Entropy **2012**, *14*, 456-479; doi:10.3390/e14030456

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

The Mathematical Structure of Information Bottleneck Methods

Tomáš Gedeon [1,*], Albert E. Parker [2] and Alexander G. Dimitrov [3]

[1] Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717, USA

[2] Center for Biofilm Engineering, Montana State University, Bozeman, MT 59717, USA; E-Mail: parker@math.montana.edu

[3] Department of Mathematics and Science Programs, Washington State University Vancouver, Vancouver, WA 98686, USA; E-Mail: alex.dimitrov@vancouver.wsu.edu

* Author to whom correspondence should be addressed; E-Mail: gedeon@math.montana.edu; Tel.: +1-406-994-5359; Fax: +1-406-994-1789.

Received: 2 December 2011; in revised form: 7 February 2012 / Accepted: 24 February 2012 / Published: 1 March 2012

Abstract: Information Bottleneck-based methods use mutual information as a distortion function in order to extract relevant details about the structure of a complex system by compression. One of the approaches used to generate optimal compressed representations is by annealing a parameter. In this manuscript we present a common framework for the study of annealing in information distortion problems. We identify features that should be common to any annealing optimization problem. The main mathematical tools that we use come from the analysis of dynamical systems in the presence of symmetry (equivariant bifurcation theory). Through the compression problem, we make connections to the world of combinatorial optimization and pattern recognition. The two approaches use very different vocabularies and consider different problems to be "interesting". We provide an initial link, through the *Normalized Cut Problem*, where the two disciplines can exchange tools and ideas.

Keywords: information distortion; spontaneous symmetry breaking; bifurcations; phase transition

---

1. Introduction

Our goal in this paper is to investigate the mathematical structure of Information Distortion methods. There are several approaches to computing the best quantization of the data, and they differ in the algorithms used, the data they are applied to, and the functions that are optimized by the algorithms. We will concentrate on the annealing method applied to two different functions: the Information Bottleneck cost function [1] and the Information Distortion function [2]. By formalizing a common framework in which to study these two problems, we will exhibit common features of, as well as differences between, the two cost functions. Moreover, the differences and commonalities we will highlight are based on the underlying structural properties of these systems rather then on the philosophy behind their derivation. All results that we present are valid for any system characterized by a probability distribution and in this sense they present fundamental structural results.

On a more concrete level, our goal is to understand why the annealing algorithms now in use work as well as they do, but also to suggest improvements to these algorithms. Some results which have been observed numerically are not expected when applying annealing to a general cost function. We want to ask what is the special feature of these systems that cause such results.

Our final goal is to provide a bridge between the world of combinatorial optimization and pattern recognition, and the world of dynamical systems in mathematics. These two areas have different goals, different sets of "natural questions" and, perhaps most crucially, different vocabularies. We want this manuscript to contribute to bridging this gap, as we believe that both sides have developed interesting and powerful techniques that can be used to expand the knowledge of the other side.

We close by introducing the optimization problems we will study. Both approaches attempt to characterize a system of interest $(\boldsymbol{X}, \boldsymbol{Y})$ defined by a probability $p(\boldsymbol{X}, \boldsymbol{Y})$ by quantizing (discretizing) one of the variables ($\boldsymbol{Y}$ here) into a reproduction variable $\boldsymbol{T}$ with few elements. One of the problems stems from the Information Distortion approach to neural coding [2,3],

$$\max_{q \in \Delta} F_H(q) := \boldsymbol{H}(T|Y) + \beta \boldsymbol{I}(X;T) \tag{1}$$

where $\boldsymbol{H}(\cdot)$ is the conditional entropy, and $\boldsymbol{I}(\cdot;\cdot)$ is the mutual information [4]. The other problem is from the Information Bottleneck approach to clustering [1,5,6]

$$\max_{q \in \Delta} F_I(q) := -\boldsymbol{I}(T;Y) + \beta \boldsymbol{I}(X;T) \tag{2}$$

which has been used for document classification [7,8], gene expression [9], neural coding [10,11], stellar spectral analysis [12], and image time-series data mining [13].

The variables (quantizers) q are conditional probabilities $q := q(t|y)$ and Δ is the space of all appropriate conditional probabilities. We will explain all of the details in the main text, but we want to sketch the basic idea of the annealing approach here. Since both functions $\boldsymbol{H}(T|Y)$ and $-\boldsymbol{I}(T;Y)$ are concave, when $\beta = 0$, both problems (1) and (2) admit a homogeneous solution $q(t|y) = 1/N$, where N is the number of elements in $\boldsymbol{T}$. Starting at this solution and increasing β slowly, the optimal solution, or quantizer, q will undergo a series of phase transitions (bifurcations) as a function of β. We will show that the parameter β, at which the first phase transition takes place, does not depend on the number of elements in the reproduction variable $\boldsymbol{T}$. Annealing in the temperature-like parameter β terminates either

at some predefined finite value of β, or goes to $\beta = \infty$. It is this process and its phase transitions that we consider in this contribution.

1.1. Outline of the Mathematical Contributions

In Section 2 we start with the optimization problems and identify the space of variables over which optimization takes place. Since these variables are constrained, we use Lagrange multipliers to eliminate equality constraints. We also present some results about convexity and concavity of the cost functions.

Our first main question is whether the approach of deterministic annealing [14] can be used for these optimization problems. Rose and his collaborators have shown that, if the distortion function in certain class of optimization problems is taken to be the Euclidean distance, the phase transitions of the annealing function can be computed explicitly. More precisely, the first phase transition can be computed explicitly, since the quantizer value is known and only the value of the temperature at which this quantizer loses stability has to be computed. In general, an implicit formula relating critical temperature and the critical quantizer at which phase transition occurs can be computed.

In Section 4 we will show that the same calculations can be done for our optimization problems. We relate the critical value of β at which the uniform quantizer $q_{\frac{1}{N}}$ loses stability to a certain eigenvalue problem. This problem can be solved effectively off-line and thus the annealing procedure can start from this value of β rather then at $\beta = 0$. As a consequence, we also show that in both optimization problems considered here, the quantizer $q_{\frac{1}{N}}$ is a local maximum for all $\beta \in [0, 1]$. In complete analogy with deterministic annealing, our results extend beyond phase transitions off $q_{\frac{1}{N}}$. As we show in Section 5, the aforementioned eigenvalue problem implicitly relates all critical values of the parameter β to critical values of the quantizer q.

We study more closely the first phase transition in Section 6. We show that the eigenvector corresponding to this phase transition solves the Approximate Normalized Cut problem for some graphs with vertices corresponding to elements of $\boldsymbol{Y}$. These graphs have considerable intuitive appeal.

In [15–17] we studied the subsequent phase transitions more closely, using bifurcation theory with symmetries. We summarize the main results here as well. The symmetry of our problems comes from the fact that the cost function is invariant under the relabeling of the elements of the representation variable $\boldsymbol{T}$. Such a symmetry is characterized by the permutation group S_N and its subgroups. Since this is a structural symmetry, it does not require the symmetry of the underlying probability distribution $p(\boldsymbol{X}, \boldsymbol{Y})$. These results are valid for arbitrary probability distributions.

2. Mathematical Formulation of the Problem

The variables q over which the optimization takes place are conditional probabilities $q(t|y)$. In order for the problems (1) and (2) to be well defined, we must fix the number of elements of $\boldsymbol{T}$. Let this number be N and let the number of elements in Y be K. Then there are NK conditional probabilities $q(t|y)$ which satisfy

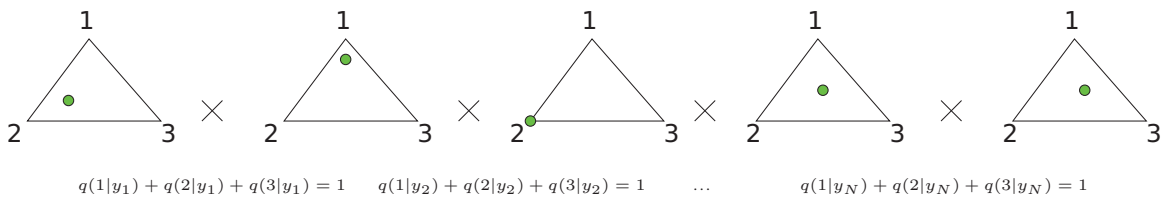
$$\sum_t q(t|y) = 1 \quad \text{for all } y \tag{3}$$

These equations form an equality constraint on the maximization problems (1) and (2). We also have to satisfy inequality constraints $q(t|y) \geq 0$ since $q(t|y)$ are probabilities. We notice that, for a fixed y,

the space of admissible values $q(t|y)$ is the unit $N-1$ simplex Δ^{N-1} in $\mathbf{R}^N$. We denote this simplex as Δ_y, to also indicate that it is related to variable y, and suppressing the dimension for simplicity of notation. It follows from (3) that the set of all admissible values of $q(t|y)$ is a product of such simplices (see Figure 1), which we call Δ,

$$\Delta := \Delta_{y_1} \times \Delta_{y_2} \times \ldots \times \Delta_{y_K}$$

Figure 1. The space Δ of admissible vectors q can be represented as a product of simplices, one simplex for each $y_i \in Y$. The figure shows the case when the reproduction variable T has three elements ($N = 3$). Each triangle represents a unit simplex in $\mathbf{R}^3$ and the constraint $q(1|y_i) + q(2|y_i) + q(3|y_i) = 1,\ \ q(t|y_i) \geq 0$. The green point represents the position of a particular q. To clarify the illustration: The part of q in simplex 4, $q(t|3)$, is almost deterministic (shown at a vertex), while the next q, $q(t|4)$ is almost uniform (shown almost at the center of simplex 4).



