

---

# Annealing and the rate distortion problem

---

**Albert E. Parker**

Department of Mathematical Sciences  
Montana State University  
Bozeman, MT 59771  
parker@math.montana.edu

**Tomáš Gedeon**

Department of Mathematical Sciences  
Montana State University  
gedeon@math.montana.edu

**Alexander G. Dimitrov**

Center for Computational Biology  
Montana State University  
alex@nervana.montana.edu

**Bryan Roosien**

Department of Mathematical Sciences  
Montana State University  
roosien@math.montana.edu

## Abstract

In this paper we introduce an algorithm to efficiently optimize a class of similar cost functions which are used in Rate Distortion Theory, Deterministic Annealing, Information Distortion and the Information Bottleneck Method. Our algorithm is efficient because it explicitly takes into account the bifurcation structure of optima of the cost functions.

## 1 Introduction

This paper analyzes a class of optimization problems

$$\max_{q \in \Delta} G(q) + \beta D(q) \quad (1)$$

where  $\Delta$  is a constraint space,  $G$  and  $D$  are real valued functions of  $q$ , and  $\max_{q \in \Delta} G(q)$  is known. The goal is to solve (1) for  $\beta = \mathcal{B} \in [0, \infty)$ .

This type of problem arises in Rate Distortion Theory [1, 2], Deterministic Annealing [3], Information Distortion [4, 5, 6] and the Information Bottleneck Method [7, 8].

The following basic algorithm, various forms of which have appeared in [3, 4, 6, 7, 8], can be used to solve (1) for  $\beta = \mathcal{B}$ .

**Algorithm 1** *Let*

$$q_0 \text{ be the maximizer of } \max_{q \in \Delta} G(q) \quad (2)$$

*and let  $\beta_0 = 0$ . For  $k \geq 0$ , let  $(q_k, \beta_k)$  be a solution to (1). Iterate the following steps until  $\beta_K = \mathcal{B}$  for some  $K$ .*

1. *Perform  $\beta$ -step: Let  $\beta_{k+1} = \beta_k + d_k$  where  $d_k > 0$ .*
2. *Take  $q_{k+1}^{(0)} = q_k + \eta$ , where  $\eta$  is a small perturbation, as an initial guess for the solution  $q_{k+1}$  at  $\beta_{k+1}$ .*

3. *Optimization: solve*

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

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

We introduce methodology to efficiently perform algorithm 1. Specifically, we implement numerical continuation techniques [9, 10] to effect steps 1 and 2. We show how to detect bifurcation and we rely on bifurcation theory with symmetries [11, 12, 13] to search for the desired solution branch. This paper concludes with the improved algorithm 5 which solves (1).

## 2 The cost functions

The four problems we analyze are from Rate Distortion Theory [1, 2], Deterministic Annealing [3], Information Distortion [4, 5, 6] and the Information Bottleneck Method [7, 8]. We discuss the explicit form of the cost function (i.e.  $G(q)$  and  $D(q)$ ) for each of these scenarios in this section.

### 2.1 The distortion function $D(q)$

Rate distortion theory is the information theoretic approach to the study of optimal source coding systems, including systems for quantization and data compression [2]. To define how well a source, the random variable  $Y$ , is represented by a particular representation using  $N$  symbols, which we call  $Y_N$ , one introduces a *distortion function* between  $Y$  and  $Y_N$

$$D(q(y_N|y)) = D(Y, Y_N) = E_{y, y_N} d(y, y_N) = \sum_y \sum_{y_N} q(y_N|y) p(y) d(y, y_N)$$

where  $d(y, y_N)$  is the *pointwise distortion function* on the individual elements of  $y \in Y$  and  $y_N \in Y_N$ .  $q(y_N|y)$  is a stochastic map or *quantization* of  $Y$  into a representation  $Y_N$  [1, 2]. The constraint space

$$\Delta := \{q(y_N|y) \mid \sum_{y_N} q(y_N|y) = 1 \text{ and } q(y_N|y) \geq 0 \forall y \in Y\} \quad (3)$$

(compare with (1)) is the space of valid quantizers in  $\mathfrak{R}^n$ . A representation  $Y_N$  is optimal if there is a quantizer  $q^*(y_N|y)$  such that  $D(q^*) = \min_{q \in \Delta} D(q)$ .

