_t h(t, \beta))\boldsymbol{u}, \tag{23}$$

which shows that $\boldsymbol{u}$ is an eigenvector of $d^2\mathcal{L}(q^* + t\hat{\boldsymbol{u}}, \lambda^*, \beta + \beta(t))$, with corresponding eigenvalue $\xi = d_t h(t, \beta)$. Using (11) and letting $\widehat{d^2F} := d^2F(q^* + t\hat{\boldsymbol{u}}, \beta + \beta(t))$, we see that (23) can be rewritten as

$$\begin{pmatrix} \widehat{d^2F} & J^T \\ J & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{u}} \\ \mathbf{0} \end{pmatrix} = \xi \begin{pmatrix} \hat{\boldsymbol{u}} \\ \mathbf{0} \end{pmatrix},$$

which shows that $\widehat{d^2F}\hat{\boldsymbol{u}} = \xi\hat{\boldsymbol{u}}$ and $J\hat{\boldsymbol{u}} = \mathbf{0}$. Thus, $\hat{\boldsymbol{u}} \in \ker J$ is an eigenvector of $d^2F(q^* + t\hat{\boldsymbol{u}}, \beta + \beta(t))$ with corresponding eigenvalue ξ. If $\xi > 0$, the desired result now follows from Theorem 4.1. □

Theorem 8.1 can be used to show that the subcritical bifurcating branch depicted in Figure 13 is not composed of solutions to the annealing problem (7). The condition in Theorem 8.1 is easy to check when using continuation along branches, since the Hessian $d^2\mathcal{L}(q^*)$ is available from the last iteration of Newton's method (see section VII).

At first glance, the fact that the stationary points on the subcritical branch in Figure 13 are not solutions of (7) may be worrisome, since we showed in Lemma 4.2 that $R_H(I_0)$ in (2) is continuous for all $I_0 \in [0, \max_{q\in\Delta} D(q)]$. By the continuity of R_H, for these I_0, there is a solution q^* of (7) and a vector of Lagrange multipliers (λ^*, β^*) such that $(q^*, \lambda^*, \beta^*)$ is a stationary point of the annealing problem (5) (KKT conditions).

However, recall from Theorem 4.1 that there may be solutions of the optimization problem (3) which are not solutions of the corresponding annealing problem (5). Thus, Theorem 8.1 does not address when a stationary point is not optimal for problems of the form (3). Theorem 4.1 indicates how to check for optimality in this case explicitly: a stationary point q^* is optimal for (3) if the Hessian $d^2F(q^*) = d^2(G(q^*) + \beta D(q^*))$ is negative definite on $\ker J_2(q^*)$, and it is not optimal if $d^2F(q^*)$ is not non-positive definite on $\ker J_2(q^*)$.

We next illustrate stationary points of the Information Distortion problem (7) which are not solutions of (2). Consider the subcritical bifurcating branch of stationary points of (7) at $\beta \approx 1.038706$ depicted in Figure 13.

By projecting the Hessian $d^2(G(q^*) + \beta D(q^*))$ onto $\ker J_1$ and also onto $\ker J_2(q_{\frac{1}{N}})$, we determined that (see Figure 13):

1) The clusterings on the "flat" branch $(q_{\frac{1}{N}}, \beta)$ before the bifurcation at $\beta \approx 1.038706$ are solutions to both (2) and (7).
2) The clusterings on the "flat" branch $(q_{\frac{1}{N}}, \beta)$ after the bifurcation at $\beta \approx 1.038706$ are not solutions of either (2) or (7).
3) The clusterings on the subcritical bifurcating branch are solutions of (2) but are **not** solutions of (7).
4) After the branch turns at the saddle-node , the associated clusterings are now solutions of both (2) and (7).

Clearly, the existence of subcritical bifurcations is tied to the existence of saddle-node bifurcations, where the branches turn around and regain optimality in the annealing problem (5). Generally speaking, the generic existence of saddle-node bifurcations (Theorem 6.9) is why annealing does not (necessarily) give a globally optimized clustering of the data for the optimization problem (3) and the corresponding annealing problem (5). It is possible that the global maximum at $\beta_{\max}$ is not connected to the maximum at $\beta = 0$, but that it vanishes in a saddle-node bifurcation at some finite β. If saddle-nodes were not possible, then the global optimizer would be connected by a continuation of stationary points to the uniform solution used as a starting point for the annealing problem.

Using the Information Distortion, Information Bottleneck, and Deterministic Annealing, the solutions corresponding to the symmetry breaking chain from $S_N \rightarrow S_{N-1} \rightarrow ... \rightarrow S_1$ are observed to be optimal, while branches with symmetry $S_m \times S_n$ are suboptimal [10], [11], [14]. This is in contrast to a model of speciation given in [57], [58]. We do not have a general theoretical result which explains this difference.

