The Lack of Convexity of the Relevance-Compression Function

Albert E. Parker
Center for Biofilm Engineering
Department of Mathematical Sciences
Montana State University
Bozeman, MT USA
parker@math.montana.edu

Tomáš Gedeon
Department of Mathematical Sciences
Montana State University
Bozeman, MT USA
gedeon@math.montana.edu

Alexander G. Dimitrov
Department of Mathematics
Department of Neuroscience
Washington State University
Vancouver, WA
alex.dimitrov@wsu.edu

Abstract

In this paper we investigate the convexity of the relevance-compression function for the Information Bottleneck and the Information Distortion problems. This curve is an analog of the rate-distortion curve, which is convex. In the problems we discuss in this paper, the distortion function is not a linear function of the quantizer, and the relevance-compression function is not necessarily convex (concave), but can change its convexity. We relate this phenomena with existence of first order phase transitions in the corresponding Lagrangian as a function of the annealing parameter.

1 Introduction

In previous work \cite{1,2,3} we have described the bifurcation structure for solutions to problems of the form

\[ \max_{q \in \Delta} G(q) \]
\[ D(q) \geq D_0 \]

where \( \Delta \) is a constraint space of valid conditional probabilities, \( G \) and \( D \) are continuous, real valued functions of \( q \), smooth in the interior of \( \Delta \), and the functions \( G(q) \) and \( D(q) \) are invariant under the group of symmetries \( S_N \). This type of problem, which arises in Rate Distortion Theory \cite{4,5} and Deterministic Annealing \cite{6}, is \( NP \) complete \cite{7} when \( D(q) \) is the mutual information \( I(X;Y_N) \) as in the Information Bottleneck \cite{8,9,10} and the Information Distortion \cite{11,12,13} methods. In this paper, we address the relationship between the bifurcation structure of solutions to \cite{1} and the relevance compression function \cite{10}. 
2 Preliminaries

2.1 Rate Distortion Theory

We assume that the random variable $X \in \{x_1, x_2, ..., x_K\}$ is an input source, and that $Y \in \{y_1, y_2, ..., y_K\}$ is an output source. In rate distortion theory \cite{4}, the random variable $Y$ is represented by using $N$ symbols or classes, which we call $Y_N$, where we assume without loss of generality that $Y_N \in \{1, 2, ..., N\}$. We denote a stochastic clustering or quantization of the realizations of $Y$ to the classes of $Y_N$ by $q(Y_N | Y)$. To find the quantization that yields the minimum information rate $I(Y; Y_N)$ at a given distortion, one can find points on the rate distortion curve for each value of $D_0 \in [0, D_{\text{max}}]$. The rate distortion curve is defined as \cite{4, 5}

$$R(D_0) := \min_{q \in \Delta} I(Y; Y_N)$$

subject to

$$D(Y, Y_N) \leq D_0,$$

where $D(q)$ is a distortion function. A quantization $q(Y_N | Y)$ that satisfies (2) yields an approximation of the probabilistic relationship, $p(X, Y)$, between $X$ and $Y$ \cite{11, 8, 9}. The constraint space $\Delta$ is the space of valid finite conditional probabilities $q(Y_N | Y)$, where we will write $q(Y_N = \nu | Y = y_k) = q_{\nu k}$.

The Information Bottleneck method \cite{8, 9, 10} uses the information distortion function

$$D(q) := I(X; Y) - I(X; Y_N).$$

Since the spaces $X$ and $Y$ are fixed, then $I(X; Y)$ is fixed, and so the rate distortion problem (2) in the case of the Information Bottleneck problem can be rewritten as

$$R_I(I_0) := \max_{q \in \Delta} -I(Y; Y_N)$$

subject to

$$I(X; Y_N) \geq I_0,$$

where $I_0 > 0$ is some information rate. The function $R_I(I_0)$ is referred to as the relevance-compression function in \cite{10}. Observe that there is a one to one correspondence between $I_0$ and $D_0$ via $I_0 = I(X; Y) - D_0$. To solve the neural coding problem, the Information Distortion method \cite{11, 12, 13} considers a problem of the form

$$R_H(I_0) := \max_{q \in \Delta} H(Y_N | Y)$$

subject to

$$I(X; Y_N) \geq I_0,$$

where $H(Y_N | Y)$ is the conditional entropy.

