Multiple Description Quantization by Deterministic Annealing

Prashant Koulgi, Shankar L. Regunathan and Kenneth Rose
Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106

September 21, 1999

Abstract

The design of vector quantizers for diversity-based communication over two or more channels of possibly differing capacities and failure probabilities, is considered. The standard iterative design method is highly dependent on initialization, especially of index assignment. A "good" initialization is known for the special case of scalar quantization of symmetric sources over identical channels, but does not generalize well to vectors, unbalanced channels or possibly asymmetric sources. Instead, we propose to pursue a deterministic annealing approach which is independent of initialization, does not assume any prior knowledge of the source density and avoids many poor local minima of the cost surface. The approach consists of iterative optimization of a random encoder at gradually decreasing levels of randomness as measured by the Shannon entropy. At the limit of zero entropy, a hard multiple description quantizer is obtained. This process is directly analogous to annealing processes in physical chemistry. The Lagrangian cost functional we use corresponds to the Helmholtz free energy of the system, and the system undergoes phase transitions (where the cardinality of the effective reproduction alphabet increases) whose critical "temperatures", or values of Lagrange multiplier, are computable. To illustrate the potential of our approach, we present simulation results that show substantial performance gains over existing design techniques. Finally, we include a brief note discussing some preliminary work on the relation between the proposed DA approach and the computation of an achievable region for the multiple description problem.

1 Introduction

We consider the design of Multiple Description Vector Quantizers (MDVQ) for use in a diversity-based communication system (hereafter referred to as a diversity system). A diversity system provides several channels for communication between the transmitter and the receiver. The MDVQ encoder encodes a fixed length block of source samples into individual indices for transmission over each of the channels, subject to separate rate constraints. Each of these channels may fail independently, and the decoder reconstruction is based on information received from the subset of channels that are in working order. Applications of multiple description source codes are currently being pursued in speech and video coding over packet-switched networks and fading multipath channels [1]. MDVQs designed for asymmetric channels are strongly motivated by packet-switched networks with priority classes. Finally, scalable quantizer design may be viewed as a special case of MDVQ design.

For simplicity, we will restrict the discussion to diversity systems with two channels. The two channels may have differing capacities and failure probabilities. (We call the special case when the two channels allow the same rates and have identical failure probabilities the case of *balanced descriptions*). When both the channels function reliably, the distortion achieved with the joint description is the "central" distortion; when one of the channels fails, the distortion achieved with the received single side description is the corresponding "side" distortion.

The general problem of jointly good descriptions for a diversity system with two channels was posed by Gersho, Ozarow, Witsenhausen, Wolf, Wyner and Ziv at the 1979 Information Theory Workshop. In [6], El Gamal and Cover constructed an achievable rate region for the multiple descriptions problem for a memoryless source and a single-letter fidelity criterion. Ozarow [7] showed that this region is, in fact, the rate distortion region for the memoryless Gaussian source under the squared-error distortion criterion. Contributions to this problem can also be found in Witsenhausen [8], Wolf, Wyner and Ziv [9], Witsenhausen and Wyner [10], Berger and Zhang [11], [12], Ahlswede [13], and Equitz and Cover [14].

The design of multiple description scalar quantizers (MDSQs) was studied by Vaishampayan in [1] whose work generated new interest in the practical applications of multiple description coding. He derived an iterative design algorithm (closely related to Lloyd's algorithm for quantizer design [3]) that minimizes the overall average distortion cost which is a weighted sum of the average central distortion and the two side distortions. The weights for this sum are determined by the channel failure probabilities , and the sizes allowed for the code books are fixed by the channel capacities. The algorithm is guaranteed to find a locally optimal solution. An extension of this algorithm for the design of vector quantizers has been proposed in [4]. The problem of good initializations for these algorithms has not been entirely solved.

The main contribution of this paper is an algorithm to design unstructured MDVQs which minimizes the overall average distortion cost. This optimization problem is non-convex, and the distortion cost surface is riddled with local minima. The problem of local minima is greatly exacerbated by the dependence of the average distortion cost on index assignment [1]. Existing iterative design methods [1], [4] that monotonically decrease the average distortion cost are likely to be trapped in poor local minima, unless "good" initial code books and initial index assignment are used. To the best of our knowledge, heuristic initializations for index assignment have only been proposed for the special cases: balanced description MDSQs [1], and balanced description lattice MDVQs [5]. These initializations do not generalize to the case of general MDVQs, nor for unbalanced descriptions, nor for arbitrary source distributions. In contrast, the algorithm we propose does not require initialization of code books or index assignment, and avoids many local minima of the cost surface. Further, no knowledge of the underlying source distribution is needed. Our approach is inspired by, and builds on the Deterministic Annealing (DA) approach for vector quantizer design [2]. It is motivated by the observation of annealing processes in physical chemistry.