B. Convexity of the Rate Distortion Curve

We have proved the generic existence of saddle-node bifurcations of stationary points to annealing problems of the form (5). We illustrated subcritical pitchfork and saddle-node bifurcations for the Information Distortion problem (7) in Figure 13. A natural question arises in the mind of the information theorist: Are there implications for the rate distortion curve, defined in [8], [13] as

$$R_{RD}(D_0) := \begin{array}{c} \min_{q \in \Delta} \boldsymbol{I}(\boldsymbol{Y};\boldsymbol{T}) \\ D(\boldsymbol{Y},\boldsymbol{T}) \leq D_0 \end{array} \tag{24}$$

where $D(\boldsymbol{Y},\boldsymbol{T})$ is a *distortion function*. This constrained problem is of the form (3), where $G(q) = -\boldsymbol{I}(\boldsymbol{Y};\boldsymbol{T})$. We now investigate the connection between the existence of saddle-node bifurcations and the convexity of the rate distortion function for $D(\boldsymbol{Y},\boldsymbol{T}) = \boldsymbol{I}(\boldsymbol{X};\boldsymbol{Y}) - \boldsymbol{I}(\boldsymbol{X};\boldsymbol{T})$. This is precisely the relevance-compression function, R_I, defined in (1), in which the constant $\boldsymbol{I}(\boldsymbol{X};\boldsymbol{Y})$ is ignored. Observe that there is a one-to-one correspondence between I_0 and D_0 via $I_0 = \boldsymbol{I}(\boldsymbol{X};\boldsymbol{Y}) - D_0$. For the Information Distortion problem the analogous function is R_H, defined in (2).

It is well known that if the distortion function $D(\boldsymbol{Y},\boldsymbol{T})$ is linear in q, then $R_{RD}(D_0)$ is continuous, strictly decreasing and convex [8], [13]. Since the distortion $D(\boldsymbol{Y},\boldsymbol{T}) = \boldsymbol{I}(\boldsymbol{X};\boldsymbol{Y}) - \boldsymbol{I}(\boldsymbol{X};\boldsymbol{T})$ is not a linear function of q, the convexity proof given in [8], [13] does not generalize to prove that either (1) or (2) is convex. This is why we proved the continuity of both (1) and (2) using other means in Lemma 4.2.

In [10], [44], using variational calculus, it is shown that $\frac{\delta R_I}{\delta D} = -\beta$. Since β is a function of I_0 (KKT conditions), then it seems reasonable to consider $\beta'(I_0)$ where $\beta(I_0)$ is differentiable. We have the following Lemma.

Lemma 8.2: If the functions $R_I(I_0)$, $R_H(I_0)$, and $\beta(I_0)$ are differentiable, then,

$$\frac{dR}{dI_0} = -\beta(I_0) \text{ and } \frac{d^2R}{dI_0^2} = -\frac{d\beta(I_0)}{dI_0}.$$

The relationship between the bifurcations of the stationary points of the annealing problem (5) and the convexity of the curves $R_I(I_0)$ and $R_H(I_0)$ is now clear:

Corollary 8.3: If there exists a saddle-node bifurcation of solutions to the Information Bottleneck problem (6) at $I_0 = I^*$, then $R_I(I_0)$ is neither concave, nor convex in any neighborhood of I^*. Similarly, the existence of a saddle-node bifurcation of solutions to the Information Distortion problem (7) at $I_0 = I^*$ implies that $R_H(I_0)$ is neither concave, nor convex in any neighborhood of I^*.

Proof: The result follows from Lemma 8.2 and the fact that $\frac{d\beta(I_0)}{dI_0}$ changes sign at the saddle-node bifurcation at $I_0 = I^*$. □

Since we have explicitly shown the existence of saddle-node bifurcations for the Information Distortion problem (7) (see $(\beta = 1.0375, D(q) = .0302)$ in Figure 13), then the Corollary shows that R_H in Figure 5 is neither concave nor convex. The convexity of $R_H(I_0)$ changes at $(I_0 = .0302, R_H = 1.9687)$.

Bachrach et al. [62] show that whenever $N > K + 1$, that $R_I(I_0)$ is convex. By Corollary 8.3, this shows that when solving (6) for $\Delta \subset \Re^{NK}$ when $N > K + 1$, that saddle-node bifurcations of stationary points can not exist: only supercritical bifurcating branches are possible.