2.2 Annealing

Using the method of Lagrange multipliers, an arbitrary problem of the form (1) is rewritten as

$$\max_{q \in \Delta} (G(q)) + \beta(D(q) - D_0).$$

As we will see, solutions of (1) are not always solutions of (5). Similarly, the problem (3) can be rewritten \cite{8, 9, 10} as

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

and problem (4) can be rewritten \cite{11, 12, 13}, in analogy with Deterministic Annealing \cite{6}, as

$$\max_{q \in \Delta} (H(Y_N | Y) + \beta I(X; Y_N)).$$

In (5), (6) and (7), the Lagrange multiplier $\beta$ can be viewed as an annealing parameter.
2.3 Bifurcation Structure of solutions

In [14], we presented an algorithm which can be used to determine the bifurcation structure of stationary points of (5) for each value of $\beta \in [0, \beta_{\text{max}})$ for some $\beta_{\text{max}} > 0$. These stationary points are quantizers $q^*(\beta) \in \mathbb{R}^{NK}$ where there exists a vector of Lagrange multipliers $\lambda^* \in \mathbb{R}^K$ such that the gradient of the Lagrangian of (1) is a vector of 0’s,

$$\nabla q,\lambda(G(q^*) + \beta D(q^*) + \sum_{k=1}^{K} \lambda_k^* \left( \sum_{\nu} q^*(\nu|y_k) - 1 \right) = 0.$$  

This condition is also known as the Karush-Kuhn-Tucker necessary condition for constrained optimality [15]. It is well known in optimization theory that a stationary point, i.e. the point satisfying (8), is a solution of (1) if the matrix of second derivatives, the Hessian $\Delta_q(G(q^*) + \beta D(q^*))$, is negative definite on the kernel of the Jacobian of the constraints [15]. We have the following results.

**Theorem 1** [1] A stationary point $q^*$, is a solution of (5) if $\Delta_q(G(q^*) + \beta D(q^*))$ is negative definite on $\ker \left( I_K \quad I_K \quad \ldots \quad I_K \right)$. A stationary point $q^*$ is a solution of (1) if $\Delta_q(G(q^*) + \beta D(q^*))$ is negative definite on $\ker \left( \nabla_q D(q^*), I_K \quad I_K \quad \ldots \quad I_K \right)$.

From Theorem 1 we see that there may be solutions of (1) which are not solutions of (5). We illustrate this fact numerically. For the Information Distortion problem [4, 11, 12, 13], and the synthetic data set composed of a mixture of four Gaussians which the authors used in [11], we determined the bifurcation structure of solutions to (4) by annealing in $\beta$ and finding the corresponding stationary points to (7) (see Figure 1).

Similar to the results which we presented in [14], the close up of the bifurcation at $\beta \approx 1.038706$ in Figure 1(B) shows a subcritical bifurcating branch (a first order phase transition) which consists of stationary points of the problem (7). By projecting the Hessian $\Delta_q(G(q^*) + \beta D(q^*))$ onto each of the kernels referenced in Theorem 1, we determined...
Figure 2: (A) The distortion curve $R_H(I_0)$ defined in (9). For each value of $I_0$, we solved the problem (4) to ascertain $H(Y_N|Y)$. Observe that the quantizers which yield $I(X;Y_N) \geq I_0$ for the values $I_0 \in [0, .03]$ correspond to solutions on the subcritical bifurcating branch shown in Figure fig:bifstructure(B). (B) For each value of $I_0$, we solved (4), and found the corresponding Lagrange multiplier. This plot shows $I_0$ as function of this Lagrange multiplier. This plot is identical to the subcritical bifurcation shown in Figure 1(B), which shows $I(X;Y_N)$ as a function of the annealing parameter $\beta$.

that the points on this subcritical branch are not solutions of (7), and yet they are solutions of (1).

Furthermore, observe that Figure 1(B) indicates that a saddle-node bifurcation occurs at $\beta \approx 1.037479$. That this is indeed the case was proved in [1]. In fact, for any problem of the form (5), there are only two types of bifurcations to be expected.

**Theorem 2** [1] Generically, for problems of the form (1), only symmetry breaking pitchfork-like and saddle-node bifurcations occur.

Clearly, the existence of saddle-node bifurcation at $\beta \approx 1.037479$ is tied to the existence of subcritical bifurcation (first order phase transition) at $\beta \approx 1.038706$. We now investigate the connection between existence of subcritical bifurcations and the convexity of the relevance-compression function.

