

## Rate-Distortion Bound for Joint Compression and Classification

Yanting Dong and Lawrence Carin  
Department of Electrical and Computer Engineering  
Duke University  
Box 90291

Durham, NC 27708-0291

Email: lcarin@ee.duke.edu PH: 919-660-5270 FAX: 919-660-5293

**Abstract** - Rate-distortion theory is applied to the problem of joint compression and classification. A Lagrangian distortion measure is used to consider both the squared Euclidean error in reconstructing the original data as well as the classification performance. The bound is calculated based on an alternating-minimization procedure, representing an extension of the Blahut-Arimoto algorithm. As an example application, we consider a hidden Markov model (HMM) source, and the objective is to quantize the source outputs and estimate the underlying HMM state sequence (based on the quantized data). We present bounds on the minimum rate required to achieve desired average distortion on signal reconstruction and state-estimation accuracy.

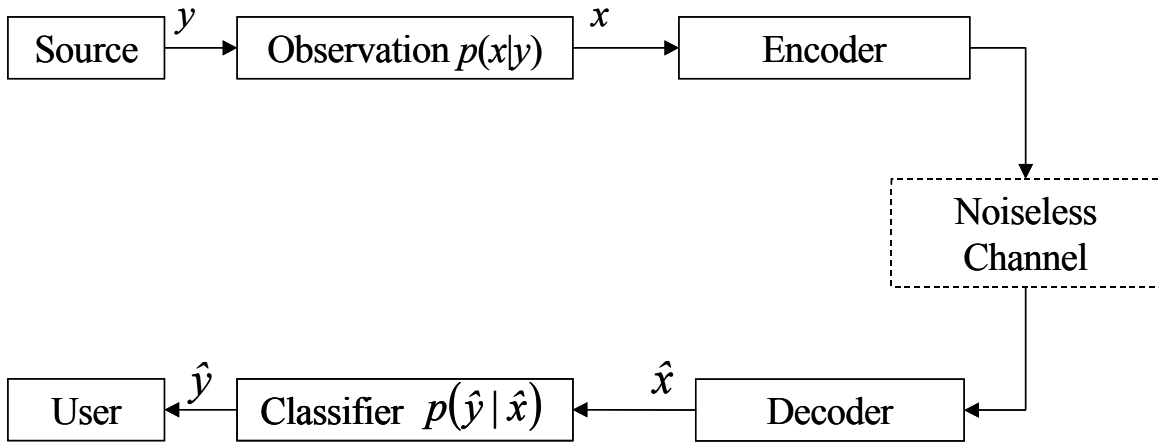
### I. Introduction

We investigate compression applications for which reliable signal reconstruction is required, and additionally good classification performance is desired based on the quantized data. This is important in many compression applications for which classification is the ultimate goal. For example in medical-image compression or digitization [1,2] it is desired that the compression be performed in a manner that yields good image reconstruction, while also accounting for the ultimate diagnostic goals. This is also of interest in sensing applications [3,4], in which one may wish to detect and/or classify a target based on data quantized by a remote sensor. In this setting it is of interest to quantize the data in a manner that accounts for the ultimate classification task. In addition, since the data may also be interpreted by a human, good signal reconstruction is desired [1-4].

We seek to quantify the minimum encoding rate  $R$  such that a prescribed average distortion  $D$  is achieved, represented as  $R(D)$  [5,6]. Here  $D$  considers both the data-

reconstruction error, quantified by the average squared Euclidean distance between the original and reconstructed data, and the classification performance, measured by the probability of classification error. Although rate-distortion theory was developed originally for data compression it has been shown that, by using proper distortion measures, this theory may be used to derive performance bounds for various problems. For example, its application to classification problems has been studied [7-9]. In this case the source is characterized by different classes and the distortion as the probability of classification error.

The previously developed structure for classification-performance analysis is extended here to the joint compression and classification problem. The key is defining a distortion measure that incorporates both the squared Euclidean distortion *and* the classification error, resulting in a Lagrangian form of distortion. This type of distortion has been developed previously in the context of a Bayes-VQ encoder [1,2]. There are three issues addressed in this paper: (i) extension of the Source Coding Theorem to the case of a distortion measure in which reconstruction and classification performance are addressed simultaneously; (ii) augmentation of the Blahut-Arimoto algorithm [10,11] for such a distortion measure; and (iii) example results for a hidden Markov model (HMM) source, in which there is interest in the accuracy of reconstructing the observed data as well as in estimating the underlying (unobserved) HMM



**Figure 1.** Structure for applying rate-distortion analysis to the joint compression and classification problem.

states. The principal difference between this work and previous studies is that the distortion matrix  $\rho(x, \hat{x})$  for source outputs  $x$  and reconstruction symbols  $\hat{x}$  is a function of the

conditional density  $p(\hat{x}|x)$ , since  $p(\hat{x}|x)$  defines the optimal mapping between the encoded  $\hat{x}$  and the estimated unobserved  $y$  (with estimate denoted  $\hat{y}$ ). This is summarized in Fig. 1. We demonstrate in Sec. II that  $R(D)$  computations apply to such sources and distortion measures, employing a relatively simple modification to the proof of the Source Coding Theorem.

The remainder of the paper is organized as follows. In Sec. II the generalized distortion measure is defined, followed by a proof of the applicability of this distortion measure to the Source Coding Theorem. The Blahut-Arimoto algorithm [10,11] is then extended in Sec. III to the case of a composite distortion measure, with example  $R(D)$  results presented in Sec. IV for an HMM source. Conclusions are discussed in Sec. V.

## II. Generalized Source Coding Theorem

### A. Problem statement

Assume a composite source emits the random variables  $(x,y)$ , with density function  $p(x,y) = p(y)p(x|y)$ , where  $y \in A_y$  represents a discrete label and  $x \in A_x$  represents a discrete random variable conditioned on the value of  $y$ . As an example,  $y$  may represent the state of a system, with state-dependent statistics for  $x$ . It is assumed that  $x$  represents the observed data while the underlying  $y$  is “hidden” (see Fig. 1). Our objective is to develop a source code for  $x$ , from which we yield the approximate reconstruction  $\hat{x}$ , with an associated squared Euclidean distortion  $(x - \hat{x})^2$ . However, we wish to encode  $x$  in a manner that optimizes our ability to estimate the underlying  $y$ , with estimate denoted  $\hat{y}$ . A natural means of characterizing the distortion between  $y$  and  $\hat{y}$  is the Hamming distortion  $h(y, \hat{y})$ , which is zero if  $y = \hat{y}$  and one otherwise.