At this point we want to comment on a successful implementation of the annealing algorithm by Slonim and Tishby [6]. In their approach they start the annealing procedure with $N = 2$ at $q(t|y) = \frac{1}{2}$ for all $t = 1, 2$ and y at $\beta = 0$. After increasing β they split $q(t|y)$, for $t = 1, 2$, into two parts, $q(t^1|y)$ and $q(t^2|y)$ by setting

$$q(t^1|y) = q(1|y)(\frac{1}{2} + \epsilon\alpha(t, y)), \quad q(t^2|y) = q(1|y)(\frac{1}{2} - \epsilon\alpha(t, y))$$

where $\alpha(t, y)$ is random perturbation and ϵ is small. If under the fixed point iteration at new value of β the values $q(t^1|y)$ and $q(t^2|y)$ converge to the same value ($1/4$ in this case), then the process is repeated; if, on the other hand, these values diverge, a presence of a bifurcation is asserted. Note, that this process changes N from 2 to 4 repeatedly. This changes the optimization problem, because the space of admissible quantizers q doubled. It is not clear a priori that phase transition detected in problem with $2K$ variables also occurs at the same value of β in problem with $4K$ variables. Numerically, however, this seems to be the case not only at the first phase transition, but at every phase transition. One of the results of Section 4 will be an explanation of this phenomena. We will show that the parameter β, at which the first phase transition takes place, does not depend on the number of elements in the reproduction variable $\boldsymbol{T}$. This provides a justification for Slonim's algorithm, at least for the first phase transition.

Since the optimization problems (1) and (2) are constrained, we first form the Lagrangian,

$$\mathcal{L} = F + \sum_{k=1}^{K} \lambda_k \left(\sum_{t=1}^{N} q(t|y_k) - 1 \right) \tag{4}$$

which incorporates the vector of Lagrange multipliers λ, imposed by the equality constraints from the constraint space Δ. Here $F = F_H$ for (1) or $F = F_I$ for (2),

Lemma 2.1 *The function $\boldsymbol{H}(T|Y)$ is a strictly concave function of $q(t|y)$ and the functions $\boldsymbol{I}(X;T)$ and $\boldsymbol{I}(Y;T)$ are convex, but not strictly convex, functions of $q(t|y)$.*

Proof. For concavity of $\boldsymbol{H}(T|Y)$ and convexity of $\boldsymbol{I}(Y;T)$, see [2]. Proof of the convexity of $\boldsymbol{I}(X;T)$ is analogous.

This Lemma implies that for $\beta = 0$ in both (1) and (2), there is a trivial solution $q(t|y) = 1/N$ for all t and y. We denote this solution as $q_{\frac{1}{N}}$.

What we want to emphasize here is that $\boldsymbol{I}(Y;T)$ and $\boldsymbol{I}(X;T)$ are not strictly convex functions. Recall that a function f is *convex* provided

$$sf(\boldsymbol{u}) + (1-s)f(\boldsymbol{v}) \leq f(s\boldsymbol{u} + (1-s)\boldsymbol{v}) \tag{5}$$

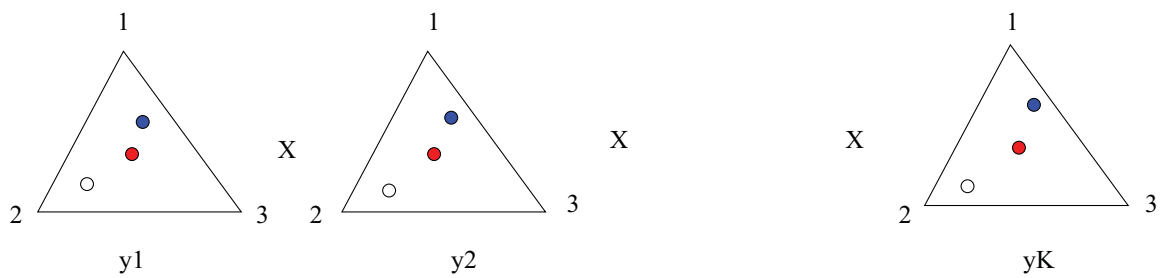
for all $\boldsymbol{u}, \boldsymbol{v}$ and $0 \leq s \leq 1$. The function f is *strictly convex* if the inequality in (5) is strict for $\boldsymbol{u} \neq \boldsymbol{v}$ and $0 < s < 1$.

To show that $\boldsymbol{I}(Y;T)$ is not strictly convex, we take $q(t|y_k) = a_t$ independent on k (see Figure 2). In order for this q to satisfy $q \in \Delta$ we require that numbers a_t are chosen with $\sum_t a_t = 1$. Using the facts that $p(y,t) = q(t|y)p(y)$ and $p(t) = \sum_y p(y,t) = a_t \sum_y p(y)$, we evaluate at $q = q_a$ the function

$$\begin{aligned} \boldsymbol{I}(T;Y) &= \sum_{y,t} p(y,t) \log \frac{p(y,t)}{p(y)p(t)} \\ &= \sum_{y,t} a_t p(y) \log \frac{a_t p(y)}{p(y) a_t} \\ &= 0 \end{aligned}$$

This implies that in Δ there is an $N-1$ dimensional linear space spanned by vectors $a = (a_1, a_2, \ldots, a_N)$ with $\sum_t a_t = 1$, such that for all q in this space $\boldsymbol{I}(T;Y)(q) = 0$. Since $\boldsymbol{I}(T;Y) \geq 0$, this function does not have a unique minimum and thus is not strictly convex. □

Figure 2. The function $I(\boldsymbol{T};\boldsymbol{Y})$ is not strictly convex. There are three vectors q depicted in the figure. The red point in the middle of each simplex represents the point $q_{\frac{1}{N}}$ with $N = 3$. The blue point and the white points have the property that $q(t|y)$ does not depend on y, only on t. At all three points the function $I(\boldsymbol{T};\boldsymbol{Y})$ is equal to zero.



This result has consequences for the function $F_I(q)$. As we will see in Lemma 3.1, $F_I(q) = 0$ at all points where $\boldsymbol{I}(T;X)(q) = 0$. This lack of strict convexity has important consequences for phase transitions for F_I. Since $\boldsymbol{H}(T|Y)$ is strictly concave, this problem will not affect the function F_H.

Maxima of (1) and (2) are critical points of the Lagrangian, that is, points q where the gradient of (4) is zero. We now switch our search from maxima to critical points of the Lagrangian. Obviously, minima and saddle points are also critical points and therefore we must always check whether a given critical point is indeed a maximum of the original problem (1) or (2). We want to use the language of bifurcation theory which deals with qualitative changes in the structure of system dynamics given by differential equations or maps. Therefore we will now reformulate the optimization problems (1) and (2) as a system of differential equations under a gradient flow,

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \nabla_{q,\lambda}\mathcal{L}(q, \lambda, \beta) \tag{6}$$

In this equation, the $NK \times 1$ vector q representing the quantizer, and the $K \times 1$ vector of the Langrange multipliers (see Equation (4)) are viewed as functions of some independent variable s, which parameterizes curves of solutions $(q(s), \lambda(s))$ to either (1) or (2). Thus, the derivatives implicit in $(\dot{q}, \dot{\lambda})$ are with respect to s. The critical points of the Lagrangian are the equilibria of (6), since those are the places where the gradient of $\mathcal{L}$ is equal to zero. By the same token, the maxima of (1) and (2) correspond to stable (in q) equilibria of the gradient flow (6). More technically, these are points for which the Hessian d^2F is negative definite on the kernel of the Jacobian of the constraints [18,19].

As β increases from 0, the solution $q_{\frac{1}{N}}$ is initially a maximum of (1) and (2). We are interested in the smallest value of β, say $\beta = \beta^*$, where $q_{\frac{1}{N}}$ ceases to be a maximum. This corresponds to a change in the number of critical points in the neighborhood of $q_{\frac{1}{N}}$ as β passes through $\beta = \beta^*$. The value β^* is called a bifurcation value and the new sets of critical points emanating from $q_{\frac{1}{N}}$ are called bifurcating branches. This question can be posed at any other point besides $q_{\frac{1}{N}}$ as well: When do such bifurcations happen? We will formulate the answer in the language of differential equations. If the linearization of the flow at equilibrium has eigenvalues with nonzero real part, the implicit function theorem implies that this equilibrium exists for all values of the parameter in a small neighbourhood. Since the number of equilibria then does not change locally, this implies that a bifurcation does not occur at such a point. Therefore, a necessary condition for bifurcation is that the real part of some eigenvalue of the linearization of the flow at an equilibrium crosses zero [20]. Therefore, we need to consider eigenvalues of the $(NK + K) \times (NK + K)$ Hessian $d^2\mathcal{L}$. Since $d^2\mathcal{L}$ is a symmetric matrix, bifurcation can only be caused by one of its real eigenvalues crossing zero, and therefore we must find values of (q, β) at which $d^2\mathcal{L}$ is singular, or, equivalently, has a nontrivial kernel.