As mentioned in the preliminaries, we have assumed no constraint in the number of clusters N. Letting $N \geq$

K allows each of the K objects of $\boldsymbol{Y}$ to be classified into its own class, so that there is potentially no compression of the data. One way to find the soft clustering which maximizes either the optimization problem (3) or the annealing problem (5) is by brute force, and to explicitly consider $q \in \Delta \subseteq \Re^{NK}$ for $N \geq K$. For the problem in Figure 3, this is at least a 52^2 dimensional space. Another, more computationally feasible approach is to anneal as is done in [9], [44]. This amounts to "space jumping", where one first considers $N = 2$ clusters (i.e. $q \in \Delta \subset \Re^{2K}$), and then larger N after each bifurcation is detected. At $N = 2$ before jumping, the bifurcation in $\Re^{2K}$ is a break of symmetry from S_2 to S_1. Once the number of potential clusters is increased (to, say, $N = 4$), the bifurcation, now imbedded in $\Re^{4K}$, corresponds to a break in symmetry from either S_4 to $S_2 \times S_2$ or from S_4 to S_3, depending on how the newly introduced clusterings in $\Re^{4K}$ are defined.

Let us consider the brute force approach, where we explicitly consider the bifurcations, when $N \geq K$, and let us compare this to the bifurcations when $N < K$, such as with the numerical results we presented in section VII, where we set $N = 4$. Finding clusterings $q \in \Delta \subset \Re^{NK}$ for such an $N < K$ can be construed as an additional constraint. Perhaps when computing the bifurcation structure for $N \geq K$, the subcritical bifurcations and the saddle-nodes will not occur for general annealing problems of the form (5), mere mathematical anomalies, and not possible when N is large enough, as is the case for the Information Bottleneck.

The argument which Bachrach et al. use to show convexity of R_I [62] relies on the explicit form of $G(q) = -\boldsymbol{I}(\boldsymbol{Y};\boldsymbol{T}) = \boldsymbol{H}(\boldsymbol{T}|\boldsymbol{Y}) - \boldsymbol{H}(\boldsymbol{Y})$ and a geometric proof given by Witsenhausen and Wyner in [63]. This argument does not hold for the Information Distortion curve R_H, since in this case $G(q) = \boldsymbol{H}(\boldsymbol{T}|\boldsymbol{Y})$, and therefore Witsenhausen's result does not apply.

In fact, the saddle-nodes and subcritical bifurcations which we have shown explicitly for the Information Distortion at $N = 4$ still occur when $N \geq K$, which is what we show next.

Consider the bifurcation of stationary points to the Information Distortion problem (7) at $\beta^* \approx 1.0387$ from the uninformative branch $(q_{\frac{1}{N}}, \lambda^*, \beta)$ depicted in Figure 13. This is a bifurcation point for any N. In Table I, we computed the discriminator $\zeta(q_{\frac{1}{N}}, \beta = 1.0387, m = 1, n = N-1)$ when $N \in \{2,3,4,5,6\}$. When $N = 2$, the branch is supercritical (since $\zeta > 0$) , but for $N = 3$, the branch becomes subcritical, and then becomes "more" subcritical as N increases (i.e. $\zeta = \beta''(0)$ becomes more negative). This trend continues for arbitrarily large N. To prove this, we note that $\zeta(q_{\frac{1}{N}}, \beta^*, m, n)$ depends on $B = d^2 f$, $\boldsymbol{b} = d^3 f$, and on $d^4 f$ (see Definition 6.4), all of which depend on $q_{\frac{1}{N}}$ only through N, which follows from the following Lemma.

Lemma 8.4: For the Information Distortion problem (7),

$$d^n f(q_{\frac{1}{N}}) = \frac{N^{n-1}}{(N-1)^{n-1}} d^n f(q_{\frac{1}{N-1}}) = \frac{N^{n-1}}{2^{n-1}} d^n f(q_{\frac{1}{2}})$$

Proof: Direct computation using the derivatives in Lemma 6.6. □

By Lemma 8.4, we have that

$$\begin{aligned} d^2 f(q_{\frac{1}{N}})^- &= B^-_{\frac{1}{N}} = \frac{2}{N} B^-_{\frac{1}{2}} \\ d^3 f(q_{\frac{1}{N}}) &= \boldsymbol{b}_{\frac{1}{N}} = \frac{N^2}{4} \boldsymbol{b}_{\frac{1}{2}}. \end{aligned}$$