With the goal of minimizing the average square Euclidian error between  $(x, \hat{x})$  and the average Hamming distortion between  $(y, \hat{y})$ , we employ the composite distortion measure

$$\rho(x, y; \hat{x}, \hat{y}) = (x - \hat{x})^2 + \lambda h(y, \hat{y}) \tag{1}$$

where  $\lambda$  is a real number characterizing the desired relative importance placed on reconstruction error for  $x$ ,  $(x - \hat{x})^2$ , and in estimation error for  $y$ ,  $h(y, \hat{y})$ .

We now assume that the source emits a *sequence* of  $n$  symbols  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ , with an underlying but unobserved sequence  $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ . It is assumed that the sequence of symbols is iid, characterized by density functions  $p(y)$  and  $p(x|y)$ . The coding process yields a reconstruction sequence  $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n\}$ , from which a classification algorithm is used to estimate the underlying  $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$ . The generalized distortion for the length- $n$  sequence is

$$\rho_n(\mathbf{x}, \mathbf{y}; \hat{\mathbf{x}}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^n \rho(x_i, y_i; \hat{x}_i, \hat{y}_i) \quad (2)$$

Considering the algorithm for estimation of  $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$ , we note that the likelihood  $p(y|\hat{x})$  is expressed in terms of the underlying source encoder, through which  $p(\hat{x}|x)$  is defined, in particular

$$p(y|\hat{x}) = \frac{\sum_{x \in A_x} p(y) p(x|y) p(\hat{x}|x)}{\sum_{y \in A_y} \sum_{x \in A_x} p(y) p(x|y) p(\hat{x}|x)} \quad (3)$$

The optimal classifier, with which (2) is minimized, is to associate  $\hat{x}$  with that  $y$  that maximizes  $p(y|\hat{x})$ , assuming equal costs for all  $y \in A_y$ , thereby yielding the mapping  $\hat{x} \rightarrow \hat{y}$ . With knowledge of the source statistics  $p(y)$  and  $p(x|y)$ , and of  $p(\hat{x}|x)$ , the optimal Bayesian classifier is simply a look-up table, with each  $\hat{x}$  mapped to a particular  $\hat{y}$ .

From (3) we note that the mapping  $\hat{x} \rightarrow \hat{y}$  is effected through knowledge of the source statistics  $p(y)$  and  $p(x|y)$ , as well as the encoder statistics  $p(\hat{x}|x)$ . Therefore, the generalized distortion  $\rho(x_n, y_n; \hat{x}_n, \hat{y}_n)$  may be simplified to  $\rho(x_n, y_n; \hat{x}_n)$ , since  $\hat{y}_n$  is directly determined from  $\hat{x}_n$ .

It is important to emphasize that given any  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$  and  $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n\}$  the average squared Euclidian error  $\frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2$  may be computed without knowledge of  $p(\hat{x}|x)$ , and it is under these conditions that Shannon's Source Coding Theorem [5] has been developed. It is therefore of interest to examine whether this theorem may be extended to the distortion measure in (1) and (2), for which  $p(\hat{x}|x)$  is implicitly required to implement the classifier  $\hat{x} \rightarrow \hat{y}$  and hence to quantify the Hamming distortion in (1). This extension of the Source Coding Theorem is discussed in the next section.

The average distortion is defined as

$$E[\rho] = E_{x,x'}\{(x - \hat{x})^2\} + \lambda E_{y,y'}\{h(y, \hat{y})\} \quad (4)$$

where we again emphasize that  $p(\hat{x}|x)$  is required for its computation. In particular,

$$p(y, \hat{y}) = \sum_{x \in A_x} \sum_{\hat{x} \in A_x} p(y)p(x|y)p(\hat{x}|x)p(\hat{y}|\hat{x}) .$$

We define  $Q_D$  as the set of  $p(\hat{x}|x)$  for which  $E[\rho] \leq D$  for a prescribed average distortion  $D$ , i.e.

$$Q_D = \{p(\hat{x}|x) : E(\rho) \leq D\} \quad (5)$$

and in the next section we demonstrate that the minimum average rate required to encode the source, such that the average distortion is less than or equal to  $D$  is expressed as

$$R(D) = \min_{p(\hat{x}|x) \in Q_D} I(\hat{x}; x) \quad (6)$$

We now prove this is true for the distortion measure described in (1).

## B. Generalized source coding theorem

A source code  $B$  of size  $K$  and blocklength  $n$  is said to have rate  $R = n^{-1} \log K$ , and the smallest size of a  $D$ -admissible code is  $K(n, D)$ , where a  $D$ -admissible code for source  $\mathbf{x}$  satisfies  $\rho(B) = E[\rho_n(\mathbf{x}|B)] \leq D$ .

**Theorem 1** (Generalized Source Coding Theorem) Let a discrete memoryless source be defined by  $\{Y_t, P_Y\}$  and  $\{X_t, P_{X|Y}\}$ , and assume the dual-letter fidelity criterion given in (1) and (2), and let  $R(D)$  be defined as in (6). Then, given  $\varepsilon > 0$  and any  $D \geq 0$ , an integer  $n$  can be found such that there exists a  $(D + \varepsilon)$ -admissible code of blocklength  $n$  with rate  $R < R(D) + \varepsilon$ . In other words, the inequality  $n^{-1} \log K(n, D + \varepsilon) < R(D) + \varepsilon$  holds for sufficiently large  $n$ .

*Proof:* The proof presented here is a generalization of that considered in [5].

We construct the desired code ensemble  $B$  by choosing  $K$  code words. In particular, the  $n$ -dimensional  $k$ th codeword is given by  $\hat{\mathbf{x}}_k = \{\hat{x}_{k,1}, \hat{x}_{k,2}, \dots, \hat{x}_{k,n}\}$ , where each element of the codeword is a member of  $A_{\hat{x}}$ . Each codeword is selected independently according to the common density function  $Q(\hat{\mathbf{x}})$ . The  $K$  codewords constituting the codebook

$$B = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_K\} \text{ are characterized by } Q(B) = \prod_{k=1}^K Q(\hat{\mathbf{x}}_k).$$