In engineering and imaging applications, the distortion function is usually chosen as the *mean squared error* [1, 3, 14],

$$\hat{D}(Y, Y_N) = E_{y, y_N} \hat{d}(y, y_N) = \sum_y \sum_{y_N} q(y_N|y) p(y) \hat{d}(y, y_N)$$

where the pointwise distortion function  $\hat{d}(y, y_N)$  is the Euclidean squared distance. In this case,  $\hat{D}(Y, Y_N)$  is a linear function of the quantizer.

In neural coding, one can model a neural *coding scheme* as a stochastic map  $p(y|x)$  from the stimulus space  $X$  to the space of neural responses  $Y$ . One approach used to determine  $p(y|x)$  is to quantize the neural responses  $Y$  into a smaller event space  $Y_N$ . Since the metric between spike trains may not coincide with Euclidean distance [15], we do not want to impose  $\hat{D}(q)$  as the distortion function. The natural measure of closeness between two distributions is the Kullback-Leibler divergence  $KL$ . For each fixed  $y \in Y$  and  $y_N \in Y_N$ ,  $p(x|y)$  and  $p(x|y_N)$  are a pair of distributions on the stimulus space  $X$ . We take  $d(y, y_N) = KL(p(x|y_N) || p(x|y))$  as a pointwise distortion function. Unlike the pointwise

distortion functions usually investigated in information theory [1, 3], this one is nonlinear, it explicitly considers a third space,  $X$ , of inputs, and it depends on the quantizer  $q(y_N|y)$  through  $p(x|y_N)$ . We define our distortion function as the expected Kullback-Leibler divergence over all pairs  $(y, y_N)$

$$D_I(q(y_N|y)) = D_I(Y, Y_N) := E_{y, y_N} KL(p(x|y_N)||p(x|y)).$$

A straightforward calculation [4, 6] shows that

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

We interpret this function as an *information distortion measure*, hence the symbol  $D_I$ . The only term in  $D_I$  which depends on the quantizer is  $I(X; Y_N)$ , so we can replace  $D_I$  with the effective distortion

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

$D_{eff}(q)$  is the function  $D(q)$  from (1) which has been considered in [4, 5, 6].

The Information Bottleneck Method is another unsupervised non-parametric data clustering technique which has been applied to document classification, gene expression, neural coding and spectral analysis [7, 8]. It also uses  $D_{eff}(q)$  as the distortion function.

## 2.2 Rate Distortion

There are two related methods used to analyze communication systems at a distortion  $D(q) \leq D_0$  for some given  $D_0 \geq 0$  [1, 2, 3]. In rate distortion theory [1, 2], the problem of finding a minimum rate at a given distortion is posed as a *minimal information rate* distortion problem:

$$\min_{q(y_N|y) \in \Delta} \begin{array}{l} I(Y; Y_N) \\ D(Y, Y_N) \leq D_0 \end{array} . \quad (4)$$

This formulation is justified by the Rate Distortion Theorem [1]. A similar exposition using the Deterministic Annealing approach [3] is a *maximal entropy* problem

$$\max_{q(y_N|y) \in \Delta} \begin{array}{l} H(Y_N|Y) \\ D(Y; Y_N) \leq D_0 \end{array} . \quad (5)$$

The justification for using (5) is Jayne's maximum entropy principle [16]. These formulations are related since  $I(Y; Y_N) = H(Y_N) - H(Y_N|Y)$ .

Let  $I_0 > 0$  be some given information rate. In [4, 6], the neural coding problem is formulated as an entropy problem as in (5)

$$\max_{q(y_N|y) \in \Delta} \begin{array}{l} H(Y_N|Y) \\ D_{eff}(q) \geq I_0 \end{array} . \quad (6)$$

which uses the nonlinear effective information distortion measure  $D_{eff}$ .