The form of $d^2\mathcal{L}$ is simple:

$$d^2\mathcal{L} = \begin{bmatrix} B_1 & 0 & \dots & I \\ 0 & B_2 & \dots & I \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & B_N & I \\ I & I & \dots & 0 \end{bmatrix}$$

where I is the identity matrix and B_i is

$$B_i := \frac{\partial^2\mathcal{L}}{\partial q(t_i|y_k)\partial q(t_i|y_l)} = \frac{\partial^2 F}{\partial q(t_i|y_k)\partial q(t_i|y_l)}$$

The block diagonal matrix consisting of all matrices B_i represents the matrix of second derivatives (Hessian) of F.

In [15,17] we showed that there are two types of generic bifurcations: saddle-node, in which a set of equilibria emerge simultaneously, and pitchfork-like, in which new equilibria emanate from an existing equilibrium. The first kind of bifurcation corresponds to a value of β, and corresponding q, for which $d^2\mathcal{L}$ is singular, but d^2F is non-singular; the second kind of bifurcation happens at β and q where d^2F is singular. Our primary focus here is on bifurcations off $q_{\frac{1}{N}}$, and more generally off an existing branch, we will focus on the second kind of bifurcation. Therefore, we will investigate only the case in which the eigenvalues of the smaller $NK \times NK$ Hessians d^2F_H and d^2F_I are zero to determine the location of pitchfork-like bifurcations.

2.1. Derivatives

In order to simplify notation we will denote

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

To determine d^2F_H and d^2F_I from (1) and (2), we need to determine the quantities $d^2\boldsymbol{H}(T|Y)$, $d^2\boldsymbol{I}(X;T)$ and $d^2\boldsymbol{I}(Y;T)$. The first two were computed in [2]:

$$\frac{\partial^2 \boldsymbol{H}(T|Y)}{\partial q_{\eta l}\partial q_{\nu k}} = -\frac{1}{\ln 2}\frac{p(y_k)}{q_{\nu k}}\delta_{\nu\eta}\delta_{kl} \tag{7}$$

and

$$\frac{\partial^2 \boldsymbol{I}(X;T)}{\partial q_{\eta l}\partial q_{\nu k}} = \frac{\delta_{\nu\eta}}{\ln 2}\left(\sum_i \frac{p(x_i,y_k)\,p(x_i,y_l)}{\sum_k q_{\nu k}p(x_i,y_k)} - \frac{p(y_k)p(y_l)}{\sum_k q_{\nu k}p(y_k)}\right) \tag{8}$$

where $\delta_{\nu\eta} = 1$ if $\nu = \eta$ and zero otherwise. We computed the derivative of the term $-d^2\boldsymbol{I}(Y;T)$ in [19]

$$\frac{-\partial^2 \boldsymbol{I}(Y;T)}{\partial q_{\eta l}\partial q_{\nu k}} = \frac{\delta_{\nu\eta}}{\ln 2}\left(\frac{p(y_k)p(y_l)}{p(\nu)} - \frac{\delta_{lk}p(y_k)}{q_{\nu k}}\right) \tag{9}$$

The formulas (7)–(9) show that we can factor $\delta_{\nu\eta}$ out of both $d^2F_H = d^2\boldsymbol{H}(T|Y) + \beta d^2\boldsymbol{I}(X;T)$ and $d^2F_I = -d^2\boldsymbol{I}(Y;T) + \beta d^2\boldsymbol{I}(X;T)$. This implies that the $NK \times NK$ matrices d^2F_H and d^2F_I are block diagonal, with N blocks, with each $K \times K$ block B_i corresponding to a particular value (class) of the reconstruction variable $\boldsymbol{T}$.

2.2. Symmetries

The optimization problems (1) and (2) have symmetry. We capitalize on this symmetry to solve these problems better. The symmetries arise from the structure of $q \in \Delta$ and from the form of the functions F_H and F_I: permuting subvectors q_ν does not change the value of F_H and F_I. This symmetry is characterized as an invariance under the action of the permutation group, S_N, or one of its subgroups $S_{l_1} \times S_{l_2} \times \ldots \times S_{l_z}$, $\sum l_k = N$.

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 solutions, to (1) and (2).

In [15] we clarified 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:

Proposition 2.2 *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 $q\in\Delta\subset\mathbf{R}^{NK}$ is decomposed into N subvectors $q_\nu\in\mathbf{R}^K$.

The annealing problems (1) and (2) satisfy this Proposition. Any F satisfying Proposition 2.2 has the following properties.

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

3. The Kernel at a Bifurcation

In this section we investigate and compare the kernels of the $NK\times NK$ Hessians d^2F_I and d^2F_B.

3.1. The Kernel of the Information Bottleneck

Our first observation is that F_I is highly degenerate as a consequence of the fact that both $\boldsymbol{I}(Y;T)$ and $\boldsymbol{I}(X;T)$ are not strictly convex in q.

Lemma 3.1 *Select a collection of numbers $a_1,\ldots,a_K$ such that $a_i\geq 0$ and $\sum_{i=1}^K a_i=1$. Let $q_a\in\mathbf{R}^{NK}$ be a vector consisting of vectors $q_a^j\in\mathbf{R}^N$, $j=1,\ldots K$ such that q_a^j is a constant vector with entries a_j. In other words, select $q(t|y)=a_t$ independent on y. Then*

$$F_I(q_a)=F_I(q_{\frac{1}{N}})=0 \quad \textit{for all } \beta$$

Proof. We evaluate at $q=q_a$ the function

$$\begin{aligned}
-\boldsymbol{I}(T;Y)+\beta\boldsymbol{I}(T;X) &= -\sum_{y,t} p(y,t)\log\frac{p(y,t)}{p(y)p(t)}+\beta\sum_{x,t} p(x,t)\log\frac{p(x,t)}{p(x)p(t)}\\
&= -\sum_{y,t} a_t p(y)\log\frac{a_t p(y)}{p(y)a_t}+\beta\sum_{x,t,y} q(t|y)p(x,y)\log\frac{\sum_y q(t|y)p(x,y)}{p(x)a_t}\\
&= \beta\sum_{x,t,y} q(t|y)p(x,y)\log\frac{a_t\sum_y p(x,y)}{p(x)a_t}\\
&= 0
\end{aligned}$$

Since $q_{\frac{1}{N}}$ is a particular case of q_a, the Lemma is proved. □

Now we prove a generalization of this Lemma. We will say that q has symmetry described by $S_{l_1} \times S_{l_2} \times \ldots \times S_{l_z}$ (a subgroup of S_N) if

$$q = (q_1, \ldots, q_1, q_2 \ldots q_2, \ldots, q_z, \ldots, q_z)^T \tag{10}$$

where z is the total number of "blocks" of sub-vectors, with the sub-vector q_i repeating l_i times in the ith block. At such vector q, the matrix d^2F has z groups of blocks B_i, and all blocks in each group are identical. In particular, the first l_1 blocks $B_1 = \ldots = B_{l_1}$ are the same, then next l_2 blocks are the same, and so on.

Theorem 3.2 *Consider an arbitrary pair (q, β), where q admits a symmetry $S_{l_1} \times S_{l_2} \times \ldots \times S_{l_z}$. Then, at a fixed value of β, there is a linear manifold of dimension*

$$(l_1 - 1) + (l_2 - 1) + \ldots + (l_z - 1)$$

passing through q, such that the function F_I is constant on this manifold.

Proof. The quantizer q must take the form given by (10). Let

$$w = (c_1 q_1, c_2 q_1, \ldots, c_{l_1} q_1, q_2 \ldots q_2, \ldots, q_z, \ldots, q_z)^T$$

where the constants c_i are nonnegative and $\sum_i c_i = l_1$. We will show that

$$F_I(q) = F_I(w)$$

We separate $F_I(w)$ into two parts

$$\begin{aligned}
F_I(w) &= -\sum_{y,t} p(y,t) \log \frac{p(y,t)}{p(y)p(t)} + \beta \sum_{x,t} p(x,t) \log \frac{p(x,t)}{p(x)p(t)} \\
&= \sum_{t \le l_1} \left[-\sum_{y} p(y,t) \log \frac{p(y,t)}{p(y)p(t)} + \beta \sum_{x} p(x,t) \log \frac{p(x,t)}{p(x)p(t)} \right] \\
&+ \sum_{t > l_1} \left[-\sum_{y} p(y,t) \log \frac{p(y,t)}{p(y)p(t)} + \beta \sum_{x} p(x,t) \log \frac{p(x,t)}{p(x)p(t)} \right] \\
&= F_I^1(w) + F_I^2(w)
\end{aligned}$$

Since the vectors w and q agree for all $t > l_1$ we have

$$F_I^2(q) = F_I^2(w)$$

Observe first that

$$\begin{aligned}
F_I^1(y) &= -\sum_{y, t \le l_1} q_1(t|y) p(y) \log \frac{q_1(t|y)p(y)}{p(y)p(t)} + \beta \sum_{x,y,t \le l_1} q_1(t|y) p(x,y) \log \frac{\sum_y q_1(t|y) p(x,y)}{p(x)p(t)} \\
&= l_1 \left(-\sum_{y} q_1(t|y) p(y) \log \frac{q_1(t|y)p(y)}{p(y)p(t)} + \beta \sum_{x,y,} q_1(t|y) p(x,y) \log \frac{\sum_y q_1(t|y) p(x,y)}{p(x)p(t)} \right) \\
&= l_1 G^1(y)
\end{aligned}$$

where $G^1(y)$ is the function inside the parentheses on the last line. Now we evaluate $F_I^1(w)$

$$\begin{aligned} F_I^1(w) &= -\sum_{y,t\le l_1} c_t q_1(t|y)p(y)\log\frac{c_t q_1(t|y)p(y)}{p(y)c_t p(t)} + \beta\sum_{x,y,t\le l_1} c_t q_1(t|y)p(x,y)\log\frac{c_t\sum_y q_1(t|y)p(x,y)}{p(x)c_t p(t)} \\ &= -\sum_{t\le l_1} c_t\sum_y p(t,y)\log\frac{p(t,y)}{p(y)p(t)} + \beta\sum_{t\le l_1} c_t\sum_{x,y} q_1(t|y)p(x,y)\log\frac{\sum_y q_1(t|y)p(x,y)}{p(x)p(t)} \\ &= \sum_{t\le l_1} c_t G^1(y) \end{aligned}$$