Consider a sequence of  $n$  elements from the iid composite source characterized by  $p(y)$  and  $p(x|y)$ , denoted  $(\mathbf{x}, \mathbf{y})$  with  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$  and  $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ . If  $\mathbf{x}$  is mapped to the codeword  $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n\}$  we effect a mapping between the individual elements in  $\mathbf{x}$  and  $\hat{\mathbf{x}}$ , i.e.  $x_1 \rightarrow \hat{x}_1, x_2 \rightarrow \hat{x}_2, \dots, x_n \rightarrow \hat{x}_n$ . Therefore, for a given  $\mathbf{x}$  and  $\hat{\mathbf{x}}$  we may tabulate the occurrence rates  $r(\alpha|\beta) = n_{\alpha|\beta} / n_\beta$ , where  $n_\beta$  represents the number of times an element in the vector  $\mathbf{x}$  takes on the value  $\beta \in A_x$  and  $n_{\alpha|\beta}$  represents the number of times an element in  $\hat{\mathbf{x}}$  takes on the value  $\alpha \in A_{\hat{x}}$  if the respective element in  $\mathbf{x}$  is  $\beta$ . By the definition of probability, as  $n \rightarrow \infty$  we have  $p(\hat{x} = \alpha | x = \beta) = r(\alpha|\beta)$ . Hence, for large enough  $n$  the block encoding  $\mathbf{x} \rightarrow \hat{\mathbf{x}}$  defines  $p(\hat{\mathbf{x}}|\mathbf{x})$ , and therefore for source vectors  $(\mathbf{x}, \mathbf{y})$  and codeword  $\hat{\mathbf{x}}$  we may compute the distortion

$$\rho_n(\mathbf{x}, \mathbf{y}; \hat{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2 + \lambda \frac{1}{n} \sum_{i=1}^n h(y_i, \hat{y}_i) \quad (7)$$

where again it is understood that knowledge of  $p(\hat{x}|x)$ , defined through  $x \rightarrow \hat{x}$  for sufficient  $n$ , dictates the lookup-table mapping  $\hat{x}_i \rightarrow \hat{y}_i$ .

We now define a set of codes  $\hat{\mathbf{x}}$

$$S(\mathbf{x}, \mathbf{y}) = \{\hat{\mathbf{x}} : \rho_n(\mathbf{x}, \mathbf{y}; \hat{\mathbf{x}}) \leq D + \delta\} \quad (8)$$

where  $\delta$  is a small number defined below. The probability that a codeword chosen at random belongs to the set  $S(\mathbf{x}, \mathbf{y})$  is denoted

$$Q[S(\mathbf{x}, \mathbf{y})] = \sum_{\hat{\mathbf{x}} \in S(\mathbf{x}, \mathbf{y})} Q(\hat{\mathbf{x}}) \quad (9)$$

and the probability that no codewords of a  $K$ -dimensional, randomly chosen code will belong to  $S(\mathbf{x}, \mathbf{y})$  is represented as

$$Q^*(\mathbf{x}, \mathbf{y}) = [1 - Q(S(\mathbf{x}, \mathbf{y}))]^K \quad (10)$$

Define

$$\rho_{\max} = \max_{x, \hat{x} \in A_x \times A_{\hat{x}}} (x - \hat{x})^2 + \lambda \quad (11)$$

which represents the maximum possible distortion in (7), in which all classifications  $\hat{y}_i$  are wrong, i.e.  $h(y_i, \hat{y}_i) = 1, \forall i$ .

The average distortion across all randomly generated codes  $B$  and source outputs  $(\mathbf{x}, \mathbf{y})$  is expressed as

$$\tilde{\rho} = \sum_{\text{all } (\mathbf{x}, \mathbf{y})} p(\mathbf{y}) p(\mathbf{x}|\mathbf{y}) \sum_{\text{all } B} Q(B) \rho_n(\mathbf{x}, \mathbf{y}|B) \quad (12)$$

where  $Q(B) = \prod_{k=1}^K Q(\hat{\mathbf{x}}_k)$ . Following along the lines in Berger [5], we have

$$\tilde{\rho} \leq D + \delta + \rho_{\max} \sum_{\text{all } (\mathbf{x}, \mathbf{y})} p(\mathbf{y}) p(\mathbf{x}|\mathbf{y}) [1 - Q(S(\mathbf{x}, \mathbf{y}))]^K \quad (13)$$

We now consider any  $p(\hat{\mathbf{x}}|\mathbf{x})$  and assign to each  $\mathbf{x}$  the set of codewords  $T(\mathbf{x})$  with

$$T(\mathbf{x}) = \{\hat{\mathbf{x}} : \frac{1}{n} \log \frac{p(\hat{\mathbf{x}}|\mathbf{x})}{p(\hat{\mathbf{x}})} \leq R(D) + \delta\} \quad (14)$$

where  $p(\hat{\mathbf{x}}) = \sum_{\mathbf{x}} p(\mathbf{x})p(\hat{\mathbf{x}}|\mathbf{x})$ . For any  $\hat{\mathbf{x}} \in T(\mathbf{x})$  by definition

$p(\hat{\mathbf{x}}) \geq p(\hat{\mathbf{x}}|\mathbf{x})\exp[-n(R(D) + \delta)]$ , and utilizing  $Q(S(\mathbf{x}, \mathbf{y})) \geq Q(S(\mathbf{x}, \mathbf{y}) \cap T(\mathbf{x}))$ , we have

$$Q(S(\mathbf{x}, \mathbf{y})) \geq \exp\{-n[R(D) + \delta]\} \sum_{\hat{\mathbf{x}} \in S(\mathbf{x}, \mathbf{y}) \cap T(\mathbf{x})} p(\hat{\mathbf{x}}|\mathbf{x}) \quad (15)$$

Again following the proof in Berger [5], (15) yields

$$[1 - Q(S(\mathbf{x}, \mathbf{y}))]^K \leq 1 - \sum_{\hat{\mathbf{x}} \in S(\mathbf{x}, \mathbf{y}) \cap T(\mathbf{x})} p(\hat{\mathbf{x}}|\mathbf{x}) + e^{-K \exp[-n(R(D) + \delta)]} \quad (16)$$

Letting  $K = \lceil \exp[n(R(D) + 2\delta)] \rceil$ , where  $\lceil x \rceil$  represents the smallest integer greater than or equal to  $x$ , the term  $e^{-K \exp[-n(R(D) + \delta)]}$  vanishes for sufficiently large  $n$ . In particular, let  $n_1$  be the block length for which  $e^{-K \exp[-n(R(D) + \delta)]} < \delta / \rho_{\max}$ , and there for  $n > n_1$  and using (16), (13) becomes

$$\tilde{\rho} \leq D + 2\delta + \rho_{\max} \left[ 1 - \sum_{\text{all } (\mathbf{x}, \mathbf{y})} p(\mathbf{y})p(\mathbf{x}|\mathbf{y}) \sum_{\hat{\mathbf{x}} \in S(\mathbf{x}, \mathbf{y}) \cap T(\mathbf{x})} p(\hat{\mathbf{x}}|\mathbf{x}) \right] \quad (17)$$

We now define the set of vectors  $\mathbf{x}, \mathbf{y}$  and  $\hat{\mathbf{x}}$  satisfying

$$S = \{(\mathbf{x}, \mathbf{y}, \hat{\mathbf{x}}) : \hat{\mathbf{x}} \in S(\mathbf{x}, \mathbf{y})\} \quad (18a)$$

and similarly the set of vectors  $\mathbf{x}$  and  $\hat{\mathbf{x}}$

$$T = \{(\mathbf{x}, \hat{\mathbf{x}}) : \hat{\mathbf{x}} \in T(\mathbf{x})\} \quad (18b)$$