Tishby et. al. [7, 8] pose an information rate distortion problem as in (4)

$$\min_{q(y_N|y) \in \Delta} \begin{array}{l} I(Y; Y_N) \\ D_{eff}(q) \geq I_0 \end{array} . \quad (7)$$

Using the method of Lagrange multipliers, the rate distortion problems (4),(5),(6),(7) can be reformulated as finding the maxima of

$$\max_{q \in \Delta} F(q, \beta) = \max_{q \in \Delta} [G(q) + \beta D(q)] \quad (8)$$

as in (1) where  $\beta = \mathcal{B}$ . For the maximal entropy problem (6),

$$F(q, \beta) = H(Y_N|Y) + \beta D_{eff}(q) \quad (9)$$

and so  $G(q)$  from (1) is the conditional entropy  $H(Y_N|Y)$ . For the minimal information rate distortion problem (7),

$$F(q, \beta) = -I(Y; Y_N) + \beta D_{eff}(q) \quad (10)$$

and so  $G(q) = -I(Y; Y_N)$ .

In [3, 4, 6], one explicitly considers  $\mathcal{B} = \infty$ . For (9), this involves taking  $\lim_{\beta \rightarrow \infty} \max_{q \in \Delta} F(q, \beta) = \max_{q \in \Delta} D_{eff}(q)$  which in turn gives  $\min_{q(y_N|y) \in \Delta} D_I$ . In Rate Distortion Theory and the Information Bottleneck Method, one is only interested in solutions to (8) for finite  $\mathcal{B}$  which takes into account a tradeoff between  $I(Y; Y_N)$  and  $D_{eff}$ .

For lack of space, here we consider (9) and (10). Our analysis extends easily to similar formulations which use the mean squared error distortion  $\hat{D}(q)$ , as in [3].

### 3 Improving the algorithm

We now turn our attention back to algorithm 1 and indicate how numerical continuation [9, 10], and bifurcation theory with symmetries [11, 12, 13] make it more efficient.

We begin by rewriting (8), now incorporating the Lagrange multipliers for the equality constraint  $\sum_{y_N} q(y_N|y_k) = 1$  from (3) which must be satisfied for each  $y_k \in Y$ . This gives the Lagrangian

$$\mathcal{L}(q, \lambda, \beta) = F(q, \beta) + \sum_{k=1}^K \lambda_k \left( \sum_{y_N} q(y_N|y_k) - 1 \right). \quad (11)$$

There are optimization schemes, such as the Fixed Point [4, 6] and projected Augmented Lagrangian [6, 17] methods, which exploit the structure of (11) to find local solutions to (8) for step 3 of algorithm 1.

#### 3.1 Bifurcation structure of solutions

It has been observed that the solutions  $\{q_k\}$  undergo *bifurcations* or *phase transitions* [3, 4, 6, 7, 8]. We wish to pose (8) as a dynamical system in order to study the *bifurcation structure* of local solutions for  $\beta \in [0, \mathcal{B}]$ . To this end, consider the equilibria of the flow

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

for  $\beta \in [0, \mathcal{B}]$ . These are points  $\begin{pmatrix} q^* \\ \lambda^* \end{pmatrix}$  where  $\nabla_{q, \lambda} \mathcal{L}(q^*, \lambda^*, \beta) = 0$  for some  $\beta$ . The Jacobian of this system is the Hessian  $\Delta_{q, \lambda} \mathcal{L}(q, \lambda, \beta)$ . Equilibria,  $(q^*, \lambda^*)$ , of (12), for which  $\Delta_q F(q^*, \beta)$  is negative definite, are local solutions of (8) [17, 18].

The  $(n + K) \times (n + K)$  Hessian of (11) is

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

where  $\mathbf{0}$  is  $K \times K$  [18].  $\Delta_q F$  is the  $n \times n$  block diagonal matrix of  $N$   $K \times K$  matrices  $\{B_i\}_{i=1}^N$  [4].  $J$  is the  $K \times n$  Jacobian of the vector of  $K$  constraints from (11),

$$J = \underbrace{\begin{pmatrix} I_K & I_K & \dots & I_K \end{pmatrix}}_{N \text{ blocks}}. \quad (13)$$

The kernel of  $\Delta_{q, \lambda} \mathcal{L}$  plays a pivotal role in determining the bifurcation structure of solutions to (8). This is due to the fact that bifurcation of an equilibria  $(q^*, \lambda^*)$  of (12) at  $\beta = \beta^*$  happen when  $\ker \Delta_{q, \lambda} \mathcal{L}(q^*, \lambda^*, \beta^*)$  is nontrivial. Furthermore, the bifurcating branches are tangent to certain linear subspaces of  $\ker \Delta_{q, \lambda} \mathcal{L}(q^*, \lambda^*, \beta^*)$  [12].

