A Mapping Approach to Rate-Distortion Computation and Analysis

Kenneth Rose, *Member, IEEE*

Abstract—**In rate-distortion theory, results are often derived and stated in terms of the optimizing density over the reproduction space. In this paper, the problem is reformulated in terms of the optimal mapping from the unit interval with Lebesgue measure that would induce the desired reproduction probability density. This results in optimality conditions that are "random relatives" of the known Lloyd optimality conditions for deterministic quantizers. The validity of the mapping approach is assured by fundamental isomorphism theorems for measure spaces. We show that for the squared error distortion, *the optimal reproduction random variable is purely discrete at supercritical distortion (where the Shannon lower bound is not tight)*. The Gaussian source is thus the only source that produces continuous reproduction variables for the entire range of positive rate. To analyze the evolution of the optimal reproduction distribution, we use the mapping formulation and establish an analogy to statistical mechanics. The solutions are given by the distribution at isothermal statistical equilibrium, and are parameterized by the temperature in direct correspondence to the parametric solution of the variational equations in rate-distortion theory. The analysis of an annealing process shows how the number of "symbols" grows as the system undergoes phase transitions. Thus, an algorithm based on the mapping approach often needs but a few variables to find the exact solution, while the Blahut algorithm would only approach it at the limit of infinite resolution. Finally, a quick "deterministic annealing" algorithm to generate the rate-distortion curve is suggested. The resulting curve is exact as long as continuous phase transitions in the process are accurately followed.**

Index Terms—**Rate-distortion theory, statistical physics, Shannon lower bound, annealing, phase transitions.**

I. Introduction

THE fundamental results of rate-distortion theory are due to Shannon [1], [2]. These are the coding theorems which provide an (asymptotically) achievable bound on the performance of source coding methods. This bound is often expressed as a rate-distortion function $R(D)$ for a given source whose curve separates the region of feasible operating points (R, D) from the region that cannot be attained by any coding system. Important extensions of the theory to more general classes of sources than those originally considered by Shannon have been developed since (see [3] and [4]).

Explicit analytical evaluation of the rate-distortion function has been generally elusive, except for very few examples of sources and distortion measures (see [5] and [6] for results on the magnitude error distortion measure). Two main approaches were taken to address this problem. The first was to develop bounds to rate-distortion functions. An important example is the Shannon lower bound [2], which is useful for difference distortion measures. The second main approach was to develop a numerical algorithm, the Blahut or Blahut–Arimoto algorithm (BA) [7], [8], to evaluate rate-distortion functions. The power of the second approach is in that the function can be approximated arbitrarily closely at the cost of complexity. The disadvantage is that the complexity may become overwhelming, particularly in the case of continuous alphabets, on which we focus in this paper, and even more so for continuous vector alphabets where the complexity could grow exponentially with the dimensions. Another disadvantage is, of course, that no closed-form expression is obtained for the function, even if a simple one happens to exist.

In this paper, a new approach to rate-distortion computation and analysis is suggested. We shall restrict our attention here to continuous alphabet sources. Much of the existing theory concerns optimization over the output density, and this is indeed the heart of the Blahut algorithm. In the new approach, we consider a mapping from the unit interval with the Lebesgue measure (i.e., uniform density) to the output space. Instead of optimizing the output density directly, we optimize this mapping. The theoretical equivalence of the mapping approach (MA) to the traditional approach is shown by isomorphism theorems for topological measure spaces. Although equivalent in principle, the MA formulation is different, and by deriving the results from this angle, some new insights are gained, as well as more efficient numerical approaches to compute the rate-distortion function. The objective of this work is to exploit the potential of MA (and the resulting statistical physics intuition) along these two lines, namely, theoretical results on the rate-distortion curve and on the optimizing output density, and numerical methods for rate-distortion function computation.

In Section II, MA is presented and its equivalence to the usual approach is discussed. The MA optimality conditions are derived, and are shown to relate directly to

Manuscript received August 12, 1992; revised April 25, 1994. This work was supported by the Engineering Foundation with the cooperation of the IEEE under Grant RI-A-92-12. This paper was presented in part at the 1993 IEEE International Symposium on Information Theory, San Antonio, TX, January 1993.

The author is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

IEEE Log Number 9406321.

0018-9448/94$04.00 © 1994 IEEE

known conditions for optimal quantizer design. They are "random coding" relatives of the Lloyd optimality conditions [9], [10]. In Section III, it is proved that for the squared error distortion, the optimizing reproduction variable is purely discrete as long as the rate-distortion function does not coincide with the Shannon lower bound. In other words, except for the case that the bound is attained (e.g., Gaussian source for all positive rates), the output density consists of singularities. This could explain why explicit expressions for the rate-distortion function are so hard to obtain. Historical credit is due to Fix [11], who showed by a different approach that the optimizing output is discrete if the source density's support is not the entire real line. This special case is covered by the results of Section III because, for such sources, the Shannon lower bound is strictly lower than the rate-distortion function at all positive distortions.

The practical implications of this result are related to the work of Benjamin [12] on rate distortion for discrete sources with variable reproducing alphabets, and to the work of Finamore and Pearlman [13] who derived a rate-distortion theory for continuous alphabet sources with finite output constraint. This is due to the fact that the optimal output alphabet is often discrete and finite. Algorithms based on MA will be related to the algorithms suggested in [13] and in [12], except that the number of symbols will grow as necessary to obtain the unconstrained rate-distortion result.

Section IV is concerned with the analysis of the evolution of the optimizing output densities as we decrease the distortion (that is, as we "crawl up" the rate-distortion curve). Here, we start by showing that the functional that is minimized to find the optimizing density is the free energy of an appropriately defined statistical mechanics system whose energy is the distortion. The slope parameter is inversely related to the temperature in the physical analogy, and the optimizing output density is given by the isothermal equilibrium distribution at the given temperature. Thus, "crawling up" the rate-distortion curve is simply an annealing process in statistical mechanics. The analysis shows that the annealing process starts with a single output symbol at $R = 0$, and consists of a sequence of phase transitions which normally increase the number of symbols (or singularities). It is shown that the last phase transition occurs when the rate-distortion curve hits the Shannon lower bound, and where the singularities split or, rather, explode into a continuous distribution. It should be noted that Berger [4, pp. 253–259] showed the equivalence between the discrete rate-distortion problem and the multiphase equilibrium problem which involves minimizing the Gibbs free energy, in support of suggested potential contributions of rate-distortion theory to the understanding of (mainly biological) information systems. That physical analogy, however, does not appear useful for our purpose here of analyzing the evolution of optimal output densities.

Section V extends the results of Sections III and IV-C to higher dimensions. It focuses on certain difficulties that complicate the derivation. These are due to possible continuity of the output over low-dimensional subspaces. This discussion was postponed to this late section in order to keep the basic derivation as simple as possible.

Section VI discusses the applicability of the mapping approach to practical computation of rate-distortion functions. It is observed that as long as no discretization is used, both MA and BA will find the globally optimal output density. However, discretization for numerical computation results in differing performance. BA optimizes over a grid of points in the output space to obtain an approximate solution (whose quality depends on the resolution of the grid). MA uses the mapping which adapts its effective grid to the source distribution and so is more efficient. Moreover, as long as the Shannon lower bound is not tight, the optimal density is discrete (and finite) so that a few variables allow MA to find the exact solution, which BA approximates using the entire grid. Note also that once the Shannon lower bound is attained, we can explicitly derive the solution, so numerical evaluation is no longer necessary. This section also includes a sketch of a mapping algorithm for squared error distortion.

To avoid confusion, a note should be made concerning our use of the term "random coding." In this paper, a probability distribution is defined over the space of all possible deterministic codes, and expectations are computed with respect to this distribution to characterize the performance of this random code, as indeed it is often done elsewhere in the Shannon theory. However, this is not used directly to prove results that are asymptotic in the block length (except, of course, for the obvious asymptotic significance of the rate-distortion function).