Following Berger [5],

$$\left[ 1 - \sum_{\text{all } (\mathbf{x}, \mathbf{y})} p(\mathbf{y})p(\mathbf{x}|\mathbf{y}) \sum_{\hat{\mathbf{x}} \in S(\mathbf{x}, \mathbf{y}) \cap T(\mathbf{x})} p(\hat{\mathbf{x}}|\mathbf{x}) \right] = 1 - p(S \cap T) = p(\bar{S} \cup \bar{T}) \leq p(\bar{S}) + p(\bar{T}) \quad (19)$$

where  $\bar{S}$  and  $\bar{T}$  are the complements respectively of  $S$  and  $T$ . We now demonstrate that  $p(\bar{S})$  and  $p(\bar{T})$  vanish for sufficiently large  $n$  and for an appropriate choice of the density function  $p(\hat{\mathbf{x}}|\mathbf{x})$ , and hence for  $Q(\hat{\mathbf{x}}) = \sum_{\text{all } \mathbf{x}} p(\hat{\mathbf{x}}|\mathbf{x})p(\mathbf{x})$ .

In particular, we choose  $p(\hat{\mathbf{x}}|\mathbf{x}) = \prod_{i=1}^n p(\hat{x}_i|x_i)$ , where the  $p(\hat{x}_i|x_i)$  is that associated with the definition of  $R(D)$ . Note that by considering  $p(\hat{\mathbf{x}}|\mathbf{x})$  we are assuming codes of sufficiently

long length  $n$  such that  $p(x = \alpha | x = \beta) \approx r(\alpha | \beta)$ , in the manner discussed above. Assuming an iid source, it follows that  $p(\mathbf{x}, \hat{\mathbf{x}}) = \prod_{i=1}^n p(x_i) p(\hat{x}_i | x_i)$ , where  $p(x_i) = \sum_y p(x_i | y) p(y)$ .

Using  $p(\hat{x}|x)$  from the definition of  $R(D)$ , the likelihood  $p(y|\hat{x})$  in (3) is defined, yielding maximum-likelihood mappings  $\hat{x}_i \rightarrow \hat{y}_i$ . We therefore may compute the distortion in (7). By the weak law of large numbers, for sufficiently large  $n$ ,  $\rho_n(\mathbf{x}, \mathbf{y}; \hat{\mathbf{x}}) \rightarrow D$  in probability. Since  $\bar{S} = \{(\mathbf{x}, \mathbf{y}; \hat{\mathbf{x}}) : \rho_n(\mathbf{x}, \mathbf{y}; \hat{\mathbf{x}}) > D + \delta\}$ , we define  $n_2$  as that for  $n > n_2$  we have  $p(\bar{S}) < \delta / \rho_{\max}$ .

We now define  $Q(\hat{\mathbf{x}}) = \prod_{i=1}^n p(\hat{x}_i)$ , where  $p(\hat{x}_i) = \sum_x \sum_y p(y) p(x|y) p(\hat{x}_i|x)$ , from which

$$i_n(\mathbf{x}; \hat{\mathbf{x}}) \equiv \frac{1}{n} \log \frac{p(\hat{\mathbf{x}}|\mathbf{x})}{p(\hat{\mathbf{x}})} = \frac{1}{n} \sum_{i=1}^n \log \frac{p(\hat{x}_i|x_i)}{p(\hat{x}_i)} = \frac{1}{n} \sum_{i=1}^n i(x_i; \hat{x}_i) \quad (20)$$

From the weak law of large numbers, for sufficiently large  $n$ ,  $\frac{1}{n} \sum_{i=1}^n i(x_i; \hat{x}_i) \rightarrow R(D)$  in probability. Hence, recognizing that  $\bar{T} = \{(\mathbf{x}, \hat{\mathbf{x}}) : i_n(\mathbf{x}; \hat{\mathbf{x}}) > R(D) + \delta\}$ , there is an integer  $n_3$  such that for  $n > n_3$  we have  $p(\bar{T}) < \delta / \rho_{\max}$ .

Therefore, for  $n > \max(n_1, n_2, n_3)$  and using (19), (17) reduces to  $\tilde{\rho} \leq D + 4\delta$ . Letting  $\delta = \varepsilon / 4$ , we have  $\tilde{\rho} \leq D + \varepsilon$  and  $K = \ll \exp[n(R(D) + \varepsilon / 2)] \gg$ , from which  $\frac{1}{n} \log K < R(D) + \varepsilon$ . Since this result is based on an average across codebooks  $B$ , there must be at least one code that satisfies the conditions in the theorem, and therefore the generalized source-coding theorem is proven.

### C. Converse to the source coding theorem

The proof presented by Berger [5] for the converse to the source coding theorem is independent of the particular distortion measure chosen, and therefore his proof is applicable

here (and not repeated). Hence, from the generalized source coding theorem in Sec. IIB and its converse we demonstrate that the  $R(D)$  result still holds for the composite distortion measure considered here: for all  $D > 0$ ,  $\lim_{n \rightarrow \infty} \frac{1}{n} \log K(n, D) = R(D)$ , with  $R(D)$  defined in (6).

### III. Generalized Blahut-Arimoto Algorithm

#### A. Summary of Blahut-Arimoto algorithm

The definition of  $R(D)$  in (6) may also be expressed as [6,10]

$$R(D) = \min_{p(\hat{x})} \min_{p(\hat{x}|x): \sum_{x, \hat{x}} p(x)p(\hat{x}|x)d'(x, \hat{x}) \leq D} \sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x) \log \frac{p(\hat{x}|x)}{p(\hat{x})} \quad (21)$$

where  $\rho'(x, \hat{x}) = \sum_y p(y|x)\rho(x, y; \hat{x})$ , which yields an iterative solution whereby

$\sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x) \log \frac{p(\hat{x}|x)}{p(\hat{x})}$  is alternatively minimized (i) in terms of  $p(\hat{x})$  with  $p(\hat{x}|x)$  fixed

and (ii) in terms of  $p(\hat{x}|x)$  with  $p(\hat{x})$  fixed. This sequence of minimizations constitutes the Blahut-Arimoto algorithm, which has been demonstrated to converge to  $R(D)$ . With regard to iteration (i), the solution for  $p(\hat{x})$  for fixed  $p(\hat{x}|x)$  is  $p(\hat{x}) = \sum_x p(x)p(\hat{x}|x)$  [6,10]. We now

consider (ii), which is a constrained minimization problem, solved using Lagrange multipliers.