### 3.2 Bifurcations with symmetry

Any solution  $q^*(y_N|y)$  to (8) gives another equivalent solution simply by permuting the labels of the classes of  $Y_N$ . For example, if  $P_1$  and  $P_2$  are two  $n \times 1$  vectors such that for a solution  $q^*(y_N|y)$ ,  $q^*(y_N = 1|y) = P_1$  and  $q^*(y_N = 2|y) = P_2$ , then the quantizer where  $\hat{q}(y_N = 1|y) = P_2$ ,  $\hat{q}(y_N = 2|y) = P_1$  and  $\hat{q}(y_N|y) = q^*(y_N|y)$  for all other classes  $y_N$  is a maximizer of (8) with  $F(\hat{q}, \beta) = F(q^*, \beta)$ . Let  $S_N$  be the algebraic group of all permutations on  $N$  symbols [19, 20]. We say that  $F(q, \beta)$  is  $S_N$ -invariant if  $F(q, \beta) = F(\sigma(q), \beta)$  where  $\sigma(q)$  denotes the action on  $q$  by permutation of the classes of  $Y_N$  as defined by any  $\sigma \in S_N$  [18]. Now suppose that a solution  $q^*$  is fixed by all the elements of  $S_M$  for  $M \leq N$ . Bifurcations at  $\beta = \beta^*$  in this scenario are called *symmetry breaking* if the bifurcating solutions are fixed (and only fixed) by subgroups of  $S_M$ .

To determine where a bifurcation of a solution  $(q^*, \lambda^*, \beta)$  occurs, one determines  $\beta$  for which  $\Delta_q F(q^*, \beta)$  has a nontrivial kernel. This approach is justified by the fact that  $\Delta_{q, \lambda} \mathcal{L}(q^*, \lambda^*, \beta)$  is singular if and only if  $\Delta_q F(q^*, \beta)$  is singular [18]. At a bifurcation  $(q^*, \lambda^*, \beta^*)$  where  $q^*$  is fixed by  $S_M$  for  $M \leq N$ ,  $\Delta_q F(q^*, \beta^*)$  has  $M$  identical blocks. The bifurcation is generic if

$$\begin{aligned} &\text{each of the identical blocks has a single 0-eigenvector, } \mathbf{v}, \\ &\text{and the other blocks are nonsingular.} \end{aligned} \quad (14)$$

Thus, a generic bifurcation can be detected by looking for singularity of one of the  $K \times K$  identical blocks of  $\Delta_q F(q^*, \beta)$ . We call the classes of  $Y_N$  which correspond to identical blocks *unresolved* classes. The classes of  $Y_N$  that are not unresolved are called *resolved* classes.

The Equivariant Branching Lemma and the Smoller-Wasserman Theorem [12, 13] ascertain the existence of explicit bifurcating solutions in subspaces of  $\ker \Delta_{q, \lambda} \mathcal{L}(q^*, \lambda^*, \beta^*)$  which are fixed by special subgroups of  $S_M$  [12, 13]. Of particular interest are the bifurcating solutions in subspaces of  $\ker \Delta_{q, \lambda} \mathcal{L}(q^*, \lambda^*, \beta^*)$  of dimension 1 guaranteed by the following theorem

**Theorem 2** [18] *Let  $(q^*, \lambda^*, \beta^*)$  be a generic bifurcation of (12) which is fixed (and only fixed) by  $S_M$ , for  $1 < M \leq N$ . Then there exists  $M$  bifurcating solutions,*

$$\begin{aligned} &\begin{pmatrix} q^* \\ \lambda^* \\ \beta^* \end{pmatrix} + \begin{pmatrix} t\mathbf{u}_m \\ \beta(t) \end{pmatrix}, \text{ where } 1 \leq m \leq M, \\ &[\mathbf{u}_m]_\nu = \begin{cases} (M-1)\mathbf{v} & \text{if } \nu \text{ is the } m^{\text{th}} \text{ unresolved class of } Y_N \\ -\mathbf{v} & \text{if } \nu \text{ is some other unresolved class of } Y_N \\ \mathbf{0} & \text{otherwise} \end{cases} \end{aligned} \quad (15)$$

and  $\mathbf{v}$  is defined as in (14). Furthermore, each of these solutions is fixed by the symmetry group  $S_{M-1}$ .