Certain chemical systems can be driven to their low energy states by annealing, which is a gradual reduction of temperature, spending a long time in the vicinity of phase transition points. Analogously, we randomize the encoding rule of the multiple description system and seek to minimize the expected distortion cost subject to a specified level of randomness measured by the Shannon entropy. This problem can be formulated as the minimization of a Lagrangian functional that is analogous to the Helmholtz free energy of chemical systems. The degree of randomness is parameterized by the "temperature" of the configuration. We start at a high degree of randomness, where we, in fact, maximize the entropy. Here, the globally minimum configuration requires that all code vectors be coincident at the centroid of the source distribution; no initialization of code book or index assignment is necessary. We then track the minimum at successively lower levels of entropy, by re-calculating the optimum locations of the reproduction points and the encoding probabilities at each stage. At the limit of zero randomness, the algorithm directly minimizes the average distortion cost, and a deterministic encoder is obtained.

We formulate the problem of MDVQ design and establish notation in the next section. In section 3 we describe the DA approach to this problem. Necessary conditions for optimality are then used to derive an iterative MDVQ design algorithm. We conclude the section by describing the "mass-constrained" form of our algorithm, which is our preferred implementation. In section 4 we present simulation results and comparisons with existing approaches. We show that DA performs consistently better than existing approaches by avoiding many poor local minima. We are currently investigating the links between the mass-constrained DA algorithm for MDVQ design and the problem of computing the convex hull of an achievable region for multiple descriptions. We sketch some preliminary observations in section 5. Phase transition analysis, including derivation of critical temperatures at which the size of the reproduction set increases, is considered in the appendix.

2 The MDVQ problem and design considerations

We are interested in encoding a source represented by a stationary and ergodic random process $\{X_k\}$. Consider a diversity system with two channels capable of transmission of information at rates R_1 and R_2 bits/source sample, respectively. Each channel may or may not be in working order, and its condition is not known at the encoder. The encoder sends a different description over each channel. The decoder forms the best estimate of the source output from the descriptions received via the channels that were functioning reliably. An MDVQ maps an n-dimensional source vector x to the n-dimensional reproduction vectors $\hat{x}^0$,

$\hat{x}^1$ and $\hat{x}^2$, which take values in the code-books $\hat{\mathcal{X}}^0 = \{\hat{x}^0_{ij}, (i,j) \in I_1 \times I_2\}$, $\hat{\mathcal{X}}^1 = \{\hat{x}^1_i, i \in I_1\}$ and $\hat{\mathcal{X}}^2 = \{\hat{x}^2_j, j \in I_2\}$, respectively. Here, $I_1 = \{1, 2, \ldots, 2^{nR_1}\}$ and $I_2 = \{1, 2, \ldots, 2^{nR_2}\}$.

The MDVQ encoder is the mapping $f : \mathcal{R}^n \mapsto I_1 \times I_2$. Given source vector x, it selects an index pair: $f(x) = (i,j)$. Each index is sent over its respective channel. The MDVQ decoder $g = (g_0, g_1, g_2)$ is, in fact, a bank of three switched decoders each performing a look-up operation: the central decoder $g_0 : I_1 \times I_2 \mapsto \mathcal{R}^n$ takes in a double index (i,j) and produces the code vector $g_0(i,j) = \hat{x}^0_{ij}$. The side decoders $g_1 : I_1 \mapsto \mathcal{R}^n$ and $g_2 : I_2 \mapsto \mathcal{R}^n$ take in the single indices i and j to produce the code vectors $g_1(i) = \hat{x}^1_i$ and $g_2(j) = \hat{x}^2_j$ respectively.

Let $\hat{x}^m(x)$, $m = 0, 1, 2$ be the central and side reproductions chosen by the quantizer when presented the input vector x_i, and let $d_m = d(x, \hat{x}^m(x))$ be the corresponding per-vector distortion. If random vector X represents the source, and random vectors $\hat{X}^m, m = 0, 1, 2$ represent the decoder outputs, the expected central and side distortions are given by

$$E\{d(X, \hat{X}^m)\} \approx (1/N) \sum_x \{d(x, \hat{x}^m(x))\} \quad m = 0, 1, 2, \tag{1}$$

where we assume that the source distribution may be approximated by a training set of N vectors.

For given values of R_1, R_2, D_1 and D_2 we wish to find an MDVQ which minimizes

$$D(f, g) = E\{d(X, \hat{X}^0)\} + \lambda_1 E\{d(X, \hat{X}^1)\} + \lambda_2 E\{d(X, \hat{X}^2)\}, \tag{2}$$

over f and g. The specific choice of λ_1 and λ_2 in a practical system is determined by the weights we wish to assign to the side distortions relative to the central distortion, which could depend on the channel failure probabilities. In the rest of this paper we shall confine our attention to squared-error distortion, i.e., $d(x, \hat{x}^m(x)) = ||x - \hat{x}^m(x)||^2$, $m = 0, 1, 2$.

Note that the expected distortion cost depends on the code vector locations. Further, it also depends on the indices assigned to codevectors since they determine which pair of side vectors are mapped to each central code vector. Locally optimal multiple description quantizer design algorithms [1], [4] must be initialized with code books and index assignment. The choice of initial index assignment constrains the algorithm to a part of the cost surface, so that "good" initial index assignment is crucial to the performance of the algorithm. But good heuristics for choosing the initial index assignment are elusive, since they depend on the particular rate and distortion constraints, and some knowledge of the source distribution. Strategies for initial index assignment are discussed in detail in [1], where good heuristics are presented for the special case of balanced descriptions and scalar quantizer. Moreover, it was shown that heuristic index assignment is asymptotically optimal (in the sense of high resolution) for the particular case of memoryless Gaussian source and balanced descriptions if (a) the quantizer is scalar [1], or (b) if the quantizer has a lattice structure [5]. These heuristics cannot be extended to the case of unstructured MDVQs nor for unbalanced descriptions.