II. The Variational Equations and the Mapping Approach

The rate-distortion curve is obtained by minimizing the mutual information subject to an average distortion constraint. Formally stated, given a continuous source alphabet $\mathscr{X}$, random variable X with a probability measure given by the density $p(x)$, and an output alphabet $\mathscr{Y}$, the problem is of optimizing

$$I(X;Y) = \int dx\, p(x) \int dy\, q(y \mid x) \log \left[\frac{q(y \mid x)}{\int dx\, p(x) q(y \mid x)} \right] \tag{1}$$

over the random encoders $q(y \mid x)$, subject to

$$\int dx\, p(x) \int dy\, q(y \mid x) d(x,y) \leq D \tag{2}$$

where $d(\cdot,\cdot)$ is the distortion measure. By replacing the above minimization by parametric variational equations (see [3], [4], [14], or [15]), this problem can be reformulated as a problem of optimization over the output density

$q(y)$. The functional to be minimized is

$$F(q) = -\frac{1}{\beta}\int dx\, p(x) \log \int dy\, q(y) e^{-\beta d(x,y)}. \tag{3}$$

The choice of the positive parameter β rather than the somewhat more common negative slope parameter $s = -\beta$ is for reasons that will become obvious when the relation to statistical mechanics is discussed (similar notation was used in [3] and [16]). Another nonessential change is that we divide the usual functional by β. Optimizing (3) subject to $\int dy\, q(y) = 1$, one gets the known conditions for optimality, and the Blahut algorithm (BA) as fixed point iterations based on these conditions.

Note that the objective of the above optimization is to determine a probability measure on the output space $\mathcal{Y}$. Let us consider an alternative approach: the mapping approach (MA). Instead of searching for the optimal $q(y)$ directly, we can search for the optimal mapping $y : [0,1] \to \mathcal{Y}$, where to the unit interval we assign the Lebesgue measure, which we denote by μ. In other words, we induce a σ-homomorphism or set mapping of $\mathcal{Y}$ into the unit interval, and consequently we induce a mapping of measures on the unit interval into measures on $\mathcal{Y}$. The equivalence of the approaches is ensured by the theory of general measures in topological spaces (see, for example, [17, ch. 15] or [18, ch. 2–3]). In particular, the Borel isomorphism theorem states that a complete separable metric space with a finite Borel measure is isomorphic to the unit interval with Lebesgue measure, and hence a corresponding point-to-point mapping exists. A more precise statement of the theorem requires special attention to the case of a measure with atoms. This will be of considerable importance to us later, but let us ignore it at this point.

We thus have to minimize the functional

$$F(y) = -\frac{1}{\beta}\int dx\, p(x) \log \int_{[0,1]} d\mu(u) e^{-\beta d[x,y(u)]} \tag{4}$$

over the mapping $y(u)$.

Before continuing, let us write the necessary condition for optimality, which is obtained by calculus of variations (see part A of the Appendix):

$$\int dx\, p(x) \left[\frac{e^{-\beta d[x,y(u)]}}{\int d\mu(u) e^{-\beta d[x,y(u)]}} \right] \frac{\partial}{\partial y} d[x, y(u)] = 0, \tag{5}$$

which is the Euler equation corresponding to the variational problem at hand. This condition must be satisfied almost everywhere with respect to $\mu(u)$. Since disagreement on a set of measure zero is of no importance here, we restrict the discussion to functions $y(u)$ that satisfy the condition everywhere. We define the support of Y as the set of values that the random variable Y can *possibly* take, that is, the range of $y(u)$. Thus, the support of Y is a subset of the points y_o satisfying

$$\int dx\, p(x) \left[\frac{e^{-\beta d(x,y_o)}}{\int d\mu(u) e^{-\beta d[x,y(u)]}} \right] \frac{\partial}{\partial y_o} d(x, y_o) = 0. \tag{6}$$

Note that, for simplicity, we have ignored the case of boundary points (if the output space is bounded) in the above derivation.

To interpret this result, we define transition probabilities as

$$q(u \mid x) = \frac{e^{-\beta d[x,y(u)]}}{\int d\mu(u) e^{-\beta d[x,y(u)]}}. \tag{7}$$

This definition will be justified in Section IV. Using Bayes' theorem and the fact that $\mu(u)$ is the Lebesgue measure on [0, 1], the condition (5) can be rewritten compactly as

$$\int dx\, p(x \mid u) \frac{\partial}{\partial y} d[x, y(u)] = 0 \tag{8}$$

for all $u \in [0,1]$, and similarly for all points of support of Y. This result is nicely related to the Lloyd necessary conditions for optimal quantizer design [9], [10]: the transition probabilities (7) are random relatives of the nearest neighbor rule, and the condition (8) is a random relative of the centroid rule. This connection is used in the deterministic annealing approach to vector quantizer design [19].

In summary, we see two approaches, namely BA and MA, whose equivalence follows from the Borel isomorphism theorem. However, these approaches are substantially different in their computational complexity and performance if we need to discretize (as we usually do). When using BA, discretization means defining a grid $\{y_i\}$ on the output space $\mathcal{Y}$. In MA, we "discretize the unit interval" (i.e., replace it by a set of indexes) and induce a grid on $\mathcal{Y}$ by our mapping. The mapping will adapt this grid to the source, and therefore less complexity will be needed to obtain the same precision. This difference between the approaches is more fundamental because the output densities are often discrete and finite, as is explained in the following sections. This gives MA the theoretical capability of producing *exact solutions* at finite resolution, while BA can only approach them at the limit of infinite resolution.

It is important to note that discretized MA, when one allows a general probability measure over the set of indexes, as indeed one should, is closely related to the alphabet-size constrained rate distortion [13] at a given β and with the "right" number of symbols, and to the work on rate-distortion of discrete sources with continuous reproductions [12]. In the sequel, it is shown how the rate-distortion curve is obtained by annealing within the MA framework, without *a priori* limitation on the number of output symbols.

III. Optimal Reproduction is Discrete Unless the Shannon Lower Bound is Tight

In this section, we assume the squared error distortion measure $d(x,y) = |x - y|^2$. For simplicity, we restrict our treatment to scalars. The derivation for higher dimensions parallels the scalar derivation, but requires more care, and will be addressed in a later section. The reader should note that the insights leading to understanding the discrete nature of the optimal reproduction variable come naturally from the mapping approach and the statistical physics analogy of the next section. However, the results can be proved directly without any essential use of the mapping approach, as we shall point out specifically where appropriate in this section.

Let us assume that the support of the reproduction random variable Y contains an interval I_o, i.e., $I_o \subset y([0,1])$. First, the condition for optimality (6) must be satisfied for all $y_o \in I_o$. For the squared error distortion, it states

$$\int dx p(x)\lambda(x)(x - y_o)e^{-\beta(x-y_o)^2} = 0 \qquad (9)$$

where

$$\lambda(x) = \left[\int d\mu(u)e^{-\beta[x-y(u)]^2}\right]^{-1} = \left[\int dy q(y)e^{-\beta(x-y)^2}\right]^{-1}.$$

We rewrite the condition as

$$\int dx p(x)\lambda(x)\frac{\partial}{\partial y_o}e^{-\beta(x-y_o)^2} = 0. \qquad (10)$$

Since the left-hand side vanishes everywhere in the interval I_o, all its derivatives must be zero at y_o. We therefore write

$$\int dx p(x)\lambda(x)\frac{\partial^n}{\partial y_o^n}e^{-\beta(x-y_o)^2} = 0, \qquad n \geq 1. \quad (11)$$

For the sake of readability, we shall not provide rigorous justification for the change in order of differentiation and integration. A more careful proof of a stronger claim is given in part B of the Appendix.

Note that the set of equations (11) can also be directly obtained by differentiating the Kuhn–Tucker optimality condition for a continuous alphabet (e.g., [15, p. 98]) at $y_o \in I_o$, instead of using our mapping approach optimality condition (6).

Next, we observe that

$$\frac{d^n}{dz^n}e^{-\beta z^2} = H_n(z)e^{-\beta z^2} \qquad (12)$$

where $H_n(z)$ is the nth Hermite polynomial with respect to the weight function $e^{-\beta z^2}$. We thus rewrite (11) as

$$\int dx p(x)\lambda(x)H_n(x - y_o)e^{-\beta(x-y_o)^2} = 0, \qquad n \geq 1. \quad (13)$$

Hence, $p(x)\lambda(x)$ is orthogonal to all Hermite polynomials of degree greater than zero. Since the Hermite polynomials form a complete orthogonal system in $L^2[e^{-\beta(x-y_o)^2}]$, we get

$$p(x)\lambda(x) = \text{constant} \qquad (14)$$

(subject to some restrictive assumption such as $p\lambda \in L^2[e^{-\beta(x-y_o)^2}]$, which will be relaxed in part B of the Appendix). The constant is determined by normalization, and the result is

$$p(x) = \sqrt{\frac{\beta}{\pi}}\int d\mu(u)e^{-\beta[x-y(u)]^2}, \qquad (15)$$

which can be rewritten directly in terms of the reproduction density (if Y has one):

$$p(x) = \sqrt{\frac{\beta}{\pi}}\int dy q(y)e^{-\beta(x-y)^2}. \qquad (16)$$

If the optimal output density $q(y)$ satisfies (16), then the Shannon lower bound equals the rate distortion function [4]. The previous equation (15) states the equivalent condition for a general output probability measure. Let us summarize the result in the form of a theorem.