For a bifurcation from the uniform quantizer,  $q_{\frac{1}{N}}$ , which is identically  $\frac{1}{N}$  for all  $y$  and all  $y_N$ , all of the classes of  $Y_N$  are unresolved. In this case,

$$\mathbf{u}_m = (-\mathbf{v}^T, \dots, -\mathbf{v}^T, (N-1)\mathbf{v}^T, -\mathbf{v}^T, \dots, -\mathbf{v}^T, \mathbf{0}^T)^T$$

where  $(N-1)\mathbf{v}$  is in the  $m^{\text{th}}$  component.

Relevant to the computationalist is that instead of looking for a bifurcation by looking for singularity of the  $n \times n$  Hessian  $\Delta_q F(q^*, \beta)$ , one may look for singularity of one of the  $K \times K$  identical blocks, where  $K = \frac{n}{N}$ . After bifurcation of a local solution to (8) has been detected at  $\beta = \beta^*$ , knowledge of the bifurcating directions makes finding solutions of interest for  $\beta > \beta^*$  much easier (see section 3.4.1).

### 3.3 The subcritical bifurcation

In all problems under consideration the solution for  $\beta = 0$  is known. For (9), (10) this solution is  $q_0 = q_{\frac{1}{N}}$ . For (4) and (5),  $q_0$  is the mean of  $Y$ . Rose [3] was able to compute explicitly the critical value  $\beta^*$  where  $q_0$  loses stability for the Euclidean pointwise distortion function. We have the following related result.

**Theorem 3** [21] *Consider problems (9), (10). The solution  $q_0 = 1/N$  loses stability at  $\beta = \beta^*$  where  $1/\beta^*$  is the second largest eigenvalue of a discrete Markov chain on vertices  $y \in Y$ , where the transition probabilities  $p(y_l \rightarrow y_k) := \sum_i p(y_k|x_i)p(x_i|y_l)$ .*

**Corollary 4** *Bifurcation of the solution  $(q_{\frac{1}{N}}, \beta)$  in (9), (10) occurs at  $\beta \geq 1$ .*

This result together with numerical evidence presented in Figure 1 strongly suggests that the bifurcation from  $(q_{\frac{1}{N}}, \beta)$  is subcritical (i.e. a first order phase transition) for (9) and (10).

This should be contrasted with the results for (4) and (5). Here the convexity and monotonicity of the rate distortion curve [2] implies continuity of this curve for  $0 \leq D(Y, Y_N) < \max D(Y, Y_N)$  which implies that all bifurcations must be supercritical (i.e. second order phase transitions). The argument showing convexity of the rate distortion curve relies on the fact that the distortion  $\hat{D}(q)$  is linear in  $q$  and does not have immediate generalization for the analogous curves in problems (9) and (10). The analogy of problem (10) to problem (4) leads [8] to suggest that all bifurcations for (10) are supercritical (i.e. second order phase transitions), which our evidence does not confirm.