Assuming fixed  $p(\hat{x})$ , we must minimize the functional

$$J = \sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x) \log \frac{p(\hat{x}|x)}{\sum_x p(x)p(\hat{x}|x)} + \gamma \sum_y \sum_x \sum_{\hat{x}} p(y)p(x|y)p(\hat{x}|x)\rho(x, y; \hat{x}) + \sum_x \nu(x) \sum_{\hat{x}} p(\hat{x}|x) \quad (22)$$

where  $\gamma$  and  $\nu(x)$  are Lagrange multipliers (see [6, p. 362]). We differentiate (22) with respect to  $p(\hat{x}_i|x_j)$ , corresponding to particular values of  $(\hat{x}, x)$  equal to  $(\hat{x}_i, x_j)$ . Setting the result equal to zero we have

$$\left\{ p(x_j) \log \frac{p(\hat{x}_i|x_j)}{p(\hat{x}_i)} + \gamma p(x_j) \rho'(x_j, \hat{x}_i) + v(x_j) \right\} + \gamma E_{x\hat{x}} \left\{ \frac{\partial \rho'(x, \hat{x})}{\partial p(\hat{x}_i|x_j)} \right\} = 0 \quad (23)$$

where  $E_{x\hat{x}} \left\{ \frac{\partial \rho'(x, \hat{x})}{\partial p(\hat{x}_i|x_j)} \right\} = \sum_x \sum_{\hat{x}} p(x) p(\hat{x}|x) \frac{\partial \rho'(x, \hat{x})}{\partial p(\hat{x}_i|x_j)}$  is the expectation of the derivative of the distortion with respect to  $p(\hat{x}_i|x_j)$ . Setting the terms within the left brackets of (22) to zero, we have [6,10]

$$p(\hat{x}|x) = p(\hat{x}) e^{-\gamma \rho'(x, \hat{x})} / \sum_{\hat{x}} p(\hat{x}) e^{-\gamma \rho'(x, \hat{x})} \quad (24)$$

where this is valid for all  $(\hat{x}, x)$ ,  $(\hat{x}_i, x_j)$  a special case.

The distortion  $\rho'(x, \hat{x})$  is explicitly

$$\rho'(x, \hat{x}) = (x - \hat{x})^2 + \lambda \sum_y h[y, \hat{y}(\hat{x})] p(y|x) = (x - \hat{x})^2 + \lambda P_e(x, \hat{x}) \quad (25)$$

where  $\hat{y}(\hat{x})$  is a lookup table that maps a given  $\hat{x}$  to  $\hat{y}$ , as defined through (3), and  $\sum_y h[y, \hat{y}(\hat{x})] p(y|x) = \sum_{y \neq \hat{y}(\hat{x})} p(y|x) = P_e(x, \hat{x})$ , where  $P_e(x, \hat{x})$  represents the probability of classification error when source output  $x$  is mapped (coded) to  $\hat{x}$ . Since  $(x - \hat{x})^2$  is independent of  $p(\hat{x}|x)$ , in the context of  $E_{x\hat{x}} \left\{ \frac{\partial \rho'(x, \hat{x})}{\partial p(\hat{x}_i|x_j)} \right\}$  we focus on the second term in (25). We therefore

have

$$E_{x\hat{x}} \left\{ \frac{\partial \rho'(x, \hat{x})}{\partial p(\hat{x}_i|x_j)} \right\} = \sum_{\hat{x}} \sum_x p(x) p(\hat{x}|x) \frac{\partial P_e(x, \hat{x})}{\partial p(\hat{x}_i|x_j)} \quad (26)$$

Let  $\tilde{p}(\hat{x}|x)$  represent the conditional density function used in (3) to define the Bayes classifier, yielding the mapping  $\hat{x} \rightarrow \hat{y}$ . Equation (26) may now be expressed as

$$E_{x\hat{x}} \left\{ \frac{\partial \rho'(x, \hat{x})}{\partial p(\hat{x}_i|x_j)} \right\} = \frac{\partial}{\partial \tilde{p}(\hat{x}_i|x_j)} \sum_{\hat{x}} \sum_x p(x) p(\hat{x}|x) P_e[x, \hat{x}; \tilde{p}(\hat{x}|x)] = \frac{\partial}{\partial \tilde{p}(\hat{x}_i|x_j)} P_{e,av}[\tilde{p}(\hat{x}|x)] \quad (27)$$

where the average probability of error is  $P_{e,av} = \sum_{\hat{x}} \sum_x p(x)p(\hat{x}|x)P_e[x,\hat{x};\tilde{p}(\hat{x}|x)]$ . When evaluated at  $\tilde{p}(\hat{x}|x) = p(\hat{x}|x)$ ,  $\frac{\partial}{\partial \tilde{p}(\hat{x}_i|x_j)} P_{e,av}[\tilde{p}(\hat{x}|x)] = 0$ , since  $\tilde{p}(\hat{x}|x) = p(\hat{x}|x)$  corresponds to the Bayes optimal classifier, for which  $P_{e,av}$  is minimized. This demonstrates that the equality in (23) is achieved when  $p(\hat{x}|x)$  satisfies (24), and when the mapping  $\hat{x} \rightarrow \hat{y}$  corresponds to the Bayes optimal classifier associated with  $p(\hat{x}|x)$ .

Note that  $p(\hat{x}|x)$  in (24) is a function of  $\rho'(x,\hat{x})$ . In addition,  $\rho'(x,\hat{x})$  is a function of  $p(\hat{x}|x)$ , via the Bayes mapping  $\hat{x} \rightarrow \hat{y}$  defined through (3). Therefore, (24) does not yield an explicit solution for  $p(\hat{x}|x)$ , although it does suggest an iterative solution. At  $k=0$  we initialize  $p^k(\hat{x}|x)$ , yielding an initial Bayes mapping  $\hat{y}^k(\hat{x})$ , from which  $\rho'^k(x,\hat{x})$  is defined. Using this  $\rho'^k(x,\hat{x})$ , we update the conditional density function, obtaining  $p^{k+1}(\hat{x}|x)$  and subsequently  $\rho'^{k+1}(x,\hat{x})$ . This sequential updating of  $p^k(\hat{x}|x)$  and  $\rho'^k(x,\hat{x})$  is repeated until convergence is achieved, yielding the solution to iteration (ii) of the Blahut-Arimoto algorithm. We discuss convergence below.