Since $\sum_{t\le l_1} c_t = l_1$ by assumption, we have $F_I^1(y) = F_I^1(w)$ and therefore

$$F_I(w) = F_I(y)$$

Since $\sum_{t\le l_1} c_t = 1$, the solutions w form a l_1-dimensional linear manifold. The same argument can be applied to $q_2,\ldots,q_z$ to finish the proof. □

Now we spell out the consequences of this degeneracy for $\dim\ker d^2F_I$. Since the manifolds of constant value of F_I are linear, the second derivative along these manifolds must vanish. Note that in Theorem 3.2 we required that the solutions lie in Δ. Therefore, $\ker d^2\mathcal{L}$ must vanish along this manifold, rather then $\ker d^2F_I$. In the following paragraphs, our first two results are concerned with $\ker d^2F_I$, the third—with $\ker d^2\mathcal{L}$.

First we will show the result for a single block of d^2F_I.

Lemma 3.3 *Fix an arbitrary quantizer q and an arbitrary class ν. Then the $K\times 1$ vector $q_\nu := q(\boldsymbol{T}=\nu|Y)$ is in the kernel of the νth block B_ν of d^2F_I for any value of β.*

Proof. To show that the vector q_ν, defined in the statement of the Lemma, is in the kernel of $d^2F_I^\nu(q)$, the νth-block of d^2F_I, we compute the lth row of this matrix. From (8) and (9) we see that

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

This shows that q_ν is in the kernel of block ν of d^2F_I. □

Corollary 3.4 *For an arbitrary pair (q,β), the dimension of $\ker d^2F_I$ is at least N, the number of classes of $\boldsymbol{T}$.*

Proof. Given q_ν as in Lemma 3.3, we define vectors $\{\boldsymbol{u}_i\}_{i=1}^N \in \mathbf{R}^{NK}$ by

$$\boldsymbol{u}_1 = \begin{pmatrix} q_1 \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \quad \boldsymbol{u}_2 = \begin{pmatrix} \mathbf{0} \\ q_2 \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \ldots, \quad \boldsymbol{u}_N = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ q_N \end{pmatrix}$$

By Lemma 3.3, $\{\boldsymbol{u}_i\}_{i=1}^N \in \ker d^2 F_I(q,\beta)$. Clearly these vectors are linearly independent. □

Now we investigate the consequences of Theorem 3.4 for the dimensionality of the kernel of $d^2\mathcal{L}$.

Theorem 3.5 *Consider an arbitrary pair (q,β), where q admits a symmetry $S_{l_1} \times S_{l_2} \times \ldots \times S_{l_z}$. Then the dimension of* $\ker d^2\mathcal{L}_I$ *at such point is at least*

$$d(q) := (l_1 - 1) + (l_2 - 1) + \ldots + (l_z - 1)$$

Proof. Since q admits the stated symmetry it has the form (10). There are $l_1 - 1$ vectors of the form

$$\boldsymbol{v}(1) := \boldsymbol{u}_1 - \boldsymbol{u}_l, \quad l = 2, \ldots, l_1.$$

Direct computation shows that, since $\boldsymbol{u}_i \in \ker d^2 F_I$, each vector $\boldsymbol{v}(1) \in \ker d^2\mathcal{L}$. Similar argument shows that there are $l_i - 1$ vectors $\boldsymbol{v}(i) \in \ker d^2\mathcal{L}$ for $i = 2, \ldots, z$.

Corollary 3.6 *If q has no symmetry, i.e., $q = (q_1, q_2, \ldots, q_K)$ and all $q_i \neq q_j$ for $i \neq j$, then the dimension of* $\ker d^2\mathcal{L}_I$ *is $d(q) = 0$. In other words, $d^2\mathcal{L}_I$ is non-singular.*

Lemma 3.7 *At a phase transition (q,β) of system (2) we have* $dim\ \ker d^2\mathcal{L}_I \geq d(q) + 1$.

Proof. This follows from the fact that the degeneracy of the kernel of dimension $d(q)$ is a consequence of the existence of a $d(q)$-dimensional manifold of solutions on which F_I is constant. The existence of kernel with this dimension therefore does not indicate a phase transition. For that, the kernel must be at least $d(q) + 1$-dimensional. □

3.2. The Kernel of the Information Distortion

We want to contrast the degeneracy of $d^2 F_I$ with the non-degeneracy of $d^2 F_H$.

Theorem 3.8 *There is* no *value of q such that the matrix $d^2 F_H(q,\beta)$ is singular for all β in some interval.*

Proof. If $\exists q$ such that for each β in some interval I, $d^2 F_H$ is singular, then

$$\left(d^2\boldsymbol{H}(T|Y) + \beta d^2\boldsymbol{I}(T;Y)\right) W(\beta) = \mathbf{0}$$

for some $NK \times 1$ vector valued function $W(\beta)$. Thus, $\frac{1}{\beta}W(\beta) = -(d^2\boldsymbol{H}(T|Y))^{-1} d^2\boldsymbol{I}(T;Y) W(\beta)$, from which it follows that $W(\beta)$ is a $\frac{1}{\beta}$-eigenvector of the fixed matrix $-(d^2\boldsymbol{H}(T|Y))^{-1} d^2\boldsymbol{I}(T;Y)$ for every $\beta \in I$. This is a contradiction, since $-(d^2\boldsymbol{H}(T|Y))^{-1} d^2\boldsymbol{I}(T;Y)$ has at most NK distinct eigenvalues. □

Lemma 3.9 *At the phase transition (q,β) for system (1) we have* $dim\ \ker d^2 L_H \geq 1$.

4. Bifurcations off the Uniform Solution $q_{\frac{1}{N}}$

In this section we want to illustrate the close analogy between Deterministic Annealing with Euclidean distortion function and Information Distortion. Our goal is to find values of (q, β) for which the problems (1) and (2) undergo phase transition. Given the joint probability distribution $p(x, y)$, we can find the values of β explicitly for $q = 1/N$ in terms of eigenvalues of a certain stochastic matrix. Secondary phase transitions that occur at values of $q \neq 1/N$ cannot be computed explicitly and we must resort to numerical continuation along the branches of equilibria. An eigenvalue problem, implicitly relating quantities $q \neq 1/N$ and β at which phase transition occurs, can still be obtained. This is completely analogous to results of Rose [14] for a different class of optimization problems.

We start by deriving a general eigenvalue problem which computes the pair (q, β). We seek to compute (q, β) for which the $NK \times NK$ matrix of second derivatives d^2F has a nontrivial kernel. This is a necessary condition for a bifurcation to occur. We first discuss the Hessian of (1), $d^2F_H := d^2F_H(q, \beta)$, evaluated at q and at some value of the annealing parameter β. Thus, we need to find pairs (q, β) where d^2F_H has a nontrivial kernel. For that, we solve the system

$$d^2F_H\boldsymbol{w} = (d^2\boldsymbol{H}(T|Y) + \beta d^2\boldsymbol{I}(X;T))\boldsymbol{w} = 0 \tag{11}$$

for any nontrivial $\boldsymbol{w} \in \mathbf{R}^{NK}$. We rewrite (11) as an eigenvalue problem,

$$(-d^2\boldsymbol{H}(T|Y))^{-1}d^2\boldsymbol{I}(X;T)\boldsymbol{w} = \frac{1}{\beta}\boldsymbol{w} \tag{12}$$

Since $-\boldsymbol{I}(Y;T) = \boldsymbol{H}(T|Y) - \boldsymbol{H}(T)$, then, for the Hessian d^2F_I, we find pairs (q, β) for which

$$d^2F_I\boldsymbol{w} = (d^2\boldsymbol{H}(T|\boldsymbol{Y}) - d^2\boldsymbol{H}(T) + \beta d^2\boldsymbol{I}(X;T))\boldsymbol{w} = 0$$

Multiplying by $(-d^2\boldsymbol{H}(T|Y))^{-1}$ leads to a generalized eigenvalue problem

$$(-d^2\boldsymbol{H}(T|Y))^{-1}d^2\boldsymbol{I}(X;T)\boldsymbol{w} = (I - (-d^2\boldsymbol{H}(T|Y))^{-1}d^2\boldsymbol{H}(T))\frac{1}{\beta}\boldsymbol{w} \tag{13}$$

Since $d^2\boldsymbol{H}(T|Y)$ is diagonal, we can explicitly compute the inverse

$$[(-d^2\boldsymbol{H}(T|Y))^{-1}]_{(\nu k),(\eta l)} = \delta_{\eta\nu}\delta_{lk}\ln 2\frac{q_{\nu k}}{p(y_k)}. \tag{14}$$