Theorem 1: If the support of the optimal reproduction random variable contains some nonempty open set, then the rate-distortion function coincides with the Shannon lower bound.

Corollary 1: If the Shannon lower bound does not hold with equality, then the optimal reproduction random variable is purely singular.

A stronger theorem which implies Theorem 1 is the following:

Theorem 2: If the support of the optimal reproduction random variable has an accumulation point, then the rate-distortion function coincides with the Shannon lower bound.

Since Theorem 2 implies Theorem 1, a more careful (and lengthier) proof for it is given in part B of the Appendix.

Corollary 2: If the Shannon lower bound does not hold with equality, then the support of the optimal reproduction random variable consists of isolated singularities. Further, if this support is bounded, then Y is discrete and finite.

Corollary 2 follows directly from Theorem 2 and the Bolzano–Weierstrass theorem. In particular, if the source has bounded support, then so does the output which must thus be discrete and finite, a result due to Fix [11], who also suggested to bound the cardinality of the support by using Jensen's theorem on the number of zeros of an entire function within a disk.

In order to obtain the general description of the continuity of the reproduction random variable as we vary the distortion, we review some known results on the coincidence of $R(D)$ and the Shannon lower bound (SLB). The first point to make is that coincidence at distortion D_o implies coincidence at all positive $D < D_o$. It is well known, and can also be seen from the convolution in (16) or in (15), that SLB coincides with $R(D)$ if and only if the source random variable can be written as the sum of two independent random variables:

$$X = Y + N$$

where Y is the reproduction and N is normal, zero-mean with variance D. This condition is referred to as the "backward channel" condition. Now, if at distortion D_o the backward channel condition is satisfied, i.e., $X = Y_o + N_o$ with appropriate Y_o and N_o, then for all $D < D_o$, we can write $N_o = N + N^*$ where N and N^* are independent zero-mean normal random variables with variances D and $D_o - D$, respectively. Thus, by choosing $Y = Y_o + N^*$, we satisfy the backward channel condition $X = Y + N$.

This implies that there is a unique distortion level, called the *critical distortion* D_c, such that $R(D) = \text{SLB}$ for all $D < D_c$ and $R(D) > \text{SLB}$ for all $D > D_c$. Moreover, at subcritical distortion $D < D_c$, the reproduction variable is *absolutely continuous* since its density can be obtained by convolution with a normal density (let D_o be such that $D < D_o < D_c$; then $Y = Y_o + N^*$ where N^* is the corresponding independent normal variable as above).

The following theorem summarizes the continuity properties of the optimal reproduction variable.

Theorem 3: The optimal reproduction random variable is absolutely continuous at subcritical distortion, and is purely discrete at supercritical distortion. At critical distortion, the optimal reproduction probability measure may have both absolutely continuous and singular parts.

At subcritical distortion, the rate-distortion problem is analytically resolved as $R(D)$ coincides with the Shannon lower bound. The open problem is thus the behavior at supercritical distortion. Having seen that at supercritical distortion the optimal reproduction variable is discrete, we next consider the evolution of these discrete solutions as we vary the distortion. We address this problem by establishing the equivalence with a fundamental problem in statistical physics.

IV. Phase Transitions in Rate-Distortion Computation

In this section, we show that our functional (4), which is optimized to find points on the rate-distortion curve, is the free energy associated with a corresponding statistical mechanics system whose energy function is the distortion. The analogy is shown by using random coding in two ways. First, we consider the random encoder for a given fixed "codebook" [fixed mapping $y(u)$], and then we define a probability distribution over all possible deterministic codes and characterize the performance by taking expectations. The latter views the random code as an appropriate Gibbs canonical ensemble, and relates to the former by marginalization. The minimization of the free energy is equivalent to reaching isothermal equilibrium. Our parameter β is inversely related to the temperature (we take the Boltzmann constant to be one). This derivation facilitates an analysis of the evolution of the output distributions as β is increased. All this eventually leads to suggesting a deterministic annealing method for computing the rate-distortion curve, which is similar to the deterministic annealing approach to vector quantization [19] and for mass-constrained clustering [20].

Please note that the recourse to statistical physics is by no means necessary here, as the mathematical theories of bifurcation would have been sufficient to derive most of the results. This choice of presentation simply reflects the source of intuition leading to the results herein, including those already described in Section III. Moreover, by observing the equivalence to fundamental problems in statistical physics, one can take advantage of the powerful mathematical tools and methods that have been developed in this field, as will be specifically pointed out.

A. Transition Probabilities

To derive the transition probabilities, we define the source as a (possibly uncountable) set of vector points $\{p\}$, whose density at $x \in \mathscr{X}$ is given by $p(x)$. We assume that each point in the set is encoded independently into a "codeword" $u \in [0, 1]$, and that the transition probabilities depend only on the location x, so we can denote them $q(u \mid x)$. We further assume that the mapping $y : [0,1] \rightarrow \mathscr{Y}$ is fixed. We can now write the distortion as

$$D = \int dx p(x) \int d\mu(u) q(u \mid x) \, d[x, y(u)]. \tag{17}$$

This is the energy in our physical analogy. To compute the distribution governing the system, we use Jaynes' formalism [21] based on the principle of maximum entropy. More precisely, we compute the transition (conditional) densities by maximizing the conditional entropy (we purposely ignore the subtleties associated with the continuous case, as one could use coarse-graining and the discrete entropy to obtain the same results, as indeed is commonly done in statistical mechanics; see [22] for such a discrete derivation)

$$H = -\int dx p(x) \int d\mu(u) q(u \mid x) \log q(u \mid x) \tag{18}$$

subject to the energy constraint (17). We, of course, obtain the Gibbs distribution of (7):

$$q(u \mid x) = \frac{e^{-\beta d[x, y(u)]}}{\int d\mu(u) e^{-\beta d[x, y(u)]}}. \tag{19}$$

The normalization function in the denominator is known as the partition function in statistical mechanics:

$$Z_x = \int d\mu(u) e^{-\beta d[x, y(u)]}. \tag{20}$$

By the assumption that each point p is independently encoded, the total partition function (accounting for encoding the entire input set $\{p\}$) is "$Z = \prod_p Z_{x(p)}$." As $\{p\}$ is possibly uncountable, we write more precisely

$$Z = e^{\int dx p(x) \log Z_x}. \tag{21}$$

The free energy is therefore

$$F = -\frac{1}{\beta}\log Z = -\frac{1}{\beta}\int dx\, p(x) \log \int d\mu(u) e^{-\beta d[x, y(u)]}. \tag{22}$$

Note that *the free-energy of our physical system is exactly the functional (4) that we need to optimize to find a point on the rate-distortion curve.*