### 3 The Relevance-Compression Function

Given the generic existence of subcritical pitchfork-like and saddle-node bifurcations of solutions to problems of the form (1), a natural question arises: What are the implications for the rate distortion curve (2)? We examine this question for the information distortion $D(q) = I(X;Y) - I(X;Y_N)$, used by the Information Bottleneck and the Information Distortion methods. Recall that the relevance-distortion function is

$$R_I(I_0) := \max_{\Delta \cap Q_{I_0}} -I(Y_N, Y).$$

where

$$Q_{I_0} := \{q \in \Delta \mid I(X, Y_N) \geq I_0\}.$$  

For the Information Distortion problem the relevance-distortion function is

$$R_H(I_0) := \max_{\Delta \cap Q_{I_0}} H(Y_N|Y).$$
In Figure 2(A), we present a plot of $R_H(I_0)$, which was computed using the same data set of a mixture of four Gaussians which we used in Figures 1(A) and (B). The plot was obtained by solving the problem (4) for each value of $I_0$.

To make explicit the relationship between the bifurcation structure shown in Figure 1 which was obtained by annealing in $\beta$, and the distortion curve shown in Figure 2(A), which was obtained by annealing in $I_0$, we present Figure 2(B). When solving (4) for each $I_0$, we computed the corresponding Lagrange multiplier $\beta$. Thus, $\beta = \beta(I_0)$, which is the curve we show in Figure 2(B). This plot matches precisely the subcritical bifurcating branch from Figure 1(B), which we obtained by annealing in $\beta$.

Lemma 3 For a fixed value of $I_0 > 0$ the solution $q^*$ of (7) and (8) satisfies the equality condition $I = I_0$.

Proof. Assume $q^*$ is a maximizer of (4) and $I(q^*) > I_0$. Then the constraint is not active and we must have $\nabla H(q^*) = 0$. Since $\nabla H(q^*) = 0$ implies $q^* = 1/N$, we get $I(q^*) = 0$. This is a contradiction and thus $I(q^*) = I_0$.

Now assume $q^*$ is a maximizer of (5) and $I(q^*) > I_0$. Then again the constraint is not active and we must have $\nabla I(Y, Y_N) = 0$. Short computation shows that the condition $\nabla I(Y, Y_N) = 0$ implies $q^* = (q_{0k})$ is of the form $q_{0k} = q_0$, i.e., $q_{0k}$ does not depend on $k$. However, at such value of $q^*$ we get $I(q^*) = 0$. This is again a contradiction and thus $I(q^*) = I_0$. \hfill \square

As a consequence of the Lemma, for each $I_0 > 0$ there exists a Lagrange multiplier $\beta(I_0)$. The existence of subcritical bifurcation branch implies that along this branch $\beta(I_0)$ is not a one-to-one function of $I_0$, and therefore not invertible.

3.1 Properties of relevance-compression function

It is well known that if the distortion function $D(q)$ is linear in $q$, that $R(D_0)$ is a non-increasing and convex [4,5]. The proof of this result first establishes that the rate-distortion curve is monotone and that it is convex. These two properties together imply continuity and strict monotonicity of the rate distortion curve. Since the information distortion $D = I(X; Y) - I(X; Y_N)$ is not a linear function of $q$, the convexity proof given in [4,5] does not generalize to prove that either (8) or (9) is convex. Therefore we need to prove continuity of the relevance-compression function using other means.

Lemma 4 The curves $R_H$ and $R_I$ are non-increasing curves on $I_0 \in [0, I_{\text{max}}]$ and are continuous for $I_0 \in (0, I_{\text{max}})$.

Proof. Observe that since $I(X, Y_N)$ is convex [11] in quantizer $q$, we have that $Q_{I_1} \subset Q_{I_2}$ whenever $I_1 \geq I_2$.

Therefore, if $I_1 \geq I_2$, then the maximization at $I_1$ happens over a smaller set than in $Q_{I_2}$, and so $R(I_1) \leq R(I_2)$.