Next, we compute the explicit forms of the $NK \times NK$ matrices

$$U(q) := (-d^2\boldsymbol{H}(T|Y))^{-1}d^2\boldsymbol{I}(X;T)$$

and

$$A(q) := (-d^2\boldsymbol{H}(T|Y))^{-1}d^2\boldsymbol{H}(T)$$

Since both of these matrices are block diagonal, with one block corresponding to a class of $\boldsymbol{T}$, we will compute the νth block of these matrices. Using (7)–(9) we get that the $(l, k)^{th}$ element of the νth block of $U(q)$ is

$$u^{\nu}_{lk} := \sum_i \frac{p(x_i, y_k)p(x_i, y_l)}{p(x_i, \nu)p(y_l)}q_{\nu l} - \frac{p(y_k)}{p(\nu)}q_{\nu l}$$

and the $(l,k)^{th}$ element of the νth block of $A(q)$ is

$$a_{lk}^{\nu} := \frac{p(y_k)q_{\nu l}}{p(\nu)} \tag{15}$$

We observe that the matrix $U(q)$ can be written as $U(q) = Q(q) - A(q)$, where the $(l,k)^{th}$ element of the νth block of matrix $Q(q)$ is

$$r_{lk}^{\nu} := \sum_i \frac{p(x_i,y_k)p(x_i,y_l)q_{\nu l}}{p(x_i,\nu)p(y_l)} \tag{16}$$

Therefore the problems (12) and (13) become generalized eigenvalue problems,

$$(Q(q) - A(q))\boldsymbol{w} = \lambda \boldsymbol{w} \quad \text{for system (1)} \tag{17}$$

and

$$(Q(q) - A(q))\boldsymbol{w} = (I - A(q))\lambda \boldsymbol{w} \quad \text{for system (2)} \tag{18}$$

respectively.

In the eigenvalue problems (17) and (18), the matrices $Q(q)$ and $A(q)$ change with q. On the other hand, we know that for all $\beta \in [0,\hat{\beta})$ for some $\hat{\beta} > 0$, both problems (1) and (2) have a maximum at the uniform solution $q_{\frac{1}{N}}$ [19], *i.e.*, when $q(t|y) = 1/N$ for all t and y. We now determine when this extremum ceases to be the maximum.

We evaluate matrices $Q(q)$ and $A(q)$ at $q_{\frac{1}{N}}$ to get

$$r_{lk}^{\nu} = \sum_i \frac{p(x_i,y_k)p(x_i,y_l)}{p(x_i)p(y_l)} = \sum_i p(y_k|x_i)p(x_i|y_l)$$

and

$$a_{lk}^{\nu} = p(y_k)$$

Let $\mathbf{1}$ be a vector of ones in $\mathbf{R}^N$. We observe that

$$A(q_{\frac{1}{N}})\mathbf{1} = \mathbf{1}$$

and that the lth component of $Q(q_{\frac{1}{N}})\mathbf{1}$

$$\begin{aligned}
[Q(q_{\frac{1}{N}})\mathbf{1}]_l &= \sum_k \sum_i p(y_k|x_i)p(x_i|y_l) \\
&= \sum_i p(x_i|y_l) \sum_k p(y_k|x_i) \\
&= \sum_i p(x_i|y_l) \\
&= 1
\end{aligned}$$

Therefore, we obtain one particular eigenvalue-eigenvector pair $(0,\mathbf{1})$ of the eigenvalue problems (17) and (18):

$$Q(q_{\frac{1}{N}}) - A(q_{\frac{1}{N}})\mathbf{1} = 0 \quad \text{and } Q(q_{\frac{1}{N}}) - A(q_{\frac{1}{N}})\mathbf{1} = 0 = (I - A(q_{\frac{1}{N}}))\mathbf{1}$$

Since the eigenvalue λ corresponds to $1/\beta$, this solution indicates a bifurcation at $\beta = \infty$. We are interested in finite values of β.

Theorem 4.1 *Let $1 = \lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots \lambda_K$ be eigenvalues of a block of the matrix $Q(q_{\frac{1}{N}})$. Then the solution $q_{\frac{1}{N}}$ of the maximization problems (1) and (2) ceases to be a maximum at $\beta = \frac{1}{\lambda_2}$. The corresponding eigenvector to λ_2 (and all λ_k for $k \geq 2$) is perpendicular to the vector $p := (p(y_1), p(y_2), \ldots, p(y_n))^T$.*

Proof. We note first that the range of matrix $A(q_{\frac{1}{N}})$ is the linear space spanned by vector $\mathbf{1}$, and its kernel is the linear space

$$W := \{\boldsymbol{w} \in \mathbf{R}^N \mid \langle \boldsymbol{p}, \boldsymbol{w} \rangle = 0\}$$

where $\boldsymbol{p} = (p(y_1), \ldots, p(y_n))$.

We now check that the space W is invariant under the matrix $Q(q_{\frac{1}{N}})$, which means that $Q(q_{\frac{1}{N}})W \subset W$. It will then follow that all eigenvectors of $Q(q_{\frac{1}{N}}) - A(q_{\frac{1}{N}})$, except $\mathbf{1}$, belong to W and are actually eigenvectors of $Q(q_{\frac{1}{N}})$ alone. So, assume $\boldsymbol{w} = (w_1, \ldots, w_N) \in W$, which means

$$\sum_k w_k p(y_k) = 0$$

We compute the l-th element $[Q(q_{\frac{1}{N}})\boldsymbol{w}]_l$ of vector $Q(q_{\frac{1}{N}})\boldsymbol{w}$:

$$[Q(q_{\frac{1}{N}})\boldsymbol{w}]_l = \sum_k \sum_i p(y_k|x_i)p(x_i|y_l)w_k$$

The vector $Q(q_{\frac{1}{N}})\boldsymbol{w}$ belongs to W if, and only if, its dot product with $\boldsymbol{p}$ is zero. We compute the dot product

$$\begin{aligned} Q(q_{\frac{1}{N}})\boldsymbol{w} \cdot \boldsymbol{p} &= \sum_{l,i,k} p(y_k|x_i)p(x_i|y_l)w_k p(y_l) \\ &= \sum_{i,k} p(y_k|x_i)w_k \sum_l p(x_i|y_l)p(y_l) \\ &= \sum_k w_k \sum_i p(y_k|x_i)p(x_i) \\ &= \sum_k w_k p(y_k) \end{aligned}$$

The last expression is zero, since $w \in W$.

This shows that all other eigenvectors of $Q(q_{\frac{1}{N}}) - A(q_{\frac{1}{N}})$, except $\mathbf{1}$, belong to W and are eigenvectors of $Q(q_{\frac{1}{N}})$ alone. Since bifurcation values β are reciprocal to eigenvalues λ_i, the result follows. □

Corollary 4.2 *The value β at which the first phase transition occurs does not depend on the number of classes, N. It only depends on the properties of the matrix Q.*

Observe that, since d^2F has N identical blocks at $q_{\frac{1}{N}}$ and each block has a zero eigenvalue at $\beta = 1/\lambda_i$, we get that

$$\dim \ker d^2 F_H \geq N$$

at such a value of β. This is a consequence of the symmetry. For the Information Bottleneck function F_I, as a consequence of Lemma 3.4, each block has a zero eigenvalue for any value of β. At the instance

of the first phase transition at $(q_{\frac{1}{N}}, \beta = 1/\lambda_i)$, each block of F_I admits an additional zero eigenvalue, and therefore

$$\dim \ker d^2 F_I \geq 2N$$

Notice that the matrix $Q(q_{\frac{1}{N}})$ is transpose of a stochastic matrix, since the sum of all elements in the lth row,

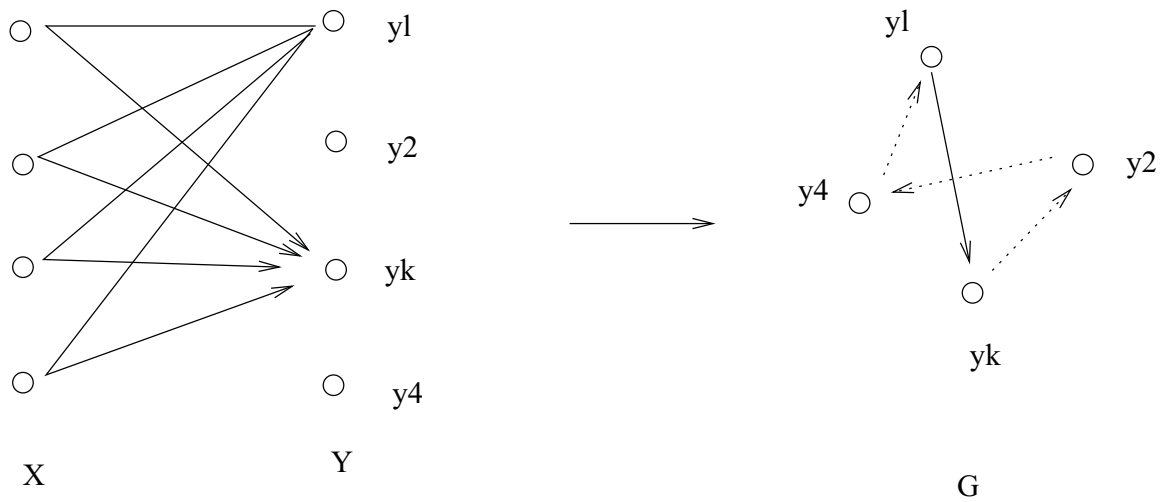
$$\sum_k \sum_i p(y_k|x_i)p(x_i|y_l) = 1$$

Therefore all eigenvalues satisfy $-1 \leq \lambda_i \leq 1$. In particular, $\lambda_2 \leq 1$. This proves

Corollary 4.3 *For both problems (1) and (2), the solution $q_{\frac{1}{N}}$ is stable for all $\beta \in [0, 1]$.*

Remark 4.4 The matrix $Q := Q(q_{\frac{1}{N}})$ has an interesting structure and interpretation (see Figure 3). Let G be a graph with vertices y_k and let the oriented edge $y_l \to y_k$ have a weight $\sum_i p(y_k|x_i)p(x_i|y_l)$. The matrix Q is the transpose of a Markov transition matrix on the elements $\{y_k\}$. The weight attached to each edge is a sum of all the contributions along all the paths $y_l \to x_i \to y_k$ over all i. This structure is key to associating the annealing problem to the normalized cut problem discussed in Section 6.