B. Randomizing Over All Codes

In this subsection, we extend the derivation, and drop the impossible assumption that the mapping $y(u)$ is fixed. Instead of considering the encoder probabilities, we consider the *probability distribution over all deterministic codes*; in other words, an ensemble of codes. A code is given by a mapping $y(u)$ where $y: [0,1] \to \mathcal{Y}$, and a deterministic encoding rule via an encoder/selector function $s(p)$ where $s: \{p\} \to [0,1]$. We write $s(p) = u_p$ where u_p is the "codeword" assigned to p, and where the reproduction value is $y(u_p)$. This notation parallels the notation in the vector quantization literature [23]. The encoding rule can be equivalently given by the choice of values for the set of variables $\{u_p\}$. Note that, in principle, we allow points p at the same location x to be encoded differently. Let $f(y,s)$ denote the probability density over the space of deterministic codes, i.e., over the product space of all possible mappings $\{y\}$ and all possible selector functions $\{s\}$. We determine $f(y,s)$ by maximizing the entropy

$$H = -\int dy \int ds f(y,s) \log f(y,s) \tag{23}$$

subject to the expected distortion constraint

$$\int dy \int ds D(y,s) f(y,s) = D \tag{24}$$

where the distortion due to the code (y,s) is

$$D(y,s) = \int dp\, d[x(p), y(s(p))]. \tag{25}$$

We obtain the Gibbs distribution

$$f(y,s) = \frac{e^{-\beta D(y,s)}}{\int dy \int ds\, e^{-\beta D(y,s)}}, \tag{26}$$

assuming that the normalizing integral exists. Next, we compute the marginal probability density over the space of mapping functions $\{y\}$:

$$f(y) = \int ds f(y,s). \tag{27}$$

We first consider the following average all possible selector functions:

$$\langle e^{-\beta D(y,s)} \rangle_s = \int e^{-\beta \int dp\, d[x(p), y(u_p)]} \prod_p d\mu(u_p), \tag{28}$$

which made use of (25). Had $\{p\}$ been a finite set where each element has unit weight (the typical training set used in practice), then we could immediately write the above average as the product of integrals:

$$\begin{aligned} \langle e^{-\beta D(y,s)} \rangle_s &= \prod_p \int_{[0,1]} d\mu(u) e^{-\beta d[x(p), y(u)]} \\ &= \prod_p Z_{x(p)}(y) \end{aligned} \tag{29}$$

where the rightmost equation used (20). As the set $\{p\}$ is generally uncountable, we rewrite (29) more precisely:

$$\langle e^{-\beta D(y,s)} \rangle_s = e^{\int dx\, p(x) \log Z_x(y)} = Z(y) \tag{30}$$

where we also used (21). Note that here we make the dependence of Z_x and Z on y explicit because, in this subsection, the mapping $y(u)$ is no longer assumed to be fixed. It follows from (26), (27), and (30) that

$$f(y) = \frac{Z(y)}{\int dy Z(y)} = \frac{e^{-\beta F(y)}}{\int dy\, e^{-\beta F(y)}} \tag{31}$$

where $F(y)$ is the free energy as defined in (4) and again in (22).

In order to find the most probable mapping $y(u)$ at a given β, we maximize (31) or, equivalently, we minimize the effective cost function $F(y)$, that is, we minimize the free energy. Note that minimum free energy defines isothermal equilibrium of the stochastic system in the physical analogy. This gives the direct relation to annealing once we gradually increase β (decrease the temperature). We thus see that the optimal output density, or the optimal mapping $y(u)$, determines equilibrium at a given temperature (i.e., at a given rate-distortion slope). A process of annealing consists of starting at $\beta = 0$ and gradually increasing β while maintaining the system at equilibrium. Hence, *annealing is equivalent to computing the rate-distortion curve*, starting at $R = 0$ and "crawling" up the curve. This is the topic of the next subsection.

C. Phase Transitions in the Annealing Process

Unlike the more general derivation of the previous two subsections, in this part, we restrict our derivation to scalars and to the squared error distortion measure $d(x,y) = (x-y)^2$. Generalization to vectors will be discussed in a later section. At supercritical distortion, the optimal reproduction variable is discrete, and for simplicity, let us assume that its support is finite.

We start by considering the case of $\beta = 0$ (extremely high temperature). It is easy to see that the optimal mapping satisfying (5) is obtained by mapping the entire unit interval to a single point y_o which minimizes $\int dx p(x)(x - y_o)^2$; hence, y_o is the center of mass of the input distribution. The average distortion is the variance of the input. At this temperature, the cardinality of the effective reproduction alphabet is one. As we lower the temperature, the cardinality will change (it will generally grow). We consider differing output cardinalities as *phases* of the physical system, and the purpose of this subsection is to provide insights on the evolution of the system via a *phase transition analysis*. Some skepticism may be raised about the notion of phase transitions, especially since the rate-distortion curve is continuous and well behaved, and no phase transitions are detected by its examination. However, the existence, and importance, of phase transitions is evident from a more appropriate (parametric) representation. In Fig. 1 we see the curve of distortion versus slope (or, equivalently, energy versus temperature) for the simple case of a source with uniform density. This is a typical phase diagram in physics.

The continuity of the $R(D)$ curve implies that the system undergoes *continuous* (second-order) phase transitions, which are, in fact, symmetry breaks. An elegant and powerful approach for the analysis of such phase transitions is the Landau theory [24] (another relevant treatment can be found in [25]) which uses the theory of symmetry groups. A continuous phase transition means that there exists a critical temperature such that, above it, the state of the system is invariant to a transformation group, while below it, it is only invariant to a subgroup. Thus, continuous phase transitions correspond to breaks in symmetry. The Landau theory addresses the analysis by introducing the basic concepts of *order parameter* and the *Landau variational free energy*. The advantages of a careful application of the Landau theory to rate-distortion analysis are still under investigation. In this subsection, some preliminary results are given to demonstrate the usefulness of addressing the rate-distortion problem by phase transition analysis.

A continuous phase transition occurs when the optimal mapping $y(u)$ stops defining the minimum of the free energy, and becomes a saddle point. Specifically, at the critical β, the mapping $y(u)$ satisfies the usual constraint

$$\frac{\partial}{\partial\epsilon}F(y+\epsilon\eta)\,|_{\epsilon=0}=0, \qquad \forall\eta(u),$$

but there exists a particular perturbation $\eta(u)$ such that

$$\frac{\partial^2}{\partial\epsilon^2}F(y+\epsilon\eta)\,|_{\epsilon=0}=0.$$

Because of the discrete nature of the reproduction distribution, this translates into a simpler condition (for a detailed derivation, see part C of the Appendix): the transition occurs when some point of support y_o satisfies both (10), which we rewrite here as

$$\int dx p(x)\lambda(x)\frac{\partial}{\partial y_o}e^{-\beta(x-y_o)^2}=0 \tag{32}$$

and

$$\int dx p(x)\lambda(x)\frac{\partial^2}{\partial y_o^2}e^{-\beta(x-y_o)^2}=0. \tag{33}$$

The first condition (32) is needed for y_o to be a point of support. The second condition (33) is equivalent to

$$1-2\beta\sigma^2_{x|y_o}=0 \tag{34}$$

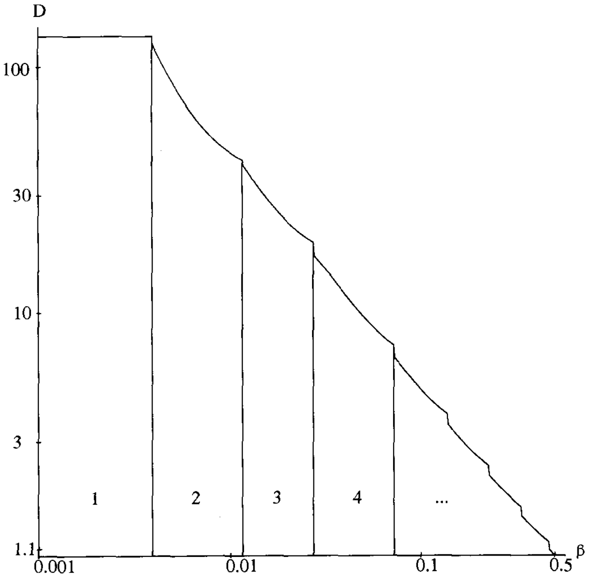

Fig. 1. This phase diagram was produced by simulation on a uniform source $[-20,20]$. Distortion versus β on logarithmic scale clearly shows the critical β where phase transitions occur. The cardinality (number of symbols) of the effective reproduction alphabet is marked within the region of the corresponding phase. Note that the apparent discontinuity of the transitions is due to the discrete jumps in β, and to the fact that in the simulation they occur slightly later than they should.

where the variance is computed based on the backward conditional density

$$\sigma^2_{x|y_o}=\int dx p(x|y_o)(x-y_o)^2.$$

Note that it is easy to show that if $y_o=y(u_o)$ for some $u_o\in[0,1]$, then

$$p(x|y_o)=p(x|u_o)=p(x)\frac{e^{-\beta(x-y_o)^2}}{\int d\mu(u)e^{-\beta[x-y(u)^2]}}.$$

From (34), the condition on the critical β is that there exists a point of support y_o such that

$$\beta_c=\frac{1}{2\sigma^2_{x|y_o}}. \tag{35}$$

In other words, a reproduction symbol will split when the variance of its inverse image in the source space becomes large enough with respect to the temperature. We thus have a description of the annealing process, where at each phase the effective reproduction alphabet consists of reproduction symbols which are centroids of "fuzzy" clusters in the input space. Phase transitions occur so as to maintain the cluster variance below $1/2\beta$. As we increase β, when a cluster's variance becomes equal to $1/2\beta$, it splits into smaller clusters, and the number of symbols increases.