The subscripts show whether the matrices are evaluated at $q_{\frac{1}{N}}$ for $N > 2$ or at $q_{\frac{1}{2}}$. Substituting these into (22), and noting that $B_{\frac{1}{N}}$ and $B_{\frac{1}{2}}$ have the same eigenpairs, then we can write $\zeta(q_{\frac{1}{N}}, \beta^*, m, n)$ in terms of functions of $q_{\frac{1}{2}}$ for arbitrarily large N, as

$$\frac{N^3}{8}\left(3\left(1 - \frac{mn}{m^2 - mn + n^2}\right)\boldsymbol{b}^T_{\frac{1}{2}} B^-_{\frac{1}{2}} \boldsymbol{b}_{\frac{1}{2}} - d^4 f(q_{\frac{1}{2}})\right).$$

This shows that if $d^4 f(q_{\frac{1}{2}}) < 0$ and if $3\boldsymbol{b}^T_{\frac{1}{2}} B^-_{\frac{1}{2}} \boldsymbol{b}_{\frac{1}{2}} > |d^4 f(q_{\frac{1}{2}})|$ as in the case for the Information Distortion at $\beta^* = 1.0387$, then for $m = 1$, $n = N-1$ and $N = 2$, the branch with symmetry $S_{N-1} = S_1$ is supercritical. But for N large enough, the $N-1$ bifurcating branches with symmetry S_{N-1} (Theorem 6.2) will become subcritical pitchforks. In a similar scenario, it could be that branches switch from subcritical to supercritical as N increases.

We have demonstrated that even for the case $N \geq K$, subcritical pitchforks and saddle-nodes exist for the Information Distortion. Thus, a potential advantage for using the Information Bottleneck over the Information Distortion method (or any annealing scheme (5)) for clustering data is that for $N > K + 1$, one is guaranteed that only supercritical bifurcations exist, and no saddle-nodes. This is relevant for the computationalist, since the existence of subcritical bifurcations and saddle-nodes can incur significant computational cost when one attempts to find optimal clusterings when using the Basic Annealing Algorithm 3.1.

IX. Conclusions

We have argued that the minimal set of assumptions that constrain the neural coding problem is that it has to be stochastic on a fine scale (due to inherent noise in the neural processing), but deterministic on a large scale (because of the evolutionary enforced need for a consistent response). Therefore a general model for the neural code, which is the correspondence between the inputs and the outputs, is a stochastic map. This map, however, becomes (almost) deterministic, when viewed

on a coarser scale, that is, as a map from clusters of inputs to clusters of outputs. This model of a neural code has a clear advantage over other models of not needing any additional assumptions on the character of the code. In this sense it is the most general such model. There are two main challenges of this approach. First, we needed to find an algorithm that would find the optimal deterministic "skeleton" of the stochastic coding map, or, equivalently, the optimal soft clustering of the set of inputs and the set of outputs that best approximates the (almost) deterministic code. The second challenge is the need for large data sets that contain the rare signals and responses in sufficient number for the stochastic map to be well represented by the data. This second challenge is not particular to our approach. More importantly, our method allows iterative refinement of the coarse groups as more data becomes available and so it scales well with data availability.

The optimality criterion for the best soft clustering comes from information theory. We seek clusters of inputs and outputs such that the induced relationship between the two clustered spaces preserves the maximum amount of the original mutual information between the inputs and outputs. It has been shown that the globally optimal solution is deterministic [12] and that the combinatorial search for the solution is NP-complete [64] and therefore computationally not feasible for large data sets. The lack of a fast algorithm that would compute the global maximum of the mutual information cost function led to the implementation of annealing as the standard algorithm for such optimization problems [9]–[11], [14].

Even though the implementation is straightforward and annealing usually finds biologically feasible solutions, our goal was to understand the annealing algorithm in more detail, the reasons for this success, and the potential for failure.

Using bifurcation theory with symmetries we have shown that the soft clustering which optimizes the cost function of interest is not an annealing solution after a subcritical bifurcation. Thus, although the curve of optimal solutions to the cost function is continuous with respect to the annealing parameter, the curve of annealing solutions is discontinuous at a subcritical bifurcation. However, since the annealing procedure is guaranteed to find a local solution eventually, the subcritical branch must turn and become optimal at some later saddle-node bifurcation, which we have shown occur generically for this class of problems.

We also discuss the number and the character of refinements that the annealing solutions undergo as a function of the annealing parameter. Generically occurring symmetry breaking pitchforks are in contrast to the symmetry breaking transcritical bifurcations of solutions to an S_N invariant model for speciation in [57], [58]. For the Information Distortion, Information Bottleneck, and Deterministic Annealing methods, the solutions corresponding to the symmetry breaking chain from $S_N \rightarrow S_{N-1} \rightarrow ... \rightarrow S_1$ are observed to be locally optimal, while branches with symmetry $S_m \times S_n$ are not [10], [11], [14]. This is another difference with the model of speciation given in [57], [58].