Now we prove continuity. Take an arbitrary $I_0 \in (0, I_{\text{max}})$. Let $M_{I_0} := \{ y \mid y = H(q) \text{ where } q \in \Delta \cap Q_{I_0} \}$ be the range (in $\mathbb{R}$) of the function $H(q)$ with the domain $\Delta \cap Q_{I_0}$. Given an arbitrary $\epsilon > 0$, let $M_{I_0}^\epsilon$ be an $\epsilon$ neighborhood of $M_{I_0}$ in $\mathbb{R}$. A direct computation shows that $\nabla q H(q) = 0$ if and only if $q$ is homogeneous, i.e., $q_{nuk} = 1/N$, where $N$ is the number of classes of $Y_N$. Since $H(q)$ is continuous on $\Delta$, then the set $H^{-1}(M_{I_0}^\epsilon)$ is a relatively open set in $\Delta$. Because by definition $H(\Delta \cap Q_{I_0}) = M_{I_0}$, we see that $Q_{I_0} \cap \Delta \subset H^{-1}(M_{I_0}^\epsilon)$. (10)
Furthermore, since $\nabla H(q) \neq 0$ for $q \in Q_{I_0}$, then, by the Inverse Mapping Theorem, $H^{-1}(M_{I_0}^t)$ is an open neighborhood of $Q_{I_0}$.

The function $I(X; Y_N)$ is also continuous in the interior of $\Delta$. Observe that

$$Q_{I_0} = I^{-1}([I_0, I_{\text{max}}])$$

is closed, and thus $Q_{I_0} \cap \Delta$ is closed and hence compact. Thus, by (10) $H^{-1}(M_{I_0}^t)$ is an relatively open neighborhood of a compact set $Q_{I_0} \cap \Delta$. Therefore, since $I(X; Y_N)$ is continuous, there exists a $\delta > 0$ such that the set

$$\text{Int} Q_{I_0 + \delta} \cap \Delta = I^{-1}((I_0 + \delta, D_{\text{max}}]) \cap \Delta$$

is a relatively open set in $\Delta$ such that

$$Q_{I_0} \cap \Delta \subset \text{Int} Q_{I_0 + \delta} \subset H^{-1}(M_{I_0}^t).$$

It then follows that

$$\max_{\Delta \cap Q_{I_0 + \delta}} H - \max_{\Delta \cap Q_{I_0}} H | < \epsilon.$$ 

By definition of the rate distortion function, this means that

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

Since $\epsilon$ was arbitrary, this implies continuity of $R(I)$ at $I = I_0$. 

3.2 The Derivative $\frac{\partial R}{\partial I}$

In [8, 10], using variational notation, it is shown that

$$\delta R_{I_0} \delta D = -\beta.$$ 

For the sake of completeness, we will reprove this, acknowledging explicitly the fact that the problems (3) and (4) are constrained problems.

**Theorem 5** If relevance-compression functions $R_I(I_0)$ and $R_H(I_0)$ are differentiable, then

$$\frac{dR}{dI_0} = -\beta(I_0) \text{ and } \frac{d^2R}{dI_0^2} = -\frac{d\beta(I_0)}{dI_0} \quad (11)$$

**Corollary 6** Since $\frac{d\beta(I_0)}{dI_0}$ changes sign at saddle-node bifurcation, then the relevance-compression functions $R_I(I_0)$ and $R_H(I_0)$ are neither concave, nor convex.

**Proof of Theorem**: We start with

$$\max_{q \in \Delta} R(q) + \beta D(q) + \sum_k \lambda_k \left(\sum_{q} q_{qk} - 1\right) \quad (12)$$

where $R(q)$ is one of $R_I(q) := H(Y_N|Y)$, $R_H(q) := -I(Y_N, Y)$. We parameterize the solution $q^*$ locally by $\beta$. This can be done everywhere except if $q^*$ is at a saddle-node bifurcation. At $q^*(\beta)$,

$$\nabla_q R + \beta \nabla_q D + \ddot{\lambda} = 0. \quad (13)$$

**Lemma 7** For $q \in \Delta$,

$$q \cdot \nabla_q R_H = R_H + 1, \quad q \cdot \nabla_q R_I = R_I, \text{ and } q \cdot \nabla_q I = I. \quad (14)$$
Proof. Direct calculation.\qed

Hence (13) implies
\[ R(q^*(\beta)) + c + \beta I(q^*(\beta)) + q \cdot \bar{x} = 0 \] (15)
Here \( c = 1 \) for \( R_H \) and \( c = 0 \) for \( R_I \) is a constant. For \( q \in \Delta \), we set
\[ \Lambda(\beta) := q \cdot \bar{x} = \sum_k \lambda_k. \]