3 Derivation of the DA algorithm

3.1 General description

We have seen that the problem of finding good code books and index assignments to initialize existing multiple description quantizer design algorithms, is hard and generally unresolved. We take a different route, and propose an algorithm which does not depend on an initial set of code books or an initial index assignment.

Our algorithm is based on DA, which is motivated by the observation of annealing processes in physical chemistry. Here, certain chemical systems are driven to their low energy states by a gradual reduction of temperature, spending a long time in the vicinity of the phase transition points. A formal derivation of DA can be based on principles of information theory or statistical physics. One approach optimizes a random encoder subject to constraints on its degree of randomness. An alternative derivation appeals to Jaynes's principle of maximum entropy for statistical inference [15]. We shall use the former approach here. Both derivations were shown to be equivalent in [17], which includes a tutorial overview of the wide applicability of DA to general optimization problems. Of particular interest here is the DA approach to computation of the rate-distortion function of continuous sources [16].

We consider a multiple description coding system where the deterministic encoder is replaced by a random encoding rule. In other words, an input vector is assigned to each possible pair of indices in probability. The probabilities are determined by finding the distribution that minimizes the expected distortion cost of (2) subject to a specified level of randomness. The level of randomness is, naturally, measured by the Shannon conditional entropy of the encoding probability distribution. This problem can be cast as the unconstrained minimization of a Lagrangian functional where the Lagrangian multiplier controls the degree of randomness of the configuration. The Lagrangian functional plays the role of the free energy, while the Lagrange multiplier is the temperature in the analogous annealing process of physical chemistry. We shall henceforth refer to them by these names.

The process of annealing consists of starting at a high temperature and gradually lowering the temperature. The free energy is minimized at each temperature to find the encoding probabilities and the optimal locations for the reproductions. In the physical analogy, it means reaching isothermal equilibrium. At very high temperatures we mainly attempt to maximize the encoding entropy. The encoding probabilities are uniformly distributed; the globally optimal central and side code books each consist of coincident code vectors at the centroid of the source distribution. As the temperature is lowered we trade entropy for reduction in expected distortion cost. There occur stages when it becomes advantageous to bifurcate sets of coincident code vectors, thus increasing the number of *distinct* reproductions (effective reproduction cardinality). These are phase transitions in our physical analogy, and the temperature at which each occurs is the corresponding "critical temperature". As the temperature approaches zero, we minimize the expected distortion cost directly. The encoding probabilities turn into hard mappings (i.e., they are either 0 or 1) as in a normal *hard quantizer*, and the index assignment arises naturally. Successive iterations of the algorithm now only reduce the distortion cost: DA becomes a standard local optimization technique at the limit of zero temperature, albeit, not with arbitrary initialization.

3.2 Encoding probabilities and reproduction points

Let us begin by assuming that the three code-books, $\hat{\mathcal{X}}^0 = \{\hat{x}^0_{ij}\}$, $\hat{\mathcal{X}}^1 = \{\hat{x}^1_i\}$ and $\hat{\mathcal{X}}^2 = \{\hat{x}^2_j\}$ are given. We use a random encoding rule, and assign input source vector x to the index pair (i, j) with probability $q(ij|x)$. The central decoder and the two side decoders output $\hat{x}^0_{ij}$, $\hat{x}^1_i$ and $\hat{x}^2_j$ when presented with indices (i, j). We can rewrite the expected distortion cost of (2) for a random encoder as

$$D = \sum_x p(x) \sum_{ij} q(ij|x)\{||x - \hat{x}^0_{ij}||^2 + \lambda_1||x - \hat{x}^1_i||^2 + \lambda_2||x - \hat{x}^2_j||^2\}, \tag{3}$$

where we drop the arguments of D for notational simplicity.

Directly minimizing D with respect to the free parameters $\{q(ij|x)\}$ would assign each input vector to the index pair minimizing $||x - \hat{x}^0_{ij}||^2 + \lambda_1||x - \hat{x}^1_i||^2 + \lambda_2||x - \hat{x}^2_j||^2$ with probability 1. However, we recast this optimization problem as that of seeking the distribution which minimizes D subject to a specified level of randomness. The level of randomness is measured by the Shannon entropy of the distribution, given as

$$H(I_1, I_2, X) = -\sum_x \sum_{ij} p(x)q(ij|x) \log p(x)q(ij|x). \tag{4}$$

The corresponding Lagrangian to minimize is

$$F = D - TH. \tag{5}$$

The Lagrangian functional, F, is analogous to the Helmholtz free energy of a physical system where D is the energy, H is the entropy and the Lagrangian multiplier, T, is the temperature. Minimizing F corresponds to seeking isothermal equilibrium of the system.