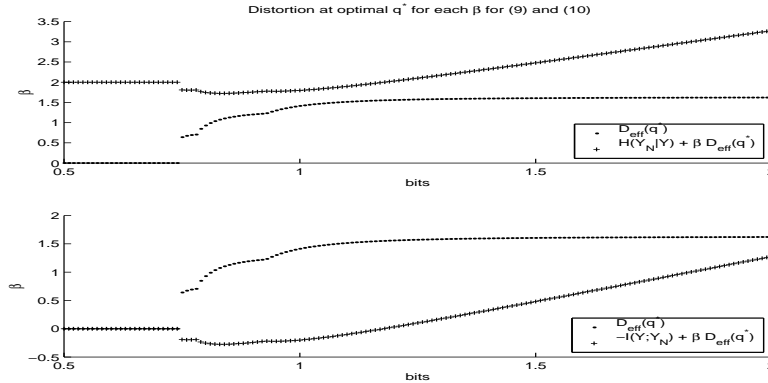


Figure 1: A joint probability space on the random variables  $(X, Y)$  was constructed from a mixture of four Gaussians as in [4]. Using this probability space, the functions (9) and (10) were maximized to determine a quantization  $q^*(y_N|y)$  of  $Y$  into a representation  $Y_N$  of 4 elements. *Top panel:* The distortion  $D_{\text{eff}}(q^*(\beta))$  and  $F(q^*, \beta)$  as defined in (9). *Bottom panel:* The distortion  $D_{\text{eff}}(q^*(\beta))$  and  $F(q^*, \beta)$  as defined in (10). At  $\beta \approx .7191$ ,  $q^* \neq q_{\frac{1}{N}}$  since  $D_{\text{eff}}(q^*) \neq 0$ . By Corollary 4, a bifurcation of the solution  $q_{\frac{1}{N}}$  does not occur until  $\beta \geq 1$ . This evidence supports the conjecture that the bifurcation from  $q_{\frac{1}{N}}$  is subcritical.

### 3.4 Numerical Continuation

Numerical *continuation* methods efficiently analyze the solution behavior of dynamical systems such as (12) [9, 10]. Continuation methods can speed up the search for the solution  $q_{k+1}$  at  $\beta_{k+1}$  in step 3 of algorithm 1 by improving upon the arbitrary choice  $q_{k+1}^{(0)} = q_k + \eta$ . First,

the vector  $(\partial_\beta q_k^T \ \partial_\beta \lambda_k^T)^T$  which is tangent to the curve  $\nabla_{q,\lambda} \mathcal{L}(q, \lambda, \beta) = \mathbf{0}$  at  $(q_k, \lambda_k, \beta_k)$  is computed by solving the matrix system

$$\Delta_{q,\lambda} \mathcal{L}(q_k, \lambda_k, \beta_k) \begin{pmatrix} \partial_\beta q_k \\ \partial_\beta \lambda_k \end{pmatrix} = \partial_\beta \nabla_{q,\lambda} \mathcal{L}(q_k, \lambda_k, \beta_k). \quad (16)$$

Now the initial guess in step 2 becomes  $q_{k+1}^{(0)} = q_k + d_k \cdot \partial_\beta q_k$  where  $d_k = \frac{\Delta s}{\|\partial_\beta q_k\| + \|\partial_\beta \lambda_k\| + 1}$  for  $\Delta s > 0$ . Furthermore,  $\beta_{k+1}$  in step 1 is found by using this same  $d_k$ . This choice of  $d_k$  assures that a fixed step along  $(\partial_\beta q_k^T \ \partial_\beta \lambda_k^T)^T$  is taken for each  $k$ . We use three different continuation methods which implement variations of this scheme: *Parameter*, *Tangent* and *Pseudo Arc-Length* [9, 18]. These methods can greatly decrease the optimization iterations needed to find  $q_{k+1}$  from  $q_{k+1}^{(0)}$  in step 3. The cost savings can be significant, especially when continuation is used in conjunction with a Newton type optimization scheme which explicitly uses the Hessian  $\Delta_q F(q_k, \beta_k)$ . Otherwise, the CPU time incurred from solving (16) may outweigh this benefit.

### 3.4.1 Branch switching

Suppose that a bifurcation of a solution  $q^*$  of (8) has been detected at  $\beta^*$ . To proceed, one uses the explicit form of the bifurcating directions,  $\{\mathbf{u}_m\}_{m=1}^M$  from (15) to search for the bifurcating solution of interest, say  $q_{k+1}$ , whose existence is guaranteed by Theorem 2. To do this, let  $\mathbf{u} = \mathbf{u}_m$  for some  $m \leq M$ , then implement a *branch switch* [9]

$$q_{k+1}^{(0)} = q^* + d_k \cdot \mathbf{u}.$$

## 4 The improved algorithm

We conclude with an efficient algorithm to solve (1). The section numbers in parentheses indicate the location in the text supporting each step.

**Algorithm 5** Let  $q_0$  be the maximizer of  $\max_{q \in \Delta} G$ ,  $\beta_0 = 1$  (3.3) and  $\Delta s > 0$ . For  $k \geq 0$ , let  $(q_k, \beta_k)$  be a solution to (1). Iterate the following steps until  $\beta_K = \mathcal{B}$  for some  $K$ .