The equation (15) defines a relation between \( R \) and \( I \). Recall, that we can always express \( \beta = \beta(I_0) \). Then the term \( I(q^*(\beta(I_0))) = I_0 \) and we have a relationship
\[ R(I_0) + c + \beta(I_0) I_0 + \Lambda(I_0) = 0. \] (16)

We differentiate (16):
\[ \frac{dR}{dI_0} + \frac{d\beta}{dI_0} I_0 + \frac{d\Lambda}{dI_0} = 0 \Rightarrow \] (17)
which shows that \( \frac{dR}{dI_0} = -\frac{d\beta}{dI_0} - \frac{d\Lambda}{dI_0} \) since \( \frac{d\Lambda}{dI_0} = \frac{d\Lambda}{d\beta} \).

In (11) (equation (10)) we have an explicit expression for \( \lambda_k \) as a function of \( \beta \):
\[ \lambda_k = p_k(1 - \ln \sum_{\nu} e^{\beta(\nabla I)_{\nu k}/p_k}). \] (18)

Differentiating this with respect to \( \beta \) yields
\[ \frac{d\lambda_k}{d\beta} = -p_k \sum_{\nu} e^{\beta(\nabla I)_{\nu k}/p_k} \frac{\nabla I_{\nu k}/p_k}{\sum_{\mu} e^{\beta(\nabla I)_{\mu k}/p_k}}. \]
Since \( \Lambda = \sum_k \lambda_k \), this implies that
\[ \frac{d\Lambda}{d\beta} = \sum_k \frac{d\lambda_k}{d\beta} = -\sum_{\nu k} \sum_{\mu} e^{\beta(\nabla I)_{\mu k}/p_k} (\nabla I)_{\nu k}. \]

For a solution \( q^* \), \[ \sum_{\nu k} e^{\beta(\nabla I)_{\nu k}/p_k} = q^*_{\nu k} \] (11), (12), hence
\[ \frac{d\Lambda}{d\beta} |_{q^*} = -\sum_{\nu k} q^*_{\nu k} (\nabla I)_{\nu k} = -q^* \cdot \nabla I = -I \]

This shows that the term \( I_0 + \frac{d\Lambda}{d\beta} = 0 \) at \( q^* \), hence from (17) we get the first part of (11). The second part follows immediately.\qed

Acknowledgments

This research is partially supported by NSF grants DGE 9972824, MRI 9871191, and EIA-0129895; and NIH Grant R01 MH57179.

References

[1] Albert E. Parker and Tomas Gedeon. Bifurcations of a class of \( S_n \)-invariant constrained optimization problems. Journal of Dynamics and Differential Equations, 16(3):629–678, July 2004. Second special issue dedicated to Shui-Nee Chow.

[2] A Parker, A Dimitrov, and T Gedeon. Symmetry breaking clusters in soft clustering decoding of neural codes. IEEE Trans. Inform. Theor., 56:901–927, 2010.
[3] T. Gedeon, A. Parker, and A. Dimitrov. The mathematical structure of information bottleneck methods. *Entropy: Special issue for the Information Bottleneck Method*, 14:456–479, 2012.

[4] Thomas Cover and Jay Thomas. *Elements of Information Theory*. Wiley Series in Communication, New York, 1991.

[5] Robert M. Gray. *Entropy and Information Theory*. Springer-Verlag, 1990.

[6] Kenneth Rose. Deteministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE*, 86(11):2210–2239, 1998.

[7] Brendan Mumey and Tomas Gedeon. Optimal mutual information quantization is np-complete. Neural Information Coding (NIC) workshop, Snowbird UT, 2003.

[8] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 617–623. MIT Press, 2000.

[9] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. The 37th annual Allerton Conference on Communication, Control, and Computing, 1999.

[10] Noam Slonim. The information bottleneck: Theory and applications. Doctoral Thesis, Hebrew University, 2002.

[11] 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.

[12] 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.

[13] Tomas Gedeon, Albert E. Parker, and Alexander G. Dimitrov. Information distortion and neural coding. *Canadian Applied Mathematics Quarterly*, 10(1):33–70, 2003.

[14] Albert Parker, Tomas Gedeon, and Alexander Dimitrov. Annealing and the rate distortion problem. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, volume 15, pages 969–976. MIT Press, Cambridge, MA, 2003.

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