For the first phase transition, (35) is easily related to known results from rate-distortion theory. For example, if the input is Gaussian, the curve hits the $R = 0$ axis with the slope

$$s = -\beta = -\frac{1}{2D_{\max}} = -\frac{1}{2\sigma_x^2}$$

where σ_x^2 is the source variance [4], as predicted by (35). We are interested in less trivial outcomes than the behavior at $R = 0$. Note that the result of (35) is not restricted to Gaussians, nor to the first phase transition. It implies that we obtain a sequence of phase transitions, corresponding to reproduction symbol splits when the variance of their inverse image in the input space becomes large enough with respect to the temperature.

Three important comments should be made at this point. First, we note that the condition for phase transition is equivalent to the first two conditions for continuity of the reproduction variable, namely, (11) or (13) for $n = 1, 2$. We say that a phase transition occurred at β_c satisfying (35) if the condition was *not* satisfied at nearby $\beta < \beta_c$. In other words, the *last* phase transition occurs at *critical distortion.* At subcritical distortion, the rate distortion curve coincides with the Shannon lower bound and (35) is satisfied everywhere. The Gaussian source is the unique source whose rate-distortion curve coincides with the Shannon lower bound for all positive rate R. It is therefore the unique source whose reproduction variable is absolutely continuous at all positive rates. From our viewpoint of phase transitions, the Gaussian's first phase transition which happens at $\beta = 1/2\sigma_x^2$ is also the last phase transition, and results in continuous output density.

Second, it is important to note that there exists another type of phase transition that we have not analyzed here. We considered symbol splits where the resulting new symbols continuously move away. Another possibility is a new symbol that grows continuously from zero mass. The continuity of the rate-distortion curve [4], [15] implies that these two are the only possible types of phase transition. The requirement that the distortion (energy) varies continuously with β implies that either the change in the mapping $y(u)$ is continuous (splitting), or that discontinuous change in mapping is over a set of measure zero (growing). We only obtained conditions for splitting because we used a linear form for perturbation $y + \epsilon\eta$ in our derivation above. The analysis of "mass growing" phase transitions is currently under investigation, and it is believed that a careful Landau theory formulation will yield a complete description of the process. At this point, it is not difficult to see that these new points of support can be anticipated and tracked by "massless" symbols long before they start growing their mass. Further discussion of this is given in Section VI.

The third comment concerns the issue of whether the cardinality of the effective reproduction alphabet grows monotonically with β. This section was written in a way that implicitly makes this assumption (by using the terms "splitting" and "mass growing"). However, this is a conjecture. In physics, there are only very few exceptional examples of transitions where the symmetry group below the critical temperature is not subgroup of the symmetry group above critical temperature [24, sect. 142]. This strongly supports our conjecture for the case of the simple elastic system we have here. In any case, this conjecture is not essential for the above derivation, nor for the computational methods to be described later, as both allow for occasional "merging" or "vanishing" of symbols.

Summarizing the current results together with the results of Section III, we state that as long as the rate-distortion curve has not merged with the Shannon lower bound, the optimizing output density is expected to be discrete, with the number of symbols changing according to a sequence of distinct phase transitions. We have derived the condition for mass-splitting bifurcation which relates the covariance of the symbol's inverse image in the input space to the annealing parameter β or to the slope of the rate-distortion curve $s = -\beta$.

V. Higher Dimensions

The results in Sections III and IV-C were derived for scalar sources. Although the treatment of higher dimensions is generally similar, some additional difficulties need to be addressed that would have somewhat obscured the basic results, and therefore had been postponed until now.

The derivation of Section III is easily extended to obtain the corresponding result: the support of the optimal reproduction contains an open set only if the rate-distortion function coincides with the Shannon lower bound. We now require the optimality condition to be satisfied in a small ball around the vector $y_o = y(u_o)$. The point y_o must satisfy

$$\int dx\, p(x)\lambda(x)(x - y_o)e^{-\beta|x-y_o|^2} = 0, \qquad (36)$$

which is a *vector* equation. Since this condition is satisfied by all points in the ball, all its directional derivatives must vanish at y_o. Let the unit vector e represent an arbitrary direction, and let $x_e = x^t e$ and $y_{oe} = y_o^t e$ be the corresponding components along e. We obtain the following condition for all $n \geq 1$:

$$\int dx\, p(x)\lambda(x)\frac{\partial^n}{\partial y_{oe}^n}e^{-\beta|x-y_o|^2} = 0, \qquad (37)$$

which we rewrite as

$$\int dx\, p(x)\lambda(x)H_n(x_e - y_{oe})e^{-\beta|x-y_o|^2}$$
$$= \int dx H_n(x_e - y_{oe})p(x \mid y_o) = 0. \qquad (38)$$

Integrating (38) over the subspace orthogonal to e, we get a condition on the marginal $p(x_e \mid y_o)$:

$$\int dx_e H_n(x_e - y_{oe})p(x_e \mid y_o) = 0, \qquad \forall n \geq 1. \qquad (39)$$

This implies that $p(x_e \mid y_o)$ is Gaussian (similarly to the derivation in part B of the Appendix):

$$p(x_e \mid y_o) = \sqrt{\frac{\beta}{\pi}}\, e^{-\beta (x_e - y_{oe})^2}. \tag{40}$$

We found that the marginal of $p(x \mid y_o)$ *along any arbitrary direction* is Gaussian; hence, it must be an isotropic Gaussian:

$$p(x \mid y_o) = \left(\frac{\beta}{\pi}\right)^{n/2} e^{-\beta |x - y_o|^2}. \tag{41}$$

But since we also have

$$p(x \mid y_o) = p(x)\lambda(x) e^{-\beta |x - y_o|^2}$$

then

$$p(x)\lambda(x) = \left(\frac{\beta}{\pi}\right)^{n/2} \tag{42}$$

and we conclude that

$$p(x) = \left(\frac{\beta}{\pi}\right)^{n/2} \int d\mu(u) e^{-\beta |x - y(u)|^2} \tag{43}$$

or, if the output density exists,

$$p(x) = \left(\frac{\beta}{\pi}\right)^{n/2} \int dy\, q(y) e^{-\beta |x - y|^2}. \tag{44}$$

In other words, the rate distortion function coincides with the Shannon lower bound. *We have thus obtained in the vector case that the reproduction variable is singular unless the Shannon lower bound is tight.* The problem with higher dimensions is that not all singularities are necessarily point singularities. To obtain the condition for other (non-point) singularities, let us assume that the reproduction density is continuous at y_o along direction e, but not necessarily along other directions. This immediately gives (40), which says that $p(x_e \mid y_o)$ is Gaussian with mean y_{oe} and variance $1/2\beta$. Further, this is also satisfied by all the other points on the line $y = y_o + ae$ where a is a scalar variable.

We now extend the derivation of the condition for phase transition to higher dimensions. At phase transition, the scalar conditions of (32) and (33) are replaced by the following conditions at a point of support y_o:

$$\int dx\, p(x)\lambda(x) \frac{\partial}{\partial y_o} e^{-\beta |x - y_o|^2} = 0 \tag{45}$$

and the (Hessian) matrix

$$\int dx\, p(x)\lambda(x) \left(\frac{\partial}{\partial y_o}\right)\left(\frac{\partial}{\partial y_o}\right)^t e^{-\beta |x - y_o|^2} \tag{46}$$

is singular (i.e., it is no longer positive definite). The above derivatives are, of course, short-hand notation for gradients. It is easy to see that the expression in (46) is (up to a positive multiplicative constant) the matrix

$$I - 2\beta C_{x \mid y_o} \tag{47}$$

where $C_{x \mid y_o}$ is the covariance matrix

$$C_{x \mid y_o} = \int dx\, p(x \mid y_o)(x - y_o)(x - y_o)^t. \tag{48}$$

The critical value for β is determined by the first eigenvalue of (47) to become zero. In fact,

$$\beta_c = \frac{1}{2\lambda_{\max}} \tag{49}$$

where $\lambda_{\max}$ is the largest eigenvalue of the covariance $C_{x \mid y_o}$. Thus, phase transitions occur when the temperature is lowered to twice the variance along the principal axis of the "cluster" or inverse image in the source space. Now, if besides satisfying the second moment condition (49), the marginal distribution of the principal component is *exactly Gaussian*, then condition (40) is satisfied, and the phase transition results in continuity along the principal axis. In other words, we have a singularity which is not a point singularity, as explained before.