Before proceeding further, we note that the joint entropy of (4) can be decomposed into two terms: $H(I_1, I_2, X) = H(X) + H(I_1, I_2|X)$, where $H(X) = -\sum_x p(x) \log p(x)$ is the source entropy and is independent of the encoding rule. We therefore drop the constant $H(X)$ from the Lagrangian definition, and confine our attention to the conditional entropy,

$$H(I_1, I_2|X) = -\sum_x \sum_{ij} p(x)q(ij|x) \log q(ij|x). \tag{6}$$

Minimizing F with respect to the encoding probabilities $q(ij|x)$ gives,

$$q(ij|x) = \frac{\exp[-(\frac{1}{T})\{||x - \hat{x}^0_{ij}||^2 + \lambda_1||x - \hat{x}^1_i||^2 + \lambda_2||x - \hat{x}^2_j||^2\}]}{Z_x}, \tag{7}$$

where the normalizing factor is

$$Z_x = \sum_{ij} \exp[-(\frac{1}{T})\{||x - \hat{x}^0_{ij}||^2 + \lambda_1||x - \hat{x}^1_i||^2 + \lambda_2||x - \hat{x}^2_j||^2\}]. \tag{8}$$

The corresponding minimum of F is obtained by plugging (7) into (5)

$$F^* = \min_{q(ij|x)} F = -T \sum_x p(x) \log Z_x. \tag{9}$$

We now find the optimal sets of reproduction vectors $\hat{\mathcal{X}}^0$, $\hat{\mathcal{X}}^1$ and $\hat{\mathcal{X}}^2$ which minimize F^* for this random encoder. These vectors satisfy,

$$\frac{\partial}{\partial \hat{x}^0_{ij}} F^* = 0 \quad \forall \quad \hat{x}^0_{ij} \in \hat{\mathcal{X}}^0, \tag{10}$$

$$\frac{\partial}{\partial \hat{x}^1_{i}} F^* = 0 \quad \forall \quad \hat{x}^1_{i} \in \hat{\mathcal{X}}^1 \tag{11}$$

and

$$\frac{\partial}{\partial \hat{x}^2_{j}} F^* = 0 \quad \forall \quad \hat{x}^2_{j} \in \hat{\mathcal{X}}^2. \tag{12}$$

For the squared error distortion case the above equations reduce to the centroid rules:

$$\hat{x}^0_{ij} = \sum_x p(x|ij)x, \quad \hat{x}^1_i = \textstyle\sum_x p(x|i)x, \quad \hat{x}^2_j = \displaystyle\sum_x p(x|j)x, \tag{13}$$

where $p(x|ij)$, $p(x|i)$ and $p(x|j)$ denote the posterior probabilities calculated using Bayes's rule.

Our algorithm consists of minimizing F^* with respect to the code vectors starting at a high temperature and tracking the minimum while decreasing the temperature. The central iteration is composed of the following two steps:

1) fix the code books and use (7) to compute the encoding probabilities.

2) fix the encoding probabilities and optimize the code books according to (13).

Clearly, this procedure is monotone non-increasing in F^*, and converges to a minimum. At high temperature, the global minimum configuration consists of all the (central and side) code vectors coincident at the centroid of the source distribution. We then gradually decrease the temperature and track the minimum. At the limit of low temperature the encoding probabilities harden: each source vector is assigned with probability 1 to the index pair (i, j) for which the distortion cost $||x - \hat{x}^0_{ij}||^2 + \lambda_1||x - \hat{x}^1_i||^2 + \lambda_2||x - \hat{x}^2_j||^2$ is minimized. Thus the index assignment arises naturally. The analogy of this procedure with the chemical process of annealing, which consists of maintaining a chemical system at thermal equilibrium while carefully lowering the temperature to reach a final, low energy state, is evident.

As the temperature is reduced from the initial high value the set of code vectors coincident at the centroid of the source distribution bifurcates into subsets for the first time at some lower temperature. We call this bifurcation the first phase transition, and the corresponding temperature the first critical temperature, in analogy with the phase transitions seen during the annealing of physical systems. As the temperature is lowered further these subsets again bifurcate, and each such bifurcation is a subsequent phase transition with its corresponding critical temperature. We analyze conditions for phase transitions and derive the critical temperatures in the appendix. A study of this phenomenon of phase transitions provides insight into the annealing process. Further, since the phase transitions are the critical points of the process, knowledge of the critical temperatures allows us to accelerate the annealing between phase transitions.

3.3 The mass-constrained approach

The observation of the phenomenon of phase transitions enables us to recast our algorithm in a more efficient form. Since all the (central and side) code vectors are coincident at high temperatures, they can be viewed as belonging to a single cluster, and this entire cluster can effectively be represented by a single index pair without affecting the expected distortion cost. When the code books bifurcate at the critical temperatures the effective number of clusters increases. Each of these clusters should be represented by a different index pair. We use this observation to derive the "mass-constrained" implementation of our algorithm.