1. (3.4) Perform  $\beta$ -step: solve (16) for  $(\partial_\beta q_k^T \ \partial_\beta \lambda_k^T)^T$  and select  $\beta_{k+1} = \beta_k + d_k$  where  $d_k = \frac{\Delta s}{\|\partial_\beta q_k\| + \|\partial_\beta \lambda_k\| + 1}$ .
2. (3.4) The initial guess for  $q_{k+1}$  at  $\beta_{k+1}$  is  $q_{k+1}^{(0)} = q_k + d_k \cdot \partial_\beta q_k$ .
3. Optimization: solve

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

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

4. (3.2) Check for bifurcation: compare the sign of the determinant of an identical block of each of

$$\Delta_q [G(q_k) + \beta_k D(q_k)] \text{ and } \Delta_q [G(q_{k+1}) + \beta_{k+1} D(q_{k+1})].$$

If a bifurcation is detected, then set  $q_{k+1}^{(0)} = q_k + d_k \cdot \mathbf{u}$  where  $\mathbf{u}$  is defined as in (15) for some  $m \leq M$ , and repeat step 3.

### Acknowledgments

Many thanks to Dr. John P. Miller at the Center for Computational Biology at Montana State University-Bozeman. This research is partially supported by NSF grants DGE 9972824, MRI 9871191, and EIA-0129895; and NIH Grant R01 MH57179.

## References

- [1] Thomas Cover and Jay Thomas. *Elements of Information Theory*. Wiley Series in Communication, New York, 1991.
- [2] Robert M. Gray. *Entropy and Information Theory*. Springer-Verlag, 1990.
- [3] Kenneth Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE*, 86(11):2210–2239, 1998.
- [4] Alexander G. Dimitrov and John P. Miller. Neural coding and decoding: communication channels and quantization. *Network: Computation in Neural Systems*, 12(4):441–472, 2001.
- [5] Alexander G. Dimitrov and John P. Miller. Analyzing sensory systems with the information distortion function. In Russ B Altman, editor, *Pacific Symposium on Biocomputing 2001*. World Scientific Publishing Co., 2000.
- [6] Tomas Gedeon, Albert E. Parker, and Alexander G. Dimitrov. Information distortion and neural coding. *Canadian Applied Mathematics Quarterly*, 2002.
- [7] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. The 37th annual Allerton Conference on Communication, Control, and Computing, 2000.
- [8] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 617–623. MIT Press, 2000.
- [9] Wolf-Jurgen Beyn, Alan Champneys, Eusebius Doedel, Willy Govaerts, Yuri A. Kuznetsov, and Bjorn Sandstede. *Handbook of Dynamical Systems III*. World Scientific, 1999. Chapter in book: Numerical Continuation and Computation of Normal Forms.
- [10] Eusebius Doedel, Herbert B. Keller, and Jean P. Kernevez. Numerical analysis and control of bifurcation problems in finite dimensions. *International Journal of Bifurcation and Chaos*, 1:493–520, 1991.
- [11] M. Golubitsky and D. G. Schaeffer. *Singularities and Groups in Bifurcation Theory I*. Springer Verlag, New York, 1985.
- [12] M. Golubitsky, I. Stewart, and D. G. Schaeffer. *Singularities and Groups in Bifurcation Theory II*. Springer Verlag, New York, 1988.
- [13] J. Smoller and A. G. Wasserman. Bifurcation and symmetry breaking. *Inventiones mathematicae*, 100:63–95, 1990.
- [14] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.
- [15] Johnathan D. Victor and Keith Purpura. Metric-space analysis of spike trains: theory, algorithms, and application. *Network: Computation in Neural Systems*, 8:127–164, 1997.
- [16] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proc. IEEE*, 70:939–952, 1982.
- [17] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2000.
- [18] Tomas Gedeon, Albert E. Parker, and Alexander G. Dimitrov. Continuation and symmetry breaking bifurcation of the distortion function. In preparation, 2002.
- [19] H. Boerner. *Representations of Groups*. Elsevier, New York, 1970.
- [20] D. S. Dummit and R. M. Foote. *Abstract Algebra*. Prentice Hall, NJ, 1991.
- [21] Tomas Gedeon and Bryan Roosien. Annealing for information distortion and markov chains. In preparation, 2002.