This process is nicely illustrated by a nonisotropic multivariate Gaussian source [4, pp. 108–111]. The output density starts as a singularity at the center of mass of the input density. When $\beta = 1/2\lambda_{\max}$, it becomes continuous along a line, the major principal axis of the input density. As β is increased, it reaches values corresponding to smaller eigenvalues, and the output density spreads along the corresponding principal directions. As it does, the output dimensionality is increasing. Only when β reaches the value $1/2\lambda_{\min}$ do we get continuity on an open subset of Y (in fact, we get continuity over the entire space), and exactly at this β, the Shannon lower bound for the vector source is attained. However, it is obvious from (40) that continuity within hyperplanes will only happen for a very restricted class of sources. Usually, we will only have point singularities to deal with. Moreover, in numerical computation where the source is discretized, the optimal reproduction variable will always be purely discrete.

VI. Solution Quality, Complexity, and Future Improvements

Both methods, BA and MA, yield the optimal solution in the continuous case if initialized strictly within the convex region "boundaries" in the probability space.[1] For BA, this requires the initialization to satisfy $q(y) > 0, \forall y$, while for MA, we require equivalently that $y : [0,1] \to \mathcal{Y}$ is onto. Discretizing the methods means that we have to initialize the iterations *on the boundary.* Clearly, the density has to vanish everywhere except at the grid points.

[1] By "probability space" here, we mean a space where each point represents a choice of output probability density. The boundary of this space consists of points representing densities which vanish at some $y \in \mathcal{Y}$.

For BA, this becomes the original discrete case [7], and is therefore ensured to converge to the optimal distribution [26] *on this boundary* (i.e., for the given grid). MA, on the other hand, can modify the grid by the mapping. It is thus not confined to distributions defined on the fixed grid. However, it is not guaranteed to obtain the globally optimal solution. This last statement should not necessarily be understood as a disadvantage with respect to BA. The solution obtained by MA is always optimal for the grid it converges to, just like BA's optimality on its original grid. Multiple local minima are caused by this ability to modify the grid, i.e., it is not guaranteed to find the optimal grid. To see the nonconvexity in the finite discrete case, consider the case of $\beta \to \infty$. In this case, our mapping solution tries to minimize the distortion subject to a fixed number of output symbols. This is exactly the vector quantizer design problem which is known to have nonconvex cost function. The existence of multiple local optima gives strong motivation for the use of annealing.

From the previous sections, we know that the output density solution often consists of a finite number of singularities. This means that MA can produce the exact solution with this finite number of variables. BA would approach this solution only at the limit of infinite grid resolution.

The annealing approach starts at the global optimum at $\beta = 0$ and tracks it while gradually increasing β. It is thus crawling up on the rate distortion curve. It will not leave the curve as long as we can follow all second-order phase transitions in the process. The open issue here is our ability to follow the mass-growing phase transitions. Some discussion of it is given later in this section.

It is hoped that the mapping approach and the statistical physics analogy would provide means for improving the analytic solution of the rate distortion problem. This is the reason why the emphasis in this paper is placed on the general approach rather than the specific derivable algorithms. Let us now consider a sketch of an algorithm based on the approach. The algorithm is given for scalars and for the squared error distortion measure so that we can give simple interpretation. Note that this example can easily be generalized to higher dimensions and to other distortion measures.

A deterministic annealing algorithm sketch:

1) Initialize $\beta = (1/2\sigma_x^2) - \epsilon$, $K = 1$, $y_1 = \int dx\,xp(x)$, and $q(y_1) = 1$.

2) Update for $i = 1,\cdots, K$:

$$y_i = \frac{\int dx\,xp(x)q(y_i \mid x)}{q(y_i)}$$

where

$$q(y_i \mid x) = \frac{q(y_i)e^{-\beta(x-y_i)^2}}{\sum_{j=1}^{K} q(y_j)e^{-\beta(x-y_j)^2}}$$

$$q(y_i) = \int dx\,p(x)q(y_i \mid x).$$

3) Compute

$$D = \int dx\,p(x) \sum_{j=1}^{K} q(y_j \mid x)(x - y_j)^2.$$

4) Convergence test. If not satisfied, go to 2.

5) Compute

$$R = \int dx\,p(x) \sum_{j=1}^{K} q(y_j \mid x) \log \left[\frac{q(y_j \mid x)}{q(y_j)}\right]$$

and save (R, D).

6) Increment β and check condition for phase transition for $i = 1,\cdots, K$. If critical β is reached for symbol j, add a new symbol $y_{K+1} = y_j + \delta$ and increment K.

7) Go to 2.

Note that the test for critical β in step 6, if considered expensive for high dimensions, can be replaced by a simple perturbation. In this case, we always keep two symbols at each location, and perturb them when we update β. Until the critical β is reached, they will be merged together by the iterations. At phase transition they will separate.

The relation to the Lloyd algorithm for quantizer design is easy to see. At a given β, the iteration is a generalization of the nearest neighbor and the centroid conditions. The relation to maximum-likelihood estimation of parameters in normal mixtures is also obvious (for the use of deterministic annealing in multiscale clustering, see [22] and [20]).

The above simple algorithm does not attempt to detect mass-growing phase transitions. In Fig. 2, we show its performance on a uniform source (where indeed such phase transitions do occur). It is indistinguishable from the curve produced by BA. In fact, the MA curve is everywhere slightly better than the BA curve, but this can only be seen by zooming in (Fig. 3). This is explained by our ability to place the reproduction values precisely where they are needed, while BA is constrained to a fixed grid. Apparently, the mass-growing phase transitions did not cause noticeable harm because, whenever they were missed, a "compensating" split happened soon afterwards. Considering complexity, MA appears to be dramatically more efficient. No extensive experimentation was performed to substantiate this. But for the above example, the rate-distortion curve produced by MA using a very demanding convergence threshold at each β (maximum squared change in the parameters less than 10^{-12}), and an exponential annealing schedule $\beta(n + 1) = 1.01\beta(n)$, in the range $0.001 \leq \beta < 0.5$, was generated in 14 h. The BA solution, using the same convergence threshold and for the same set of β, using a fixed grid of only 16 points, ran for over a week on the same machine (Sun SparcStation 2). Of course, these results do depend to some extent on the efficiency of the respective programs, but they do suggest that there is much to be gained by the new approach.

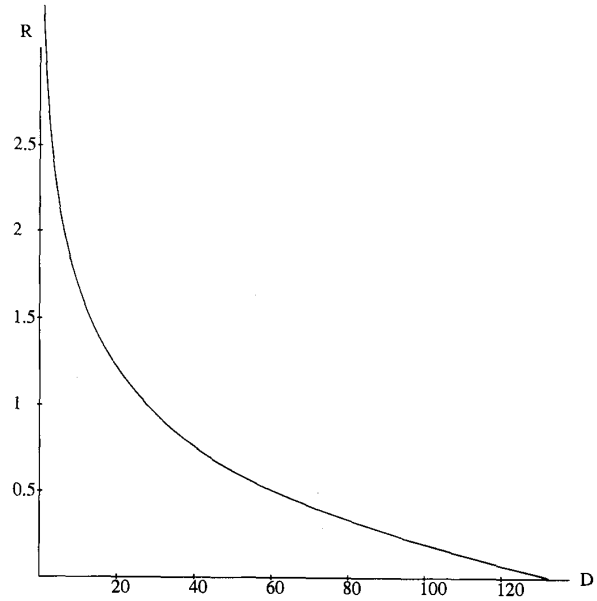

Fig. 2. Rate versus distortion for the same uniform source. The result using BA is not distinguishable from the result of MA at this scale.