Let us assume an unlimited supply of code vectors and index pairs. The fraction of code vectors of the first side code book $\hat{\mathcal{X}}^1$ which are coincident at some point can be assigned a common first index for transmission over one of the channels. Let this common index be i, the corresponding fraction of code vectors is labeled $q(i)$ (the cluster "prior" or "mass"), and the point where the code vectors are coincident is $\hat{x}_i^1$. Similarly, we assign a common index j to the mass $q(j)$ of code vectors of the second side code book coincident at $\hat{x}_j^2$. Consequently, a fraction $q(i)q(j)$ of all index pairs are assigned the index pair (i, j), and the central reproduction corresponding to this index pair is $\hat{x}_{ij}^0$. We can recast the expression for the encoding probability $q(ij|x)$ in (7) as

$$q(ij|x) = \frac{q(i)q(j)\exp[-(\frac{1}{T})\{||x-\hat{x}_{ij}^0||^2 + \lambda_1||x-\hat{x}_i^1||^2 + \lambda_2||x-\hat{x}_j^2||^2\}]}{Z_x}, \tag{14}$$

where Z_x is now modified to

$$Z_x = \sum_{ij} q(i)q(j)\exp[-(\frac{1}{T})\{||x-\hat{x}_{ij}^0||^2 + \lambda_1||x-\hat{x}_i^1||^2 + \lambda_2||x-\hat{x}_j^2||^2\}], \tag{15}$$

and F^* is

$$F^* = -T\sum_x p(x)\log Z_x, \tag{16}$$

as in (9), except that Z_x is now given by (15). F^* is to be minimized under the obvious constraints on the masses: $\sum_i q(i) = 1$ and $\sum_j q(j) = 1$. Minimizing F^* with respect to the central and side reproduction points again gives the update formulae of (13), with the encoding probabilities of (14) used to calculate the posterior probabilities. The optimum masses minimizing F^* are calculated from the following update rules:

$$q(i) = \sum_{xj} p(x)q(ij|x), \quad q(j) = \textstyle\sum_{xi} p(x)q(ij|x) \tag{17}$$

In other words, the distribution of masses on the indices is identical to the probability distribution induced on the indices via the encoding rule.

The mass-constrained approach increases the effective number of index pairs only when it is needed, i.e., at a phase transition. Thus it is computationally more efficient than the earlier "unconstrained" approach. At the limit of low temperatures the two approaches converge to the same descent process for the expected distortion cost, since their encoding probabilities are identical at the limit (they assign each data point to a single index pair (i, j) with probability 1). We use the mass-constrained approach for all our simulations.

4 Simulation results

The proposed DA-based design algorithm may be used to design unstructured MDVQs with unequal rate and distortion constraints on the two channels. Note that the distortion constraints may be determined by the individual channel failure probabilities, while the rate constraints would be fixed by the individual channel capacities. We illustrate the wide applicability of the DA algorithm by considering three examples: 1) two-dimensional vector quantizer design for equal rate and distortion constraints, 2) scalar quantizer design for unequal rate and equal distortion constraints, and 3) scalar quantizer design for equal rate and unequal distortion constraints.

For comparison, we consider the performance of the existing iterative MDVQ design technique [1], [4], which we call the "Lloyd approach" (LA) as it is directly based on Lloyd's algorithm for conventional scalar quantizer [3] and its vector extension [18]. Recall that the performance of LA depends heavily on the initialization. We use twenty different *random* initializations for the LA in our simulations. The initialization proposed in [1] is for MDSQ design with equal rate and distortion constraints. In particular, this initialization does not generalize to vectors or for unequal rate constraints. However, as an additional comparison, we used this initialization for MDSQ design with equal rate but unequal distortion constraints.

In all the three examples, the quantizer designed by DA is seen to yield a significantly lower expected distortion cost than LA with random/heuristic initializations. Further, the wide variation in performance of the quantizers designed by LA illustrates and emphasizes the significance of the problem of local minima even for simple low rate quantizers.

In Figure 1, we present the results for the design of two-dimensional vector quantizers for a Gauss-Markov source with autocorrelation coefficient $\rho = 0.9$ and unit-variance per dimension. A training set of 5000 vectors was used. The rate and distortion constraints were: $R_1 = R_2 = 1.5$ bits/source sample (i.e., each side code book has eight 2-d code vectors) and $\lambda_1 = \lambda_2 = 0.01$. We compare the performance of DA design with quantizers produced by LA for twenty different random initializations. The distortion cost of the MDVQ designed by DA is ~ 0.6 dB below the distortion of the "best" quantizer produced by random initializations of LA. Note that the heuristic index assignment proposed in [1] cannot be generalized to this case.

In Figure 2, we present the results for the design of scalar quantizers with unequal distortion constraints for a unit-variance Gaussian source. The constraints were: $R_1 = R_2 = 3$ bpss and $\lambda_1 = 0.006, \lambda_2 = 0.012$. The training set consisted of 5000 samples. The quantizers produced by LA with different random initializations show wide variation in performance (the best and the worst of these designs differ by ~ 3 dB in terms of the expected distortion cost). Note that LA initialized with the heuristic proposed in [1] yields significant gains over random initialization, and demonstrates the benefits of a good heuristic. However, MDSQ designed via the proposed DA approach outperforms by ~ 0.5 dB even LA quantizer initialized with this clever heuristic.