Figure 3. The graph G with vertices labelled by elements of Y. The oriented edges in G have weights obtained from the weights in the graph of the joint distribution $p(X, Y)$. The weight of the solid edge in G is computed by summing the edges on the left side of the picture.



5. Bifurcations in the General Case

To find the discrete values of the pairs (q, λ) that solve the eigenvalue problems (17) and (18) for a general value of q, we transform the problems (17) and (18) one more time. Let C be a block diagonal matrix of size $NK \times NK$ whose ν^{th} block is a $K \times K$ diagonal matrix, $\mathrm{diag}(p(y_k))$. Instead of the eigenvalue problem (17), we consider

$$C(Q(q) - A(q))C^{-1}\boldsymbol{w} = \lambda \boldsymbol{w} \tag{19}$$

and instead of the problem (18), we consider

$$C(Q(q) - A(q))C^{-1}\boldsymbol{w} = C(I - A)C^{-1}\lambda\boldsymbol{w} \tag{20}$$

Clearly, these problems have the same eigenvalues as the problems (17) and (18) respectively, and the eigenvectors are related via the diagonal matrix C.

Let

$$V(q) := CQ(q)C^{-1} \text{ and } B(q) := CA(q)C^{-1}$$

Then the $(l,k)^{th}$ element of the νth block of the $NK \times NK$ matrix $V(q)$ is

$$v_{lk}^{\nu} := \sum_i \frac{p(x_i, y_k)p(x_i, y_l)q_{\nu l}}{p(x_i, \nu)p(y_k)} \tag{21}$$

and for the $NK \times NK$ matrix B, we have that

$$b_{lk}^{\nu} = \frac{p(\nu, y_l)}{p(\nu)} \tag{22}$$

Lemma 5.1 *The matrix $V(q)$ is stochastic for any value of q.*

Proof. We sum the kth column of $V(q)$ to get

$$\begin{aligned}
\sum_l v_{lk} &= \sum_l \sum_i \frac{p(x_i, y_k)p(x_i, y_l)q_{\nu l}}{p(x_i, \nu)p(y_k)} \\
&= \sum_i \frac{p(x_i, y_k)}{p(x_i, \nu)p(y_k)} \sum_l p(x_i, y_l)q_{\nu l} \\
&= \sum_i \frac{p(x_i|y_k)p(y_k)}{p(x_i, \nu)p(y_k)} \sum_l p(x_i|y_l)p(y_l, \nu) \\
&= \sum_i \frac{p(x_i|y_k)}{p(x_i, \nu)} p(x_i, \nu) \\
&= \sum_i p(x_i|y_k) \\
&= 1
\end{aligned}$$

□

Lemma 5.2 *The νth block of the matrices in the eigenvalue problems (19) and (20) have solutions $\{\lambda = 0, P_\nu = (p(y_1, \nu), \ldots, p(y_K, \nu))^T\}$ that correspond to the eigenvalue-eigenvector pair $(1, P_\nu)$ of the stochastic matrix $V^\nu(q)$, the νth block of the $NK \times NK$ matrix $V(q)$. All other eigenvalues are eigenvalues of the problem*

$$V^\nu(q)u = \lambda u$$

and the corresponding eigenvectors lie in the space $W^\nu = \{u \in \mathbf{R}^K \mid \sum_j [u]_j = 0\}$.

Proof. To show the first part of the Lemma, we multiply the lth row of νth block of $V(q) - B(q)$ by the vector $P_\nu := (p(y_1, \nu), \ldots, p(y_K, \nu))^T$. We get

$$\begin{aligned}
(V(q) - B(q))^\nu P_\nu &= \sum_{i,k} \frac{p(x_i, y_k)p(x_i, y_l)q_{\nu l}p(y_k, \nu)}{p(x_i, \nu)p(y_k)} - \sum_k \frac{p(y_l)q_{\nu l}p(y_k, \nu)}{p(\nu)} \\
&= \sum_i \frac{p(x_i, y_l)q_{\nu l}}{p(x_i|\nu)p(\nu)} \sum_k \frac{p(x_i|y_k)p(y_k)p(y_k, \nu)}{p(y_k)} \\
&\quad - \frac{p(y_l)q_{\nu l}}{p(\nu)} \sum_k p(y_k, \nu) \qquad (23) \\
&= \sum_i \frac{p(x_i, y_l)q_{\nu l}}{p(x_i|\nu)} p(x_i|\nu) - p(y_l)q_{\nu l} \\
&= q_{\nu l} \sum_i p(x_i, y_l) - p(y_l)q_{\nu l} = 0 \\
&= 0
\end{aligned}$$

Observe that the above computation shows that $V^\nu(q)P_\nu = P_\nu$, and so P_ν is a 1-eigenvector of the stochastic matrix $V^\nu(q)$. This finishes the first part of the proof.

To prove the second case, we will show that $W^\nu = \ker B^\nu(q)$ and that W^ν is invariant under $V^\nu(q)$. To see that $W^\nu = \ker B^\nu(q)$, it is enough to realize that every row of $B^\nu(q)$ is a multiple of $\mathbf{1}$, the $K \times 1$ vector of ones. In other words, b^ν_{kl} from (22) is independent of k. Clearly, $\mathbf{1}$ is perpendicular to W^ν. Since the range of $B^\nu(q)$ is one-dimensional, $\dim \ker B^\nu(q) = K - 1$. It follows easily that

$$\ker\ B^\nu(q) = W^\nu \qquad (24)$$

To finish the proof, we show that W is invariant under any stochastic matrix, and in particular to the matrix $V^\nu(q)$. Let S be a $K \times K$ stochastic matrix. Then, if $\boldsymbol{w} \in W$ then

$$S\boldsymbol{w} = (s_{1,1}[\boldsymbol{w}]_1 + \ldots + s_{1,K}[\boldsymbol{w}]_K, s_{2,1}[\boldsymbol{w}]_1 + \ldots + s_{2,K}[\boldsymbol{w}]_K, \ldots, s_{K,1}[\boldsymbol{w}]_1 + \ldots + s_{K,K}[\boldsymbol{w}]_K)^T.$$

Adding up the elements in vector $S\boldsymbol{w}$, we get

$$\begin{aligned}
\sum_i S[\boldsymbol{w}]_i &= s_{1,1}[\boldsymbol{w}]_1 + \ldots + s_{1,K}[\boldsymbol{w}]_K + \ldots s_{K,1}[\boldsymbol{w}]_1 + \ldots + s_{K,K}[\boldsymbol{w}]_K \\
&= (s_{1,1} + \ldots + s_{K,1})[\boldsymbol{w}]_1 + + \ldots + (s_{1,K} + \ldots + s_{K,K})[\boldsymbol{w}]_K \qquad (25) \\
&= [\boldsymbol{w}]_1 + [\boldsymbol{w}]_2 + \ldots + [\boldsymbol{w}]_K = 0,
\end{aligned}$$

and so $S\boldsymbol{w} \in W$. □

Theorem 5.3 *Fix an arbitrary $q \in \Delta$. Let $1 = \lambda_1 = \lambda_2 = \ldots = \lambda_N \geq \lambda_{N+1} \geq \lambda_{N+2} \ldots \geq \lambda_{KN}$ be a union of eigenvalues of the stochastic matrices $V^\nu(q)$ for all ν. Then the values of β for which $d^2F_H(q)$ has a nontrivial kernel (or where $\dim\ \ker d^2F_I(q) \geq d(q) + 1$, see Lemma 3.5) are*

$$\frac{1}{\lambda_{N+1}} \leq \frac{1}{\lambda_{N+2}} \leq \ldots \leq \frac{1}{\lambda_{KN}}$$

Proof. The only difference between d^2F_H and d^2F_I is the N dimensional kernel of the latter matrix. Therefore we will only consider d^2F_H in this proof.