Previously we have shown that the annealing solution converges to a deterministic local maximum [12]. The main problem of whether the globally optimal solution can always be reached by the annealing process from the uniform solution remains open. This is because we can not rule out either the existence of saddle-node bifurcations which do not connect to the original uniform solution, or the existence of locally sub-optimal bifurcating branches which do connect the uniform solution to the global one. To our best knowledge, for the Information Distortion, Information Bottleneck, and Deterministic Annealing methods, such bifurcating branches have never been observed [10], [11], [14], although rigorous results are still lacking. We hasten to add that proving that the globally optimal solution can always be reached by the annealing process from the uniform solution would be equivalent to an $NP = P$ statement and therefore such a proof is unlikely. Despite this, the relatively straightforward annealing problem can be a fruitful method for approaching NP-hard problems. Although each iteration of annealing is more computationally intensive than the cost function evaluation needed by the combinatorial search to solve the NP-hard deterministic clustering, the overall complexity of the locally optimal annealing solution branch grows only linearly with the number of classes. We have shown here that there are only $N - 1$ bifurcations for N clusters. Compare this to the combinatorial explosion of the size of the search space in the deterministic clustering. Thus, even though we believe it unlikely that it can be proven that a branch of locally optimal annealing solutions connects from the uniform solution to the global deterministic optimum in all cases, the profoundness of such a result should still encourage work in this area.

In addition our results can be of interest for Information Theory. In contrast to rate distortion theory where the rate distortion curve is always convex, the analogous function for the Information Bottleneck and Information Distortion methods is non-convex when a saddle-node bifurcation occurs. The difference stems from the fact that both in the Information Bottleneck and Information Distortion methods the distortion function is the mutual information, which is a non-linear function of the quantizer. In Deterministic Annealing and Rate Distortion theory, the distortion function is an expectation of a pairwise distance function and hence *linear* in the quantizer.

Future work

Future works involves expanding these analytic results in two directions. We would like to extend the results from the current one-sided clustering or quantization to joint quantization of both stimulus and response spaces [20]. Joint quantization, which clusters both sides of a system jointly, has a cost function that is invariant to $(S_M \times S_N)$, where S_M acts on the space of clustered stimuli, and S_N acts on the space of clustered responses. This added complexity poses different challenges in the analytic development. Initial observations in this area show that the simplest symmetry breaking is of the kind $S_M \times S_N \rightarrow S_{M-1} \times S_{N-1}$ and not for example to $S_M \times S_{N-1}$ or $S_{M-1} \times S_N$. This is easy to understand intuitively - if either soft clustering is uniform, the cost function does not increase as no classes are resolved. However, subsequent bifurcations of the joint problem are not well understood. Specifically, we do not know at what stages a finer quantization of one space occurs relative to the other and why. Multi-quantization, another extension of the Information Bottleneck problem [65], [66], used for network analysis, has an even richer symmetry structure, with the cost function being invariant under the symmetry group $\bigotimes_i S_i$, and its bifurcation structure is completely unknown.

The approach could be further extended as a model of brain development. It shows a very definite and dramatic way in which parts of the sensory world that were previously unresolved can be separated into discriminable portions, by taking a part of a system that is uniform in its properties and splitting it into portions that perform different functions, while maximizing the information between the sensory environment ($\boldsymbol{X}$) and the neural representation ($\boldsymbol{Y}$). This is similar to the latest ideas of how a portion of the brain, previously dedicated to the same task, bifurcates into distinct parts delegated to different tasks [67], [68]. This is accomplished by the duplication of a homeobox gene which causes a replication of a whole neural subsystem, which that gene regulates. For example, it is hypothesized that the multitude of primate visual cortices [69] emerged in this manner. Applying the distortion-based methods described here to questions about evolutionary development of brain structures could provide firm quantitative foundations to such theories of brain evolution. If, for instance, the Right Fusiform Gyrus (RFG) area and the Inferior Temporal (IT) cortex emerged by duplication of a single cortical region, both cortices likely performed the same function of visual object recognition. Given enough time and evolutionary pressure, they eventually bifurcated to the current state, in which the IT cortex performs general visual object recognition, while the RFG is specialized to face discrimination.

More generally, specific realizations of this general method have been used in very diverse fields with the same goal in mind: break down a complex system into simpler components in a manner that is consistent with the structure of the complex system, then study the components separately. This is essentially the process of reductionism, used successfully in the sciences, but posed here in a formal manner, and supplied with tools that can automate it. This implies that the distortion based procedures outlined here could be used as a general system identification and analysis methodology. These methods are general enough to be used for models of arbitrary input-output systems: quantize to a simpler system, characterize the simpler system, then refine the quantization for a finer description.