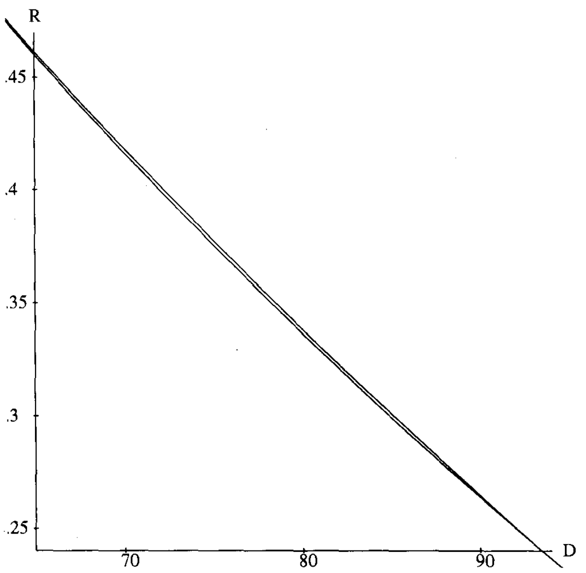

Fig. 3. Zooming in to show the rate distortion at distortion between 70 and 90. BA produced the higher curve (but will improve at higher grid resolution).

Let us now briefly discuss "mass-growing" phase transitions. As mentioned before, this issue is currently under investigation. However, some "hand-waving" arguments can be given to explain the expected solution to the problem. The basic idea is that our optimality condition (9) gives not only the points of support of the reproduction at given β, but also all "invisible" points that may grow mass at lower temperature. To see this, we note that solutions of (9) also give optima of

$$c(y) = \int dx p(x) \lambda(x) e^{-\beta(x-y)^2}.$$

If y_o maximizes $c(\cdot)$ and $c(y_o)$ takes the value one, then y_o is a point of support of $\mathcal{Y}$. All other solutions of (9) can be tracked using massless variables which will start growing mass only when $c(y_o)$ becomes one. Note that we are still dealing with a discrete set of variables. One practical approach under consideration consists of tracking the zeros of derivatives of $c(y)$.

On the other hand, work on a theoretical analysis is underway to determine the precise conditions for a "mass-growing" phase transition, and to predict the critical β and the location of such growth. These ideas are being researched, as well as their implications on the problem of vector quantizer design.

VII. Summary

A mapping approach has been suggested for rate distortion computation and analysis. For continuous source alphabet, it is equivalent to the Blahut algorithm in principle, but its fixed-point iterations are extensions of the Lloyd algorithm to random quantization. Using MA, we establish an analogy to statistical mechanics, where crawling up the rate distortion curve is equivalent to annealing of the corresponding physical system. As long as the Shannon lower bound does not hold with equality, the optimizing reproduction variable is discrete, and the number of symbols generally grows as the system undergoes phase transitions. The "pathological" case of Gaussian input results in a continuous output density at all positive rates. Normally, however, the discrete nature of the solution makes the discretized mapping approach very attractive, as few variables are sufficient to obtain the exact solution. Deterministic annealing can be used to generate the rate distortion curve, and the result is exact as long as second-order transitions are accurately followed. The analysis of phase transitions corresponding to symbol splits has been performed for the squared distance measure.

Appendix

A. Necessary Condition for Optimal Mapping

We wish to minimize the functional

$$F(y) = -\frac{1}{\beta}\int dx p(x) \log \int_{[0,1]} d\mu(u) e^{-\beta d[x,y(u)]} \qquad (50)$$

over the mapping $y(u)$. In order to obtain the necessary condition, we apply the standard procedure in variational calculus. We require

$$\frac{\partial}{\partial \epsilon} F(y+\epsilon\eta)\,|_{\epsilon=0} = 0$$

for all admissible[2] perturbation functions $\eta(u)$. Applying the condition to (50), we get

$$\int dx\, p(x) \frac{\int d\mu(u)\eta(u)\frac{\partial}{\partial y}d[x, y(u)]e^{-\beta d[x,y(u)]}}{\int d\mu(u)e^{-\beta d[x,y(u)]}} = 0,$$

which can be rewritten as

$$\int d\mu(u)\eta(u) \cdot \left\{ \int dx\, p(x)\left[\frac{e^{-\beta d[x,y(u)]}}{\int d\mu(u)e^{-\beta d[x,y(u)]}}\right]\frac{\partial}{\partial y}d[x, y(u)]\right\} = 0.$$

The equality for all admissible perturbations η requires the expression in braces to vanish almost everywhere with respect to $\mu(u)$. This gives the necessary condition for optimality as the corresponding Euler equation:

$$\int dx\, p(x)\left[\frac{e^{-\beta d[x,y(u)]}}{\int d\mu(u)e^{-\beta d[x,y(u)]}}\right]\frac{\partial}{\partial y}d[x, y(u)] = 0,$$

μ-almost everywhere.

B. Proof of Theorem 2

The theorem states that if the optimal reproduction support has an accumulation point, then the rate distortion function coincides with the Shannon lower bound.

Let us assume that the support of the reproduction random variable Y has an accumulation point y_o, i.e., for every $\epsilon > 0$, there exists a δ such that $0 < |\delta| < \epsilon$ and $y_o - \delta$ is a point of support of Y. It must satisfy the condition for optimality (6), which we rewrite for the squared error distortion:

$$\int dx\, p(x)\lambda(x)(x - y_o + \delta)e^{-\beta(x-y_o+\delta)^2} = 0 \qquad (51)$$

where

$$\lambda(x) = \left[\int d\mu(u)e^{-\beta[x-y(u)]^2}\right]^{-1}.$$

We rewrite the condition as

$$\frac{\partial}{\partial\delta}\int dx\, p(x)\lambda(x)e^{-\beta(x-y_o+\delta)^2} = 0. \qquad (52)$$

(Note that the integrals in (51) and (52) are uniformly convergent.) Next, we observe that

$$e^{-\beta(z+\delta)^2} = e^{-\beta z^2}\sum_{n=0}^{\infty}\frac{1}{n!}H_n(z)\delta^n \qquad (53)$$

where $H_n(z)$ is the nth Hermite polynomial with respect to the weight function $e^{-\beta z^2}$. Substituting into (52), we get

$$\frac{\partial}{\partial\delta}\int dx\, p(x)\lambda(x)e^{-\beta(x-y_o)^2}\sum_{n=0}^{\infty}\frac{1}{n!}H_n(x - y_o)\delta^n = 0. \qquad (54)$$

The next step is to interchange summation with integration. For justification, consider the series

$$\sum_{n=0}^{\infty}\frac{1}{n!}H_n(z)\delta^n$$

as a series of functions of z. The series of absolute values of all terms is the expansion in series of the function $e^{\beta(|\delta|^2+2|\delta||z|)}$. Thus, in particular,

$$\left|\sum_{n=0}^{N}\frac{1}{n!}H_n(z)\delta^n\right| \leq e^{\beta(|\delta|^2+2|\delta||z|)}, \qquad \forall N.$$

Hence, to use Lebesgue's dominated convergence theorem, all we need is to show that

$$\int dx\, p(x)\lambda(x)e^{-\beta(x-y_o)^2}e^{\beta(|\delta|^2+2|\delta||x-y_o|)}$$

is integrable, which we do by breaking into two parts:

$$e^{2\beta|\delta|^2}\left[\int_{y_o}^{\infty}dx\, p(x)\lambda(x)e^{-\beta(x-y_o-|\delta|)^2} + \int_{-\infty}^{y_o}dx\, p(x)\lambda(x)e^{-\beta(x-y_o+|\delta|)^2}\right].$$

Each part is integrable since, by the Kuhn–Tucker conditions for the optimal reproduction random variable (e.g., [15]),

$$\int_{-\infty}^{\infty}dx\, p(x)\lambda(x)e^{-\beta(x-z)^2} \leq 1, \qquad \forall z.$$

This establishes the conditions for Lebesgue's dominated convergence theorem.

We can now return to (54) and obtain

$$\sum_{n=1}^{\infty}\frac{1}{(n-1)!}\delta^{(n-1)} \cdot \int dx\, p(x)\lambda(x)H_n(x - y_o)e^{-\beta(x-y_o)^2} = 0. \qquad (55)$$

It is obvious from the power series on the left-hand side that for $\epsilon > 0$ small enough, it must be either identically zero, or nowhere zero for all $0 < |\delta| < \epsilon$. Since the latter contradicts our basic assumption that y_o is an accumulation point, it must be identically zero at some neighborhood of y_o (which is the situation in Theorem 1). Consequently, all terms in (55) must vanish, and we write

$$\int dx\, p(x)\lambda(x)H_n(x - y_o)e^{-\beta(x-y_o)^2} = 0, \qquad n \geq 1, \qquad (56)$$

i.e., $p(x)\lambda(x)$ is orthogonal to all Hermite polynomials of degree greater than zero. Using the fact that

$$p(x \mid y_o) = p(x)\lambda(x)e^{-\beta(x-y_o)^2} \qquad (57)$$

this result can be rewritten as

$$\int dx\, H_n(x - y_o)p(x \mid y_o) = 0, \qquad n \geq 1. \qquad (58)$$

An obvious solution to this set of equations is the Gaussian density

$$p(x \mid y_o) = \sqrt{\frac{\beta}{\pi}}\, e^{-\beta(x-y_o)^2}. \qquad (59)$$

[2] Since we derive a necessary condition, we do not need to be too careful about how restrictive our definition of admissibility is. Hence, we simply require that admissible functions be measurable, that the integrals exist, and that changing the order of integration and differentiation (where needed) is allowed.