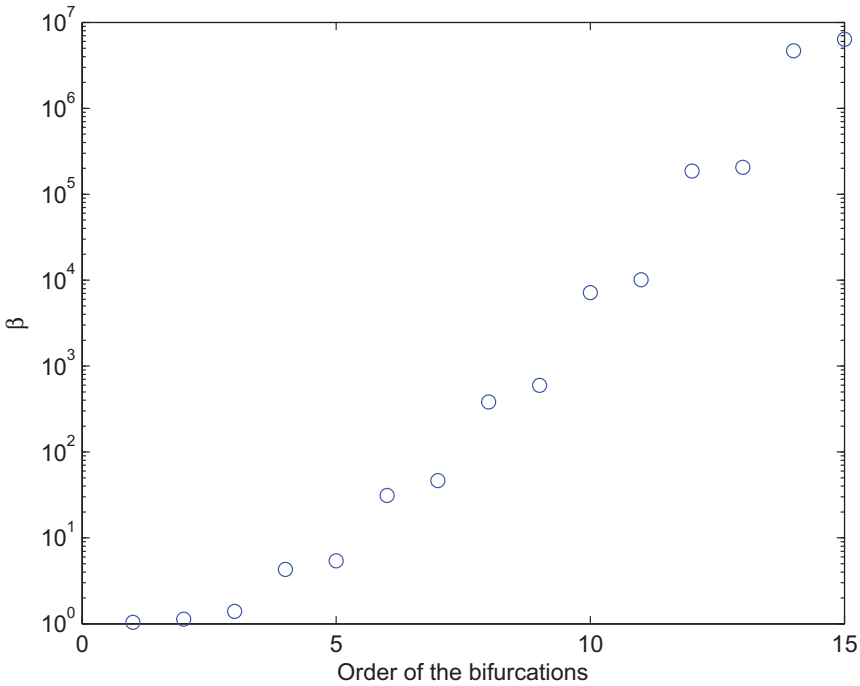
As discussed above, $d^2F_H(q)$ has a nontrivial kernel if and only if there is a block which has a nontrivial kernel. We will use the previous Lemma to discuss such a block.

Note that $\lambda = 0$ corresponds to $\beta = \infty$, and so this scenario is unimportant for the bifurcation structure of the problems (1) and (2).

Since W^ν is a $K-1$ dimension invariant subspace of $\mathbf{R}^K$, there must be $K-1$ eigenvectors of $V^\nu(q)$ in W^ν. The 0-eigenvector P_ν is not in W^ν, so all other eigenvectors not corresponding to $\lambda = 0$ must be in W^ν. Since $V^\nu(q)$ is stochastic and $\lambda = 0$ corresponds to the eigenvalue 1 of $V^\nu(q)$, then the β values at which bifurcation occurs are reciprocals to the eigenvalues of $V^\nu(q)$ for each ν. That means $\beta \leq 1$. Since there are N blocks, there will be at least N eigenvalues of $d^2F_H(q)$ equal to 1. □

We used Theorem 5.3 to determine the β values where bifurcations occur from the uniform solution branch $(q_{\frac{1}{N}}, \beta)$. The results are presented in Figure 4.

Figure 4. Theorem 5.3 can be used to determine the β values where bifurcations can occur from $(q_{\frac{1}{N}}, \beta)$. A joint probability space on the random variables (X, Y) was constructed from a mixture of four Gaussians as in [2]. For this data set, and for either $F = F_H$ or $F = F_I$, we predict bifurcation from the branch $(q_{\frac{1}{4}}, \beta)$, at each of the 15 β values given in this figure. By Theorem 4.1, $q_{\frac{1}{4}}$ ceases to be a solution at $\beta \approx 1.038706$.



6. Normalized Cuts and the Bifurcation off $q_{\frac{1}{N}}$

There is a vast literature devoted to problems of clustering. Many clustering problems can be formulated in the language of graph theory. Objects which one desires to cluster are represented as a set of nodes V of a graph $G = (V, E)$, and the weights w associated to edges represent the degree of similarity of two adjacent nodes. Finding a good clustering in such a formulation is equivalent to finding

a *cut* in the graph G, which divides the set of nodes V into sets representing individual clusters. A cut in the graph is simply a collection of edges that are removed from the graph.

A bi-partitioning of the graph is the problem in which a cut divides the graph into two parts, A and B. We define

$$cut(A, B) = \sum_{u \in A, v \in B} w(u, v) \tag{26}$$

There are efficient algorithms to solve *minimal cut* problem, where one seeks a partition into sets A and B with minimal *cut* value. When using the minimal cut as a basis for a clustering algorithm, one often finds that the minimal cut is achieved by separating one node from the rest of the graph G. Including more edges into the cut increases the cost, hence these singleton solutions will be favored.

To counteract that, Shi and Malik [21] studied image segmentation problems and proposed a clustering based on minimizing the *normalized cut (Ncut)*:

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \tag{27}$$

where

$$assoc(A, V) = \sum_{u \in A, t \in V} w(a, t)$$

Shi and Malik [21] have shown that the problem of minimizing the normalized cut is NP-complete. However, they proposed an approximate solution, which can be found efficiently. We briefly review their argument: Let

$$d(i) = \sum_j w(i, j)$$

be the total connection from node i to all other nodes. Let $n = |V|$ be the number of nodes in the graph and let D be an $n \times n$ diagonal matrix with values $d(i)$ on the diagonal. Let W be an $n \times n$ symmetric matrix with

$$W(i, j) = w_{ij}$$

Let x be an indicator vector with $x_i = 1$ if node i is in A, and $x_i = -1$ otherwise. Then Shi and Malik [21] show that the minimal cut can be computed by minimizing the Rayleigh quotient over a discrete set of admissible vectors y:

$$\min_x Ncut(x) = \min_y \frac{y^T (D - W) y}{y^T D y} \tag{28}$$

with components of y satisfying $y_i \in \{1, -b\}$ for some constant b, and under the additional constraint

$$y^T D \mathbf{1} = 0 \tag{29}$$

If one relaxes the first constraint $y_i \in \{1, -b\}$ and allows for a real valued vector y, then the problem is computationally tractable. The computation of the real valued vector y is the basis of the *Approximate normalized cut*. Once this vector is computed, vertices of G which correspond to positive entries of y will be assigned to the set A, and vertices which correspond to negative entries of y will be assigned to the set B. The relaxed problem is solved by the solution of a generalized eigenvalue problem,

$$(D - W) y = \mu D y \tag{30}$$

that satisfies the constraint (29). We repeat here an argument of Shi and Malik's [21], which shows that (28) with the constraint (29) is solved by the second smallest eigenvector of the problem (30). In fact, the smallest eigenvalue of (30) is zero and corresponds to an eigenvector $y_0 = \mathbf{1}$. The argument starts with rewriting (30) as

$$D^{-1/2}(D - W)D^{-1/2}z = \mu z$$

and realizing that $z_0 = D^{1/2}\mathbf{1}$ is a 0-eigenvector of this equation. Further, since $D^{-1/2}(D - W)D^{-1/2}$ is symmetric, all other eigenvectors are perpendicular to z_0. Translating back to problem (30), one gets the corresponding vector y_0 and all other eigenvectors satisfying $y^T D\mathbf{1} = 0$. We want to observe that this is the only place when the symmetry of matrix W is used.

In Theorem 4.1 we showed that the bifurcating direction v of one block of d^2F is the eigenvector corresponding to the second largest eigenvalue of a stochastic matrix Q. In Remark 4.4 we interpreted the matrix Q^T as a transition matrix of a Markov chain and we associated a directed graph G to this Markov chain. The graph G had vertices labelled by the elements of Y and the weight of the edge $y_l \to y_k$ was defined by

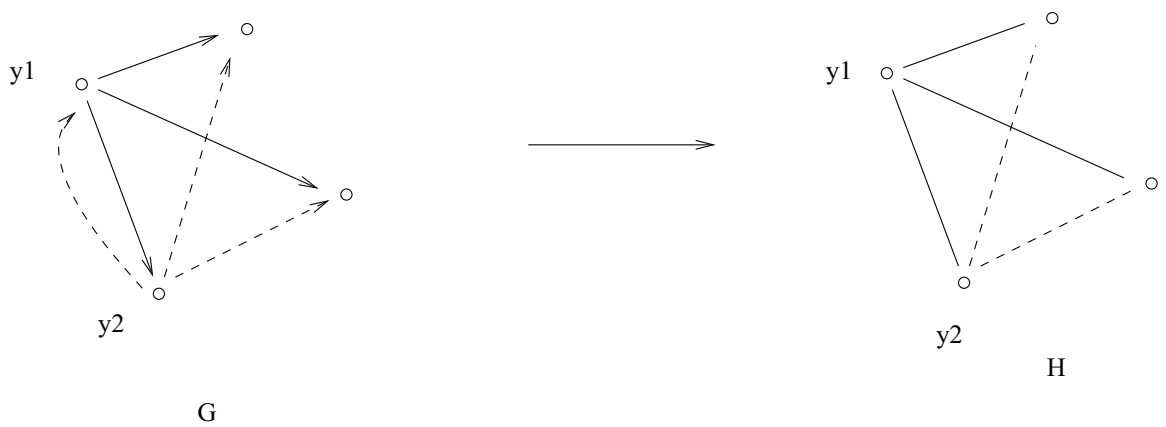
$$[Q]_{lk} = \sum_x p(y_k|x)p(x|y_l)$$

Note that these weights are not symmetric. We will symmetrize the graph G by multiplying the weight matrix $Q := Q(q_{\frac{1}{N}})$ by a diagonal matrix $C = diag(p(y_k))$. The resulting graph H (Figure 5) has a weight matrix CQ whose lk^{th} element is

$$\sum_x p(y_k|x)p(x|y_l)p(y_l) = \sum_x \frac{p(y_k, x)p(x, y_l)}{p(x)} \tag{31}$$

We form an undirected graph H with vertices labelled by elements of Y and the edge weight w_{lk} given by (31).