ACKNOWLEDGEMENT

The authors would like to gratefully acknowledge their colleagues John Miller and Zane Aldworth in the Center for Computation Biology at Montana State University in Bozeman for sharing experimental results which we use to illustrate the applicability of our method to neurophysiological data from the cricket cercal sensory system.

REFERENCES

[1] F. Rieke, D. Warland, R. R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the neural code*. The MIT Press, 1997.

[2] R. E. Kass, V. Ventura, and E. N. Brown, "Statistical issues in the analysis of neural data," *J. Neurophys.*, vol. 94, pp. 8–25, 2005.

[3] L. Paninski, J. Pillow, and E. Simoncelli, "Maximum likelihood estimation of a stochastic integrate-and-fire neural model," *Neur. Comp.*, vol. 17, pp. 1480–1507, 2005.

[4] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. Litke, E. Simoncelli, and E. Chichilnisky, "Spatio-temporal correlations and visual signaling in a complete neuronal population." *Nature*, vol. 454, pp. 995–999, 2008.

[5] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 623–656, 1948.

[6] H. B. Barlow, "Possible princilples underlying the transformation of sensory messages," in *Sensory Communications*, W. A. Rosenblith, Ed. MIT Press, Cambridge, MA, 1961.

[7] T. W. Kjaer, J. A. Hertz, and B. J. Richmond, "Decoding cortical neuronal signals: Network models, information estimation and spatial tuning." *J. Comp. Neurosci*, vol. 1, no. 1-2, pp. 109–139, 1994.

[8] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley Series in Communication, 1991.

[9] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. MIT Press, 2000, vol. 12, pp. 617–623.

[10] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," The 37th annual Allerton Conference on Communication, Control, and Computing, 1999.

[11] A. G. Dimitrov and J. P. Miller, "Neural coding and decoding: communication channels and quantization," *Network: Computation in Neural Systems*, vol. 12, no. 4, pp. 441–472, 2001.

[12] T. Gedeon, A. E. Parker, and A. G. Dimitrov, "Information distortion and neural coding," *Canadian Applied Mathematics Quarterly*, vol. 10, no. 1, pp. 33–70, 2003.

[13] R. M. Gray, *Entropy and Information Theory*. Springer-Verlag, 1990.

[14] K. Rose, "Deteministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proc. IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998.

[15] W. Bialek, R. R. de Ruyter van Steveninck, and N. Tishby, "Efficient representation as a design principle for neural coding and computation," in *Information Theory, 2006 IEEE International Symposium on*, 2006, pp. 659–663.

[16] G. Chechick, A. Globerson, N. Tishby, M. Anderson, E. D. Young, and I. Nelken, "Group redundancy measures reveals redundancy reduction in the auditory pathway," in *Advances in Neural Information Processing Systems*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14, 2002.

[17] A. G. Dimitrov and J. P. Miller, "Analyzing sensory systems with the information distortion function," in *Pacific Symposium on Biocomputing 2001*, R. B. Altman, Ed. World Scientific Publishing Co., 2000.

[18] A. G. Dimitrov, J. P. Miller, Z. Aldworth, T. Gedeon, and A. E. Parker, "Analysis of neural coding through quantization with an information-based distortion measure." *Network: Computations in Neural Systems*, vol. 14, pp. 151–176, February 2003.

[19] A. G. Dimitrov, J. P. Miller, Z. Aldworth, and T. Gedeon, "Non-uniform quantization of neural spike sequences through an information distortion measure," *Neurocomputing*, vol. 38-40, pp. 175–181, 2001.

[20] B. Mumey, A. Sarkar, T. Gedeon, A. G. Dimitrov, and J. P. Miller, "Finding neural codes using random projections," *Neurocomputing*, vol. 58-60, pp. 19–25, 2004.

[21] E. Schneidman, N. Brenner, N. Tishby, R. R. de Ruyter van Steveninck, and W. Bialek, "Universality and individuality in a neural code," in *NIPS*, 2000, pp. 159–165. [Online]. Available: citeseer.ist.psu.edu/305279.html

[22] U. Alon, N. B. D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, 1999.

[23] L. Chen, "Multiple protein structure alignment by deterministic annealing," Aug. 2003, pp. 609–610.

[24] L. Chen, T. Zhou, and Y. Tang, "Protein structure alignment by deterministic annealing," *Bioinformatics*, vol. 21, no. 1, pp. 51–62, 2005.

[25] R. M. Hecht and N. Tishby, "Extraction of relevant speech features using the information bottleneck method," in *Proceedings of "InterSpeech, (Lisbon) 2005"*, 2005.

[26] K.-M. Lee, T.-S. Chung, and J.-H. Kim, "Global optimization of clusters in gene expression data of dna microarrays by deterministic annealing," *Genomics and Informatics*, vol. 1, no. 1, pp. 20–24, 2003.