In Figure 3, we present results for scalar quantizer design under unequal rate constraints. The constraints were: $R_1 = 3$ bpss, $R_2 = 2$ bpss and $\lambda_1 = \lambda_2 = 0.01$. The training set consisted of 5000 samples of a unit-variance Gaussian source. The DA design is compared with randomly initialized designs of LA. The DA design gains ~ 1 dB over the best of the latter in terms of the expected distortion cost. Note that the heuristic index assignment cannot be extended to this case.

5 Interesting links with R-D theory and directions for future work

In [16], a DA algorithm has been proposed to calculate the rate-distortion curve for continuous alphabet sources. It was shown that the deterministic annealing process as parameterized by the temperature directly corresponds to parametric solution of the variational equations of rate-distortion theory. Further, gradually lowering the temperature starting at a high value is equivalent to starting at zero slope and climbing up the rate-distortion curve. For continuous sources under the squared error distortion criterion, the optimal reproduction alphabet has been shown to be discrete if the Shannon lower bound to the rate-distortion curve is not tight [16]. In this case, DA can exactly compute the rate-distortion curve by avoiding discretization for numerical computation (see [16] for details). In fact, the DA algorithm for VQ design tracks the rate-distortion curve until the size of the optimal reproduction alphabet equals the (pre-specified) VQ codebook size. We are currently exploring similar links between the parametric determination of the convex hull of an achievable region for multiple descriptions and the DA algorithm for MDVQ design. We present some preliminary observations here.

The multiple descriptions problem is concerned with finding all quintuples $(R_1, R_2, D_0, D_1, D_2)$ which are achievable in a rate-distortion sense. In [6] El Gamal and Cover proposed sufficient conditions for the achievable region. These conditions superseded an earlier characterization of an achievable region:

The quintuple $(R_1, R_2, D_0, D_1, D_2)$ is achievable if there exist random variables I_1 and I_2 jointly distributed with a generic source random variable X such that

$$R_1 \geq I(X; I_1), \quad R_2 \geq I(X; I_2) \tag{18}$$

$$R_1 + R_2 \geq I(X; I_1, I_2) + I(I_1; I_2), \tag{19}$$

and there exist side and central reproductions $\hat{X}^1$, $\hat{X}^2$ and $\hat{X}^0$ which can be expressed as (deterministic) functions of I_1 and I_2 in the following way

$$\hat{X}^1 = g_1(I_1), \quad \hat{X}^2 = g_2(I_2), and \quad \hat{X}^0 = g_0(I_1, I_2), \tag{20}$$

such that

$$E\{d(X, \hat{X}^m)\} - D_m \leq 0, \quad m = 0, 1, 2, \tag{21}$$

where d is a single letter distortion measure.

Zhang and Berger in [12] and Witsenhausen in [8] attribute this characterization to El Gamal and Cover. In [12], Zhang and Berger show that the region characterized by this earlier set of conditions is a subset of the region characterized in [6]. Note that the random variables I_1 and I_2 and the functions g_1, g_2 and g_0 lend themselves to easy interpretation in terms of a coding system: realizations of I_1 and I_2 may be interpreted as the indices transmitted over the two channels, while g_0, g_1 and g_2 are respectively the central and side decoder functions. This interpretation is precise if the reproduction is discrete. This was shown to be the case for squared error distortion for the ordinary (single description) R-D function. For the moment we conjecture it to be the case for multiple descriptions also.

We investigate the problem of finding quintuples on the convex hull of this region, denoted by $\mathcal{R}$. We formulate this problem as the minimization of the central distortion, $E\{d(X, \hat{X}^0)\}$, subject to constraints on the side distortions and the rates, as given above. This is an optimization problem with inequality constraints, and we use the Kuhn-Tucker theorem [19] for its solution.

Lemma 1: If $R \geq I(X; I_1, I_2) + I(I_1; I_2)$ then there exist R_1 and R_2 s.t. $R_1 + R_2 = R$ and $R_1 \geq I(X; I_1)$, $R_2 \geq I(X; I_2)$.

Proof:

$$\begin{aligned} I(X; I_1, I_2) + I(I_1; I_2) - I(X; I_1) - I(X; I_2) &= I(X; I_2|I_1) + I(I_1; I_2) - I(X; I_2), \\ &= I(X; I_2|I_1) + H(I_2|X) - H(I_2|I_1), \\ &= H(I_2|X) - H(I_2|I_1, X) = I(I_2; I_1|X) \geq 0, \end{aligned} \tag{22}$$

by the non-negativity of mutual information. Hence, $I(X; I_1, I_2) + I(I_1; I_2) \geq I(X; I_1) + I(X; I_2)$. So, $R \geq I(X; I_1, I_2) + I(I_1; I_2)$ implies $\exists R_1, R_2$ such that $R_1 + R_2 = R$ and $R_1 \geq I(X; I_1)$, $R_2 \geq I(X; I_2)$.