## B. Modified Blahut-Arimoto algorithm

We may summarize the solution to (21) in a modified form of the Blahut-Arimoto algorithm. We first initialize  $p_t(\hat{x})$  for iteration  $t=0$ . With  $p_0(\hat{x})$  fixed, we now minimize (21) with respect to  $p_t(\hat{x}|x)$ . As discussed above this involves an iterative solution between  $p_t(\hat{x}|x)$  and  $\rho'_t(x,\hat{x})$ . Using notation from the previous section, we iterate over index  $k$  until convergence is achieved for  $p_t^k(\hat{x}|x)$  and  $\rho_t'^k(x,\hat{x})$ , yielding  $p_t(\hat{x}|x)$ . Now, from step (i) of the algorithm  $p_{t+1}(\hat{x}) = \sum_x p_t(\hat{x}|x)p(x)$ . With  $p_{t+1}(\hat{x})$  known,  $p_{t+1}(\hat{x}|x)$  is updated (step (ii)), again

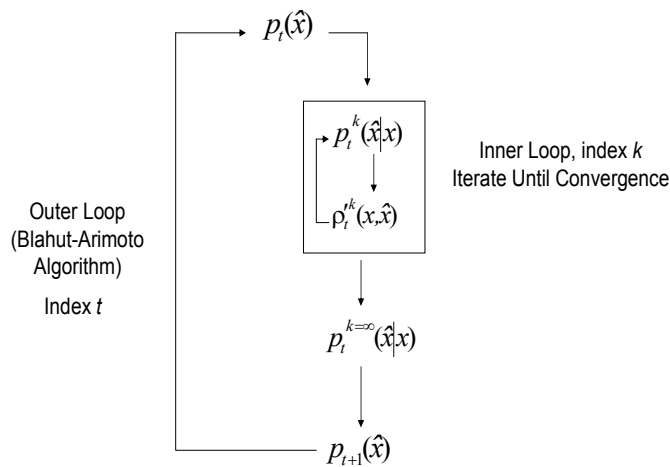
performing iterations in index  $k$  until convergence is achieved for  $p_{t+1}^k(\hat{x}|x)$ . We iterate between  $p_t(\hat{x})$  and  $p_t(\hat{x}|x)$  until convergence is achieved (see Fig. 2).

The “outer” index  $t$  is as in the original Blahut-Arimoto algorithm, alternating between the two minimizations in (21). The “inner” index  $k$  controls the iterative solution for  $p_t(\hat{x}|x)$ , required for the minimization of (22). The outer iterations are known to converge to  $R(D)$ , as in the traditional Blahut-Arimoto algorithm [10,11]. We now address convergence of the “inner” iterations, in index  $k$  between  $p_t^k(\hat{x}|x)$  and  $\rho_t^k(x, \hat{x})$ .

As discussed by Blahut [10] and Csiszar [12], the rate-distortion computation may be represented as an iterative search for  $p(\hat{x}|x)$  and  $p(\hat{x})$ , defined by  $J_\gamma = \inf_{p(\hat{x}|x), p(\hat{x})} J_\gamma[p(\hat{x}|x), p(\hat{x})]$  with

$$J_\gamma[p(\hat{x}|x), p(\hat{x})] = \sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x) \log \frac{p(\hat{x}|x)}{\sum_x p(x)p(\hat{x}|x)} + \gamma \sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x) \rho'(x, \hat{x}) \quad (28)$$

The expression in (28) is a simplification of (22), yielding identical results for  $R(D)$ . The function  $J_\gamma$  is computed by alternately minimizing (28) with respect to  $p(\hat{x}|x)$  and  $p(\hat{x})$ . For the



**Figure 2.** Summary of modified Blahut-Arimoto algorithm.

distortion measure introduced here, minimization of (28) with respect to  $p(\hat{x}|x)$  cannot be performed explicitly, and as discussed above its solution involves a separate alternating minimization of (28) with respect to  $p(\hat{x}|x)$  and  $\rho'(x, \hat{x})$ . Upon inspection of (28), for fixed  $p(\hat{x})$ , each sequential minimization of (28) with respect to  $p(\hat{x}|x)$  and  $\rho'(x, \hat{x})$  minimizations the overall value of  $J_\lambda$ . For example,

minimization of (28) with respect to  $\rho^k(x, \hat{x})$ , which involves updating the Bayes mapping  $\hat{x} \rightarrow \hat{y}$  in terms of  $p^{k-1}(\hat{x}|x)$ , reduces the average distortion reflected in the second term to the right of the equality in (28). With  $\rho^k(x, \hat{x})$  fixed, minimization of (28) with respect to  $p^k(\hat{x}|x)$  yields (24). Hence consecutive iterations in  $k$  between  $p_t^k(\hat{x}|x)$  and  $\rho_t^k(x, \hat{x})$  sequentially reduce (28), which is bounded below, and therefore these iterations must converge, yielding  $p_t(\hat{x}|x) = p_t^{k=\infty}(\hat{x}|x)$ . This and the known convergence of the iterations in  $t$  between  $p_t(\hat{x})$  and  $p_t(\hat{x}|x)$  [12] proves overall convergence of the algorithm discussed in the previous subsection. Finally, since the overall algorithm converges, and due to the known global joint minima for  $p_t(\hat{x})$  and  $p_t(\hat{x}|x)$  [12], we have assurance of convergence to the function  $R(D)$ .

### C. Sources with memory

To simplify the above discussion we have assumed that the composite source  $p(x, y) = p(y)p(x|y)$  is iid, and therefore memoryless. There are many sources with memory for which  $x$  is observable and  $y$  is “hidden” (unobservable), and one is interested in encoding  $x$  to  $\hat{x}$  with low squared error  $(x - \hat{x})^2$ , with the additional goal of estimating  $y$  (an estimate of which is denoted  $\hat{y}$ ). For example, the composite source  $p(x, y) = p(y)p(x|y)$  may correspond to a hidden Markov model (HMM), in which  $y$  corresponds to the “hidden” HMM states and  $x$  to the state-dependent observables. In this case the sequence  $p(y_1, y_2, \dots, y_n)$  is characterized by a Markov source, which has memory. It has been demonstrated that the definition of  $R(D)$  may be readily extended to the case of sources with memory [5,6], as

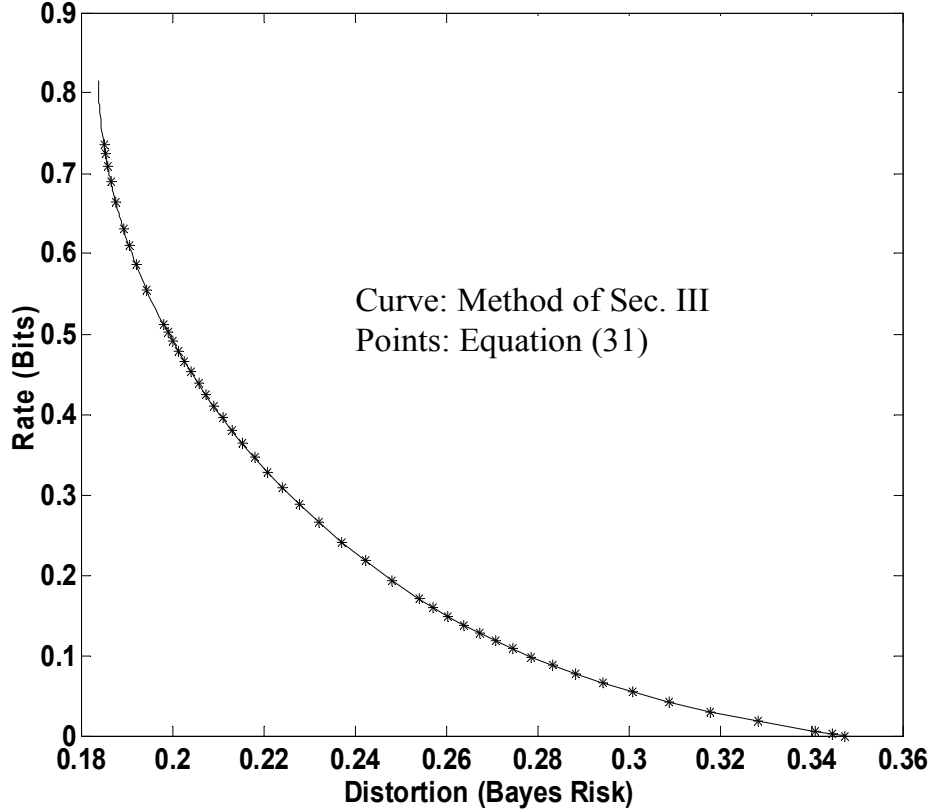
$$R(D) = \lim_{n \rightarrow \infty} R_n(D) \quad (29)$$