Figure 5. Graph G on the left is an oriented graph. We obtain unoriented graph H on the right by multiplying all edges emanating from y_i by $p(y_i)$. In the figure all weighs along solid edges are multiplied by $p(y_1)$ and all weights along the dashed edges are multiplied by $p(y_2)$.



The following Theorem, relating the bifurcating direction v_2 of matrix Q to the solution of the Approximate Normalized Cut of graph H, was proved in [22]. We use the notation of Theorem 4.1

Theorem 6.1 ([22]) *The eigenvector v_2, along which the solution $q = 1/N$ bifurcates at $\beta_2 = 1/\lambda_2$, induces the Approximate Normal Cut of the graph H.*

This Theorem shows that the bifurcating eigenvector solves the Approximate Normal Cut for the graph H, rather than the original graph G. This suggest an important inverse problem. Given a graph H for which we want to compute the Approximate Normal Cut, can we construct the graph G (given by the set of vertices, edges and weights), such that the bifurcating eigenvector would compute the Approximate Normal Cut for H? This problem, which is beyond the scope of this paper, was addressed in [22], where an annealing algorithm was designed to compute the Approximate Normal Cut using these techniques. The reader is referred to the original paper for more details.

Remark 6.2 In [15] we show that the bifurcating direction for $d^2\mathcal{L}_H$ at the first phase transition from $q_{\frac{1}{N}}$ is a vector of the form

$$\boldsymbol{u} := ((N-1)v, -v, \ldots, -v)^T$$

where $v := v_2$ is the second eigenvector of the block B_1 (all the block are identical by symmetry). In this expression $v \in \mathbf{R}^N$ and there are K vectors of size K in vector $\boldsymbol{u}$. Then the quantizer q shortly after passing a bifurcation value of β has the form

$$q = q_{\frac{1}{N}} + \epsilon \boldsymbol{u} \tag{32}$$

Let us denote by A the set of y_i such that the i-th component of v is negative, and by B the set of y_i such that the i-th component of v is positive. Note that A and B correspond to the Approximate Ncut for both graphs G and H. If we verbalize $q(t|y)$ as "the probability that y belongs to class t", then (32) shows that, after bifurcation

- the probability that $y \in A$ belongs to class 1 is less than $1/N$ and the probability that it belongs to classes $2, \ldots, N$ is more then $1/N$;
- the probability that $y \in B$ belongs to class 1 is more than $1/N$ and the probability that it belongs to classes $2, \ldots, N$ is less then $1/N$.

This describes the correspondence between the first bifurcation and Approximate Ncut.

7. Conclusions

The main goal of this contribution was to show that information-based distortion annealing problems have an interesting mathematical structure. The most interesting aspects of that mathematical structure are driven by the symmetries present in the cost functions—their invariance to actions of the permutation group S_N, represented as relabeling of the reproduction classes. The second mathematical structure that we used successfully was bifurcation theory, allowing us to identify and study the discrete points at which the character of the solutions to the cost function changed. The combination of those two tools allowed us to compute explicitly in Section 4 the value of the annealing parameter β at which the initial maximum $q_{\frac{1}{N}}$ of (1) and (2) loses stability. We concluded that, for a fixed system $p(\boldsymbol{X}, \boldsymbol{Y})$, this value is the same for both problems, that it does not depend on the number of elements of the reproduction

variable $\boldsymbol{T}$ and that it is always greater than 1. In Section 5 we further introduced an eigenvalue problem which links together the critical values of β and q for phase transition off arbitrary intermediate solutions.

Even though the cost functions F_I and F_H have similar properties, they also differ in some important aspects. We have shown that the function F_I is degenerate since its constitutive functions $\boldsymbol{I}(Y;T)$ and $\boldsymbol{I}(X;T)$ are not strictly convex. That introduces additional invariances that are always preserved, which makes phase transitions more difficult to detect, and post-transition directions more difficult to determine. Specifically, in addition to actions by the group of symmetries, the cost function F_I is invariant to altering a solution by a vector in the ever-present kernel (identified in Corollary 3.4). In contrast, F_H is strictly convex except at points of phase transitions. The theory we developed allows us to identify bifurcation directions, and determine their stability. Despite the presence of a high dimensional null space at bifurcations, the symmetries restrict the allowed transitions to multiple 1-dimensional transition, all related by group transformations.

Finally, in Section 6 we showed that the direction in which a phase transition occurs can be linked to an Approximate Normalized Cut problem of graphs arising naturally from the data structure given by $p(\boldsymbol{X},\boldsymbol{Y})$. This connection will allow future studies of information distortion methods to include powerful approximate techniques developed in Graph Theory. It will also allow the transition of the methods we developed here into tools that may be used to create new approximations for the Approximate Normalized Cut problem.

Previously we have shown that for both problems the global optimum ($\beta \to \infty$) is deterministic [3], and that the combinatorial search for the solution is NP-complete [23]. The main problem that still remains unresolved is whether the global optimum can always be achieved by the annealing process from the uniform starting solution. Proving this may be equivalent to stating that $NP = P$, so it is unlikely. However, the relatively straightforward annealing problem, when combined with the power of equivariant bifurcation theory, may be a fruitful method for approaching NP-hard problems.

Acknowledgments

This research was partially supported by NSF grants CMMI 0849433 and DMS-081878.

References and Notes

1. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. In Proceedings of the 37th annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, September 22-24, 1999.
2. Dimitrov, A.G.; Miller, J.P. Neural coding and decoding: Communication channels and quantization. *Netw. Comput. Neural Syst.* **2001**, *12*, 441–472.
3. Gedeon, T.; Parker, A.E.; Dimitrov, A.G. Information distortion and neural coding. *Can. Appl. Math. Q.* **2003**, *10*, 33–70.
4. Cover, T.; Thomas, J. *Elements of Information Theory*; Wiley Series in Communication: New York, NY, USA, 1991.

5. Slonim, N.; Tishby, N. Agglomerative information bottleneck. In *Advances in Neural Information Processing Systems*; Solla, S.A., Leen, T.K., M'uller, K.R., Eds.; MIT Press: Boston, MA, USA, 2000; Volume 12, pp. 617–623.
6. Slonim, N. The information bottleneck: Theory and applications. Ph.D. Thesis, Hebrew University, Jerusalem, Israel, November 2002.
7. Pereira, F.; Tishby, N.Z.; Lee, L. Distributional clustering of english words. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, Newark, DE, USA, 28 June–2 July 1992; pp. 183–190.
8. Bekkerman, R.; El-Yaniv, R.; Tishby, N.; Winter, Y. Distributional word clusters *vs.* words for text categorization. *J. Mach. Learn. Res.* **2003**, *3*, 33–70
9. Mumey, B.; Gedeon, T.; Taubmann, J.; Hall, K. Network dynamics discovery in genetic and neural systems. In Proceedings of the ISMB 2000, La Jolla, CA, USA, 2000.
10. Bialek, W.; de Ruyter van Steveninck, R.R.; Tishby, N. Efficient representation as a design principle for neural coding and computation. In Proceedings of the 2006 IEEE International Symposium on Information Theory, Seattle, WA, USA, 9–14 July 2006; pp. 659–663.
11. Schneidman, E.; Slonim, N.; Tishby, N.; de Ruyter van Steveninck, R.R.; Bialek, W. Analyzing neural codes using the information bottleneck method. In *Advances in Neural Information Processing Systems*; MIT Press: Boston, MA, USA 2003; Volume 15.
12. Slonim, N.; Somerville, R.; Tishby, N.; Lahav, O. Objective classification of galaxy spectra using the information bottleneck method. *Mon. Not. R. Astron. Soc.* **2001**, *323*, 270–284.
13. Gueguen, L.; Datcu, M. Image time-series data mining based on the information-bottleneck principle. *IEEE Trans. Geosci. Rem. Sens.* **2007**, *45*, 827–838.
14. Rose, K. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE* **1998**, *86*, 2210–2239.
15. Parker, A.; Dimitrov, A.G.; Gedeon, T. Symmetry breaking clusters in soft clustering decoding of neural codes. *IEEE Trans. Inform. Theor.* **2010**, *56*, 901–927.
16. Parker, A.; Gedeon, T.; Dimitrov, A. Annealing and the rate distortion problem. In *Advances in Neural Information Processing Systems 15*; Becker, S.T., Obermayer, K., Eds.; MIT Press: Cambridge, MA, USA, 2003; Volume 15, pp. 969–976.
17. Parker, A.E.; Gedeon, T. Bifurcation structure of a class of S_N-invariant constrained optimization problems. *J. Dynam. Differ. Equat.* **2004**, *16*, 629–678.
18. Nocedal, J.; Wright, S.J. *Numerical Optimization*; Springer: New York, NY, USA, 2000.
19. Parker, A.E. Symmetry Breaking Bifurcations of the Information Distortion. Ph.D. Thesis, Montana State University, Bozeman, MT, USA, April 2003.
20. Golubitsky, M.; Schaeffer, D.G. *Singularities and Groups in Bifurcation Theory I*; Springer Verlag: New York, NY, USA, 1985.
21. Shi, J.; Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905.
22. Gedeon, T.; Campion, C.; Parker, A.E.; Aldworth, Z. Annealing an information type cost function computes the normalized cut. *Pattern Recogn.* **2008**, *41*, 592–606.

23. Mumey, B.; Gedeon, T. Optimal mutual information quantization is NP-complete. In Proceedings of the Neural Information Coding (NIC) workshop, Snowbird, UT, USA, March 2003.