This shows that we can use only the constraint on the total rate $R_1 + R_2$ (18) and the constraints on the distortions (20) to find some quintuples on the convex hull of $\mathcal{R}$. It is then always possible to partition the total rate in such a way as to satisfy the constraints on the individual rates. The individual constraints on the side rates need to be considered together only to find those points on the convex hull of $\mathcal{R}$ where the individual descriptions are on their respective rate-distortion curves. But this case is of limited interest in practice. So we formulate the problem as the unconstrained minimization of the Lagrangian functional

$$J = E\{d(X, g_0(I_1, I_2))\} + \lambda_1 E\{d(X, g_1(I_1))\} + \lambda_2 E\{d(X, g_2(I_2))\} + T(I(X; I_1, I_2) + I(I_1; I_2)), \tag{23}$$

over the joint probability density of X, I_1 and I_2 and the decoder functions g_0, g_1 and g_2. T, λ_1 and λ_2 are Lagrangian multipliers.

Minimizing J over the decoder functions simply gives the centroid rules of (13) above. The conditional probability distribution of I_1 and I_2 given X that minimizes J is of the same form as $q(ij|x)$ for the mass-constrained MDVQ design approach ((14) above). Thus the mass-constrained form of the DA algorithm for MDVQ design can be interpreted as tracking a trajectory on the boundary of $\mathcal{R}$ until the (pre-specified) code book sizes are reached. Future work includes the characterization of conditions for discretization of the reproductions, and the exploration of the potential contribution of the DA viewpoint to the determination of the achievable region.

6 Conclusion

Deterministic annealing is proposed for the design of multiple description vector quantizers when the two channels need not have identical capacities or failure probabilities. This approach eliminates the dependence on initial configuration and avoids many poor local minima of the cost surface. Further, no knowledge is assumed on the underlying probability distribution of the source. DA is motivated by analogy to statistical physics and is derived from principles of information theory. A random encoding rule is used, and the encoding probabilities are determined by minimization of the expected distortion cost at a specified level of entropy. The algorithm starts at the global minimum at high temperature and tracks the minimum while lowering the temperature. A multiple description quantizer is obtained at the limit of low temperature. We compared our approach with existing methods, and obtained consistent, substantial improvements.

Appendix

A phase transition occurs when the temperature is reduced below a critical value, if the existing solution changes from a minimum of the Lagrangian functional F^* of (9) to a saddle point or a local maximum. We use this condition, and variational calculus to derive an expression for the critical temperatures.

Let us consider the perturbed central and side code books given by $\hat{\mathcal{X}}^0 + \epsilon\Psi^0 = \{\hat{x}_{ij}^0 + \psi_{ij}^0, (i,j) \in I_1 \times I_2\}$, $\hat{\mathcal{X}}^1 + \epsilon\Psi^1 = \{\hat{x}_i^1 + \psi_i^1, i \in I_1\}$ and $\hat{\mathcal{X}}^2 + \epsilon\Psi^2 = \{\hat{x}_j^2 + \psi_j^2, j \in I_2\}$, where ψ_{ij}^0, ψ_i^1 and ψ_j^2 are the perturbation vectors, and the non-negative scalar ϵ is used to scale the perturbations. We denote (Ψ^0, Ψ^1, Ψ^2) by Ψ and the vector of concatenated perturbations, $(\psi_{ij}^0 \;\; \psi_i^1 \;\; \psi_j^2)$, by ψ_{ij}. Further, we define the concatenation of central and side error vectors as $e_{ij} = ((x - \hat{x}^0_{ij}) \ \ (x - \hat{x}^1_i) \ \ (x - \hat{x}^2_j))$. In the subsequent derivations, unless transposed, all vectors are row vectors.

In terms of the Lagrange functional of (9) evaluated with the perturbed codebooks, $F^*(\hat{\mathcal{X}}^0 + \epsilon\Psi^0, \hat{\mathcal{X}}^1 + \epsilon\Psi^1, \hat{\mathcal{X}}^2 + \epsilon\Psi^2)$, we can write the necessary condition for the optimality of the code books $(\hat{\mathcal{X}}^0, \hat{\mathcal{X}}^1, \hat{\mathcal{X}}^2)$ as

$$\frac{d}{d\epsilon} F^*(\hat{\mathcal{X}}^0 + \epsilon\Psi^0, \hat{\mathcal{X}}^1 + \epsilon\Psi^1, \hat{\mathcal{X}}^2 + \epsilon\Psi^2)|_{\epsilon=0} = 0, \tag{24}$$

for all choices of finite perturbation Ψ. (Note that this leads directly to the centroid rules of (13)). We must also require a condition on the second derivative to ensure the minimum is stable:

$$\frac{d^2}{d\epsilon^2} F^*(\hat{\mathcal{X}}^0 + \epsilon\Psi^0, \hat{\mathcal{X}}^1 + \epsilon\Psi^1, \hat{\mathcal{X}}^2 + \epsilon\Psi^2)|_{\epsilon=0} \geq 0, \tag{25}$$

for all choices of finite perturbation Ψ. A necessary condition for bifurcation is equality in (25). Applying straightforward differentiation, we obtain the following condition for equality in (25):

$$\begin{aligned} &T \sum_x p(x) [\sum_{ij} (\frac{2}{T}) q(ij|x) e_{ij} L^2 \psi^t_{ij}]^2 \\ &+ 2 \sum_{ij} q(ij) \psi_{ij} L [I_{3n} - (\frac{2}{T}) L C_{x|ij} L] L \psi^t_{ij} = 0 \quad , \end{aligned} \tag{26}$$

where $q(ij|x)$ is given by (7). I_{3n} is the $(3n \times 3n)$ identity matrix.