[27] S. O'Rourke, G. Chechik, R. Friedman, and E. Eskin, "Discrete profile comparison using information bottleneck," *BMC Bioinformatics*, vol. 7, no. Suppl 1, p. S8, 2006, http://www.biomedcentral.com/1471-2105/7/s1/s8.

[28] S. F. Taylor, N. Tishby, and W. Bialek, "Information and fitness," arXiv:0712.4382v1 [q-bio.PE], Dec. 2007.

[29] A. Zhang, *Advanced Analysis of Gene Expression Microarray Data*. Singapore: World Scientific Publishing, 2006.

[30] P. Andritsos, R. Miller, and P. Tsaparas, "Information-theoretic tools for mining database structure from large data sets," in *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. ACM, New York, NY, 2004, pp. 731 – 742, http://doi.acm.org/10.1145/1007568.1007650.

[31] A. E. Parker and T. Gedeon, "Bifurcation structure of a class of s_n-invariant constrained optimization problems," *Journal of Dynamics and Differential Equations*, vol. 16, no. 3, pp. 629–678, July 2004, second special issue dedicated to Shui-Nee Chow.

[32] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, pp. 939–952, 1982.

[33] M. Kanou and T. A. Shimozawa, "Threshold analysis of cricket cercal interneurons by an alternating air-current stimulus." *J. Comp. Physiol. A*, vol. 154, pp. 357–365, 1984.

[34] J. P. Miller, G. A. Jacobs, and F. E. Theunissen, "Representation of sensory information in the cricket cercal sensory system. I. Response properties of the primary interneurons." *J. Neurophys*, vol. 66, pp. 1680–1689, 1991.

[35] F. Theunissen, J. C. Roddey, S. Stufflebeam, H. Clague, and J. P. Miller, "Information theoretic analysis of dynamical encoding by four primary sensory interneurons in the cricket cercal system," *J. Neurophy.*, vol. 75, pp. 1345–1359, 1996.

[36] F. E. Theunissen and J. P. Miller, "Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning curve width of four primary interneurons." *J. Neurophysiol.*, vol. 66, pp. 1690–1703, 1991.

[37] D. Bodnar, J. P. Miller, and G. A. Jacobs, "Anatomy and physiology of identified wind-sensitive local interneurons in the cricket cercal sensory system." *J Comp Physiol A*, vol. 168, pp. 553–564, 1991.

[38] G. A. Jacobs and R. K. Murphey, "Segmental origins of the cricket giant interneuron system." *J Comp Neurol*, vol. 265, pp. 145–157, 1987.

[39] H. Clague, F. Theunissen, and J. P. Miller, "The effects of adaptation on neural coding by primary sensor interneurons in the cricket cercal system," *J. Neurophysiol.*, vol. 77, pp. 207–220, 1997.

[40] J. C. Roddey, B. Girish, and J. P. Miller, "Assessing the performance of neural encoding models in the presence of noise." *J. Comp. Neurosci.*, vol. 8, pp. 95–112, 2000.

[41] D. S. Reich, F. Mechler, K. P. Purpura, and J. D. Victor, "Interspike intervals, receptive fields, and information encoding in primary visual cortex." *J. Neurosci.*, vol. 20, pp. 1964–74, 2000.

[42] J. D. Victor, "Binless strategies for estimation of information from neural data," *Phys. Rev. E.*, vol. 66, p. 051903, 2002.

[43] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 03)*, 2003.

[44] N. Slonim, "The information bottleneck: Theory and applications," Doctoral Thesis, Hebrew University, 2002.

[45] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory II*. Springer Verlag, New York, 1988.

[46] D. S. Dummit and R. M. Foote, *Abstract Algebra*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[47] G. James and A. Kerber, *The Representation Theory of the Symmetric Group*, ser. The Encyclopedia of Mathematics and Applications, G.-C. Rota, Ed. Addison-Wesley, Reading, Massachusetts, 1981, vol. 16.

[48] A. Dimitrov, T. Gedeon, B. Mumey, R. Snider, A. E. Parker, and J. P. Miller, "Derivation of natural stimulus feature set using a data-driven model," in *International Conference on Computational Science*, ser. Lecture Notes in Computer Science, P. M. A. Sloot, D. Abramson, A. V. Bogdanov, J. Dongarra, A. Y. Zomaya, and Y. E. Gorbachev, Eds., vol. 2660. Springer, 2003, pp. 337–345.

[49] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, New York, 2000.

[50] A. Vanderbauwhede, "Local bifurcation and symmetry," Habilitation Thesis, Rijksuniversiteit Gent., 1980.