We want to verify that it is the (substantially) unique distribution satisfying (58). We note that (58) determines all the moments of $p(x \mid y_o)$, which therefore must be the same as the moments of the Gaussian of (59). But the moment sequence associated with Gaussians satisfies Carleman's general criterion [27], [28, p. 19], and therefore it uniquely determines the corresponding distribution. Hence, we have that (58) implies (59).

Combining (57) and (59), we get

$$p(x)\lambda(x) = \sqrt{\frac{\beta}{\pi}} \tag{60}$$

and

$$p(x) = \sqrt{\frac{\beta}{\pi}} \int d\mu(u) e^{-\beta[x-y(u)]^2}, \tag{61}$$

which imply that the Shannon lower bound coincides with the rate distortion function.

C. Necessary Condition for Phase Transition

A necessary condition for y to be the optimal mapping (minimum of F) is

$$\frac{\partial}{\partial \epsilon} F(y + \epsilon\eta) \mid_{\epsilon=0} = 0, \qquad \forall \eta(u) \tag{62}$$

and

$$\frac{\partial^2}{\partial \epsilon^2} F(y + \epsilon\eta) \mid_{\epsilon=0} \geq 0. \tag{63}$$

A necessary condition for bifurcation is to have exact equality in (63) for some perturbation η (we disregard the question of higher derivatives as we are concerned with necessary conditions). After straightforward differentiation, we get the condition for equality in (63):

$$\begin{aligned} &\int dx p(x)\lambda(x) \int d\mu(u)\eta^2(u) \\ &\quad \cdot [1 - 2\beta(x - y(u))^2] e^{-\beta(x-y(u))^2} \\ &+ \int dx p(x)\lambda^2(x) \\ &\quad \cdot \left[\int d\mu(u)\eta(u)(x - y(u)) e^{-\beta(x-y(u))^2}\right]^2 = 0. \end{aligned} \tag{64}$$

We claim that the sum is positive for all η *if and only if* the first term is. The "if" part is trivial since the second term is obviously nonnegative. To prove the "only if" part, we use the knowledge that the output is discrete. For the first term to be nonpositive, there must be at least one point of support y_o of nonzero mass such that

$$\int dx p(x)\lambda(x)\left[1 - 2\beta(x - y_o)^2\right] e^{-\beta(x-y_o)^2} \leq 0.$$

The mass of y_o is the measure of the subset $I_o \in [0, 1]$ that is mapped to y_o, i.e., $y(I_o) = y_o$. Let us choose $\eta(u)$ such that $\eta(u) = 0$, $\forall u \notin I_o$ to ensure that the first term is nonpositive. Since $y(u) = y_o$ for all u such that $\eta(u) \neq 0$, the second term in (64) becomes

$$\int dx p(x)\lambda^2(x)\left[(x - y_o) e^{-\beta(x-y_o)^2} \int_{I_o} d\mu(u)\eta(u)\right]^2.$$

But we have not yet specified the function $\eta(u)$ at $u \in I_o$, which we can always define so that

$$\int_{I_o} d\mu(u)\eta(u) = 0.$$

Hence, whenever the first term is not positive, we can choose η such that the second term vanishes. The conclusion is that *we have strict inequality in (63) for all η iff the first term of (64) is positive.*

The condition for phase transition—equality in (63)—can be restated as follows: there exists some point of support y_o for which we have

$$\int dx p(x)\lambda(x)\left[1 - 2\beta(x - y_o)^2\right] e^{-\beta(x-y_o)^2} = 0$$

or, more compactly using the backward transition density definition,

$$1 - 2\beta\sigma^2_{x \mid y_o} = 0 \tag{65}$$

where

$$\sigma^2_{x \mid y_o} = \int dx (x - y_o)^2 p(x)\lambda(x) e^{-\beta(x-y_o)^2}.$$

The critical value for β is thus

$$\beta_c = \frac{1}{2\sigma^2_{x \mid y_o}}. \tag{66}$$

It is straightforward to extend the above derivation to the vector case. In the vector case, the condition for phase transition is that there exists a point of support y_o for which the matrix

$$I - 2\beta C_{x \mid y_o},$$

where $C_{x \mid y_o}$ is the covariance matrix of $p(x \mid y_o)$, is no longer positive definite. The critical value for β is therefore

$$\beta_c = \frac{1}{2\lambda_{\max}} \tag{67}$$

where $\lambda_{\max}$ is the largest eigenvalue of $C_{x \mid y_o}$.

Acknowledgment

The author is grateful to R. M. Gray who prompted him to find deeper relations between rate-distortion theory and the deterministic annealing approach to vector quantization, and for making insightful comments on an early draft of this paper. He is also grateful to T. Berger and an anonymous reviewer for a very thorough review that considerably improved the paper. W. A. Pearlman pointed out the reference to Fix's paper [11].

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[2] ——, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Nat. Conv. Rec.*, no. 4, 1959, pp. 142–163.

[3] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[4] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[5] H. H. Tan and K. Yao, "Evaluation of rate-distortion functions for a class of independent identically distributed sources under an absolute magnitude criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 59–64, Jan. 1975.

[6] K. Yao and H. H. Tan, "Absolute error rate-distortion functions for sources with constrained magnitudes," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 499–503, July 1978.

[7] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.

[8] S. Arimoto, "An algorithm for calculating the capacity of an arbitrary discrete memoryless channel," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 14–20, Jan. 1972.

[9] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 129–137, Mar. 1982. (Reprint of the 1957 paper.)

[10] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.

[11] S. L. Fix, "Rate distortion functions for squared error distortion measures," in *Proc. 16th Annu. Allerton Conf. Commun., Contr., Comput.*, Oct. 1978.

[12] T. W. Benjamin, "Rate distortion functions for discrete sources with continuous reproductions," Master's thesis, Cornell Univ., Ithaca, NY, 1973.

[13] W. A. Finamore and W. A. Pearlman, "Optimal encoding of discrete-time continuous-amplitude memoryless sources with finite output alphabets," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 144–155, Mar. 1980.

[14] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.

[15] R. M. Gray, *Source Coding Theory*. Boston, MA: Kluwer Academic, 1990.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[17] H. L. Royden, *Real Analysis*, 3rd ed. New York: Macmillan, 1988.

[18] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag, 1988.

[19] K. Rose, E. Gurewitz, and G. C. Fox, "Vector quantization by deterministic annealing," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1249–1257, July 1992.

[20] ——, "Constrained clustering as an optimization method," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 785–794, Aug. 1993.

[21] E. T. Jaynes, "Information theory and statistical mechanics," in *Papers on Probability, Statistics and Statistical Physics*, R. D. Rosenkrantz, Ed. Dordrecht, The Netherlands: Kluwer Academic, 1989. (Reprints of the original 1957 papers in *Phys. Rev.*)

[22] K. Rose, E. Gurewitz, and G. C. Fox, "Statistical mechanics and phase transitions in clustering," *Phys. Rev. Lett.*, vol. 65, pp. 945–948, 1990.

[23] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer Academic, 1991.

[24] L. D. Landau and E. M. Lifshitz, *Statistical Physics, Part 1*, 3rd ed. Oxford: Pergamon, 1980.

[25] J.-C. Toledano and P. Toledano, *The Landau Theory of Phase Transitions*. Teaneck, NJ: World Scientific, 1987.

[26] I. Csiszàr, "On the computation of rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 122–124, Jan. 1974.

[27] T. Carleman, "Sur le problème des moments," *Comptes Rendus Acad. Sci. Paris*, vol. 174, pp. 1680–1682, 1922.

[28] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*, 2nd ed. Providence, RI: Amer. Math. Soc., 1950.