where

$$R_n(D) = \frac{1}{n} \min_{p(\hat{x}|x) \in Q_D} I(x; \hat{x}) \quad (30a)$$

$$Q_D = \{p(\hat{x}|x) : \sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x)\rho_n'(x, \hat{x}) \leq D\} \quad (30b)$$

$$\rho_n(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_{t=1}^n \rho'(x_t, \hat{x}_t) \quad (30c)$$



**Figure 3.** Rate-distortion curves, computed two ways, for an HMM source and Bayes-risk distortion.

where  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$  and  $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n\}$ . Therefore, the theory developed here is applicable to HMM sources, which have found a wide range of applications, including speech processing, target tracking and target classification [4,13]. As demonstrated when presenting examples (Sec. IV), the  $R(D)$  theory developed here allows one to address quantization of a sequence of observed data  $\mathbf{x}$ , while addressing reconstruction quality and accuracy in estimating the underlying HMM states. The parameter  $\lambda$  in (1) controls the relative importance placed on these two goals.

## IV. Example Results

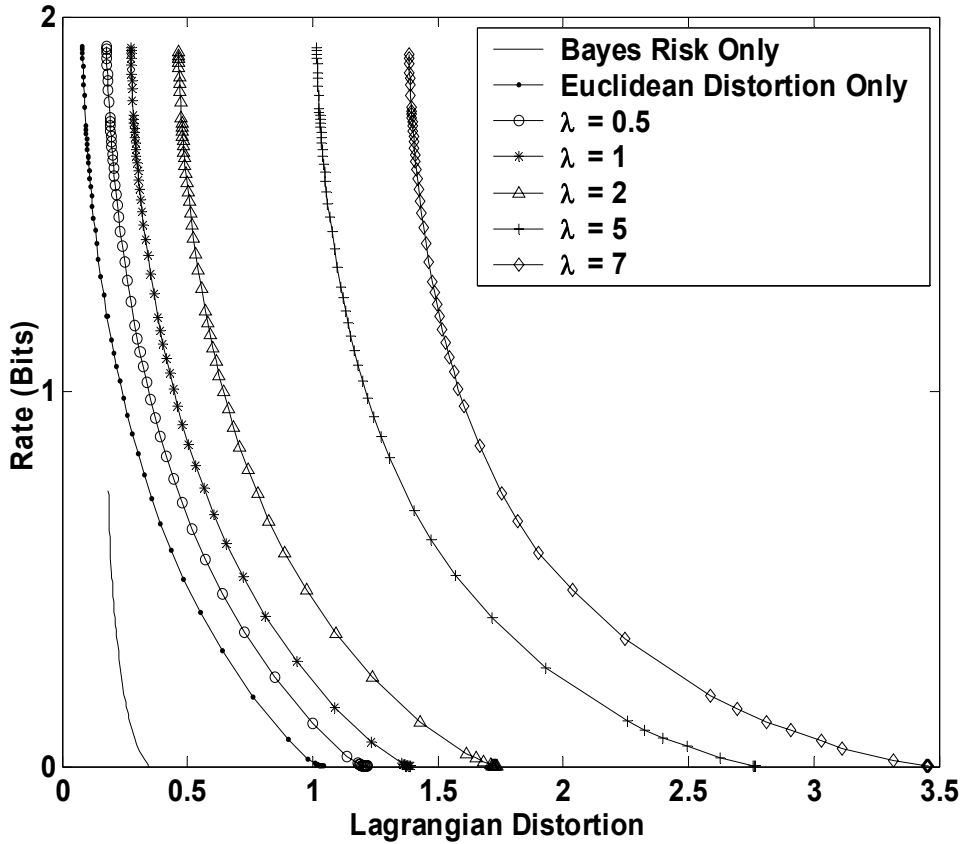
As a simple example we consider a two-state hidden Markov (HMM) source. The state-transition matrix is  $A = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}$ , where  $A_{ij}$  represents the probability of transitions from state  $i$  to state  $j$ . The initial state distribution probabilities are  $\pi = [0.2 \quad 0.8]$ , where  $\pi_i$  represents the probability that the first observation is in state  $i$ . For each state  $s$ , the associated continuous observation  $o$  satisfies a Gaussian distribution:  $P(o|s=1) \sim N(3, 1)$ ,  $P(o|s=2) \sim N(5, 2)$ . We have only proven in Sec. III the case of sources with a discrete alphabet, and in these examples we present results for finely quantized realizations of these observables. Berger [5] has also discussed extension of  $R(D)$  computations to continuous alphabets. In the notation of Sec. II, the states  $s$  correspond to the source  $y$  (binary alphabet), and the continuous observations  $o$  correspond to  $x$ . For a given rate  $R$ , the goal is to minimize the average squared Lagrangian distortion, characterized by the squared Euclidian distance between  $o$  and its reconstruction, and the Bayes error in estimating the underlying HMM states.

Ten thousand sequences of data were synthesized from the source distribution, and they were used by the modified Blahut-Arimoto algorithm to calculate the rate-distortion curve. Since this is a correlated source, the rate-distortion curve at different sequence lengths is calculated, and as the sequence length  $n$  increases  $R(D)$  converges as expected [5]. For this source we found approximate convergence for  $n \geq 4$ .

In our first example, we assume that the distortion is only characterized by Bayes risk, since this allows a comparison between the  $R(D)$  computation methods discussed in Sec. III with a simpler form of the distortion which is amenable to conventional Blahut-Arimoto computation. This corresponds to the limit  $\lambda \rightarrow \infty$  in (1). Specifically, in the context of remote sources, Berger [5] has introduced the distortion measure

$$\rho(x, \hat{y}) = \frac{1}{p(x)} \sum_y p(y) p(x|y) h(y, \hat{y}) \quad (31)$$

The result is shown in Fig. 3, and an observation sequence of length four ( $n=4$ ) is used. It is observed that the two distinct methods of computing  $R(D)$ , for the Bayes component alone, yield very similar results.