[51] G. Cicogna, "Symmetry breakdown from bifurcation," *Lettere Al Nuevo Cimento*, vol. 31, pp. 600–602, 1981.

[52] ——, "Bifurcation and symmetries," *Bollettino Un. Mat. Ital.*, pp. 787–796, 1982.

[53] M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory I*. Springer Verlag, New York, 1985.

[54] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Perseus Books: Cambridge, Massachusetts, 1998.

[55] M. Golubitsky and I. Stewart, *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*. Birkhauser Verlag, Boston, 2002.
[56] J. R. Schott, *Matrix Analysis for Statistics*. John Wiley and Sons, New York, 1997.
[57] J. Cohen and I. Stewart, "Polymorphism viewed as phenotypic symmetry-breaking," in *Non-linear Phenomena in Biological and Physical Sciences*, S. K. Malik, M. K. Chandrashekaran, and N. Pradhan, Eds. Indian National Science Academy: New Delhi, 2000, pp. 1–63.
[58] I. Stewart, "Self-organization in evolution: a mathematical perspective," *Philosophical Transactions of The Royal Society*, vol. 361, pp. 1101–1123, 2003.
[59] W. J. Beyn, A. Champneys, E. Doedel, W. Govaerts, Y. A. Kuznetsov, and B. Sandstede, "Numerical continuation and computation of normal forms," in *Handbook of Dynamical Systems III*. World Scientific, 1999.
[60] E. Doedel, H. B. Keller, and J. P. Kernevez, "Numerical analysis and control of bifurcation problems in finite dimensions," *International Journal of Bifurcation and Chaos*, vol. 1, pp. 493–520, 1991.
[61] H. B. Keller, "Numerical solutions of bifurcation and nonlinear eigenvalue problems," in *Applications of Bifurcation Theory*, P. Rabinowitz, Ed. Academic Press, New York, 1977, pp. 359–384.
[62] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," *COLT*, pp. 595–609, 2003.
[63] H. S. Witsenhausen and A. D. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. on Information Theory*, vol. IT-21, pp. 493–501, September 1975.
[64] B. Mumey and T. Gedeon, "Optimal mutual information quantization is np-complete," Neural Information Coding (NIC) workshop, Snowbird UT, 2003.
[65] G. Elidan and N. Friednam, "Learning hidden variable networks: The information bottleneck approach," *J. Machine Learning Research*, vol. 6, pp. 81–127, 2005.
[66] N. Slonim, N. Friedman, and N. Tishby, "Multivariate information bottleneck," *Neural Computation*, vol. 18, pp. 1739–1789, 2006.
[67] L. Krubitzer and K. Huffman, "Arealization in the neocortex of mammals: Genetic and epigenetic contributions to the phenotype." *Brain, Behavior and Evolution*, vol. 55, pp. 322–335, 2000.
[68] L. Krubitzer and J. Kaas, "The evolution of the neocortex in mammals: how is phenotypic diversity generated?" *Curr. Opin. Neurobiol.*, vol. 15, pp. 444–453, 2005.
[69] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex." *Cereb. Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

Albert Parker received a B.S. in mathematics from Bridgewater State College in Massachusetts, an M.S. in mathematics from the University of Vermont, and an M.S. in statistics and a Ph.D. in mathematics from Montana State University. He completed a post-doc with Curt Vogel and the Center for Adaptive Optics based in Santa Cruz in 2005. He is currently a post-doc with Colin Fox and the New Zealand Institute of Mathematics at the University of Auckland, and a Research Engineer and statistician with the Center for Biofilm Engineering at Montana State University. His research interests include iterative sampling from high dimensional densities, and modeling of complex biological systems.

Alex Dimitrov received his B.S. in Physics from Sophia University in Bulgaria in 1991, and an M.S. in Physics and a Ph.D. in Applied Mathematics from the University of Chicago in 1998. In 1998 he joined joined the Center for Computational Biology at Montana State University, where he is currently an Assistant Professor in the Department of Cell Biology and Neuroscience. His research interests include information-theoretic and probabilistic approaches to neural computing and cognitive processes, mathematical neuroscience, and non-linear neuronal models.

Tomáš Gedeon received his B.A. and M.Sc. in Mathematics in 1989 at Comenius University in Bratislava, Slovak Republic. After receiving Ph.D. in Mathematics from Georgia Institute of Technology in 1994, he spent a one-year post-doc at Northwestern University. In 1995 he joined the Department of Mathematical Sciences at Montana State University, where he is currently a Professor of Mathematics and a member of Center for Computational Biology. His research interests include information based methods of clustering and dynamics of complex systems in neuroscience and gene regulation.