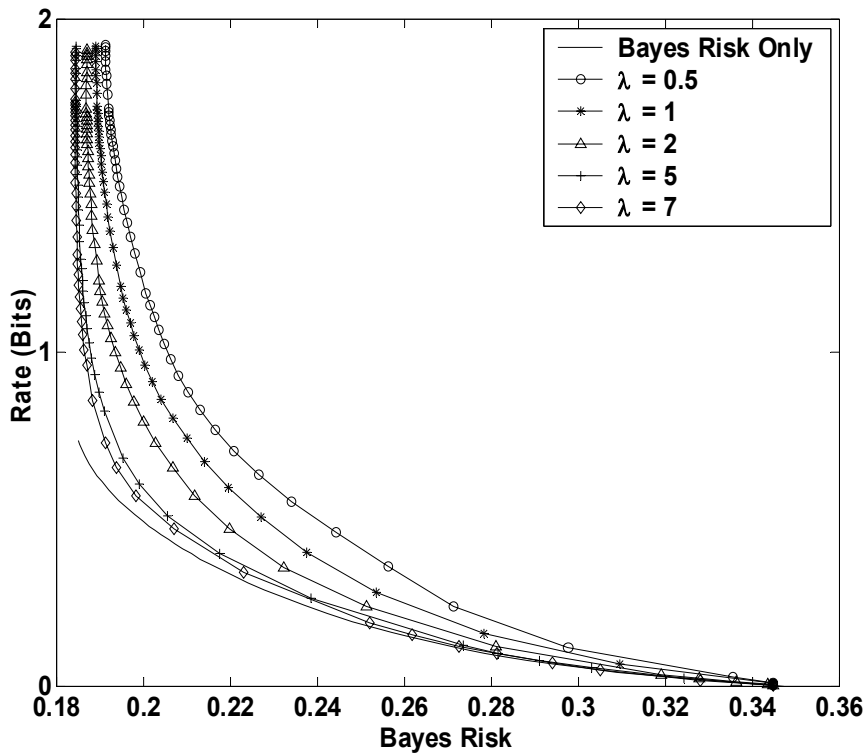
**Figure 4.** Rate-distortion curves, for Lagrangian distortion, plotted as a function of Lagrange multiplier. The limiting cases of pure squared Euclidean and pure Bayes risk distortion are also depicted.

The rate-distortion curve using the Lagrangian form of distortion is calculated for different choices of  $\lambda$ , as plotted in Fig. 4. Here a sequence length of four is used. The curve for Bayes risk only and Euclidean distortion only are also plotted for comparison. We observe that with an increase in  $\lambda$ , the Lagrangian distortion increases for a given rate  $R$ . In Fig. 5 we only plot the Bayes risk component of the Lagrangian distortion, as a function of the Lagrange multiplier  $\lambda$ , and as expected we observe that the associated  $R(D)$  approaches the results from Fig. 3, for

Bayes-risk distortion alone. A similar phenomenon happens with regard to the Euclidian component of the Lagrangian distortion, when  $\lambda$  diminishes toward zero.

## V. Conclusions

A rate-distortion framework has been developed for the joint compression and classification problem. The Lagrangian distortion measure addresses both the signal-reconstruction and classification error. It has been demonstrated that such a distortion measure is



**Figure 5.** Rate-distortion curves as in Fig. 4, but here we only plot the Bayes component of the Lagrangian distortion. Also shown is  $R(D)$  for the case of a purely Bayes-risk distortion.

appropriate for the Source Coding Theorem, and an iterative algorithm is developed based on an alternating minimization procedure, representing a generalization of the classic Blahut-Arimoto algorithm. As an example the analysis has been applied to an HMM source, wherein the goal is to encode the continuous source outputs and, based on the quantized source outputs, estimate the underlying HMM state sequence.

It should also be noted that there have been other attempts to consider bounds on encoding, accounting for both classification and reconstruction. In particular the Information Bottleneck (IB) [14] is related to the approach considered here. In the IB the mutual information used to define the rate is augmented in a Lagrangian formulation to account for classification accuracy. By contrast, in the approach presented here the Lagrangian formulation is employed on the distortion and the mutual information used to define the rate is left unchanged. The IB does not explicitly yield a coding rate and it suffers from multiple minima (non-global convergence).

### References

- [1] K. O. Perlmutter, S. M. Perlmutter, R. M. Gray, R. A. Olshen and K. L. Oehler, "Bayes risk weighted vector quantization with posterior estimation for image compression and Classification," *IEEE Trans. Image Proc.*, vol. 5, No. 2, pp. 347-360, Feb. 1996.
- [2] K. L. Oehler and R. M. Gray, "Combining image compression and classification using vector quantization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, No. 5, pp. 461-473, May 1995.
- [3] J. D. Gorman, "Vector quantizer designs for joint compression and terrain categorization of multispectral imagery," *Pro. NASA Conf. 3255, 1994 Space Earth Sci. Data Compression Workshop*, Salt Lake City, UT, Apr. 1994.
- [4] P. Runkle, P. Bharadwaj, and L. Carin, "Hidden Markov model for multi-aspect target classification," *IEEE Trans. Signal Proc.*, vol. 47, No. 7, pp. 2035-2040, July 1999.
- [5] T. Berger, *Rate Distortion Theory*, Prentice-Hall, 1971.
- [6] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.
- [7] J. Principe, D. Xu and J. Fisher III, "Pose estimation in SAR using an information-theoretic criterion," *Proc. SPIE*, vol. 3370, pp. 218-229, 1998.
- [8] P.A. Chou, *Application of Information Theory to Pattern Recognition and the Design of Decision Trees and Trellises*, Ph. D. dissertation, Stanford University, 1988.
- [9] S.D. Morgera, and M.R. Soleymanl, "A structural look at pattern recognition from the point of view of rate-distortion theory," *Pattern Recognition and Artificial Intelligence*, E.S. Gelsema and L.N. Kanal (Editors), Elsevier Science, 1988, pp. 257-275.
- [10] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, No. 4, pp. 460-473, Jul. 1972.
- [11] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 357-359, Jan. 1973.
- [12] I. Csiszar, "On the computation of rate-distortion functions," *IEEE Trans. Information Theory*, vol. 20, pp. 122-124, Jan. 1974.

- [13] L. R. Rabiner, "A tutorial on hidden Markov model and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, No. 2, pp. 257-286, Feb. 1989.
- [14] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," *Proc. of the 37th annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 22 - 24, 1999.