$L = \begin{pmatrix} I_n & 0 & 0 \\ 0 & \sqrt{\lambda_1} I_n & 0 \\ 0 & 0 & \sqrt{\lambda_2} I_n \end{pmatrix}$, where I_n is the $(n \times n)$ identity matrix.

$C_{x|ij} = \sum_x p(x|ij) e^t_{ij} e_{ij}$ is the covariance matrix of the posterior distribution $p(x|ij)$ of the cluster corresponding to the index pair (i, j).

We claim that the left hand side of (25) is positive for all perturbations iff the second term of (26) is positive. The "if" part is trivial since the first term of (26) is non-negative. We prove the "only if" part. Consider a subset of index pairs, $\mathcal{C}$, with coincident central and side code vectors. This subset bifurcates if the matrix $I_{3n} - (\frac{2}{T}) L C_{x|ij} L$ loses positive definiteness, in which case the second term on the l.h.s. of (26) can be non-positive. (Note that $C_{x|ij}$ is the same for all the index pairs of this subset.) We now show a particular perturbation that makes the first term vanish in this case. It is easily verified that any perturbation satisfying

$$\psi_{ij} = 0, \ \ \forall \ \ (i, j) \notin \mathcal{C} \ \ and \ \sum_{(i,j) \in \mathcal{C}} \psi_{ij} = 0, \tag{27}$$

makes the first term of (26) vanish. So the subset $\mathcal{C}$ bifurcates at temperature T if the conditional distribution $p(x|ij)$ satisfies the condition

$$\det[I_{3n} - (\frac{2}{T}) L C_{x|ij} L] = 0. \tag{28}$$

The above condition is implicit in the critical temperature. The critical temperature for the first phase transition (i.e., when the code vectors coincident at the centroid of the source distribution move apart for the first time) can be explicitly evaluated, giving

$$T_{c_1} = 2(1 + \lambda_1 + \lambda_2)\alpha_{\max} \tag{29}$$

where $\alpha_{\max}$ is the largest eigenvalue of the source covariance matrix. This critical temperature may be compared with the critical temperature for the first phase transition when DA is used for single-description VQ design [2]: $T_{c_1}^{SDVQ} = 2\alpha_{\max}$, and

$$T_{c_1}^{MDVQ} = (1 + \lambda_1 + \lambda_2)T_{c_1}^{SDVQ}. \tag{30}$$

This result is expected since the MDVQ design algorithm degenerates into the DA algorithm for VQ design [2] if $\lambda_1 = \lambda_2 = 0$.

Acknowledgment

The authors wish to thank Ajit Rao for fruitful discussions and insightful comments at an early stage of this work.

References

[1] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 821-834, May 1993.

[2] K. Rose, E. Gurewitz and G. C. Fox, "Vector Quantization by Deterministic Annealing," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1249-1257, July 1992.

[3] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 129-137, Mar. 1982.

[4] M. Fleming and M. Effros, "Generalized multiple description vector quantization," in *Proceedings of the Data Compression Conference*, pp.3-12, Snowbird, UT, Mar. 1999. IEEE.

[5] S. D. Servetto, V. A. Vaishampayan and N. J. A. Sloane, "Multiple description lattice vector quantization," in *Proceedings of the Data Compression Conference*, pp.13-22, Snowbird, UT, Mar. 1999. IEEE.

[6] A. A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, pp.851-857, Nov. 1982.

[7] L. Ozarow, "On a source coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, pp. 1909-1921, Dec. 1980.

[8] H. Witsenhausen, "On source networks with minimal breakdown degradation," *Bell Syst. Tech. J.*, vol. 59, no. 6, pp. 1083-1087, July-Aug. 1980.

[9] J. Wolf, A. Wyner and J. Ziv, "Source coding for multiple descriptions," *Bell Syst. Tech. J.*, vol. 59, no. 8, pp. 1417-1426, Oct. 1980.

[10] H. S. Witsenhausen and A. D. Wyner, "Source coding for multiple descriptions II: A binary source," Bell Lab. Tech. Rep. TM-80-1217, Dec. 1980.

[11] T. Berger and Z. Zhang, "Minimum breakdown degradation in binary source encoding," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 807-814, Nov. 1983.

[12] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 502-521, July 1987.

[13] R. Ahlswede, "The rate-distortion region for multiple descriptions without excess rate," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 721-726, Nov. 1985.

[14] W. H. R. Equitz and T. M. Cover, "Successive refinement of information", *IEEE Trans. Inform. Theory*, vol. 37, pp. 269-275, Mar. 1991.

[15] E. T. Jaynes, "Information theory and statistical mechanics," in *Papers on Probability, Statistics and Statistical Physics*, R. D. Rosenkratz, Ed. Dordrecht, The Netherlands: Kluwer, 1989.

[16] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1939-1952, Nov. 1994.

[17] K. Rose, "Deterministic Annealing for clustering, compression, classification, regression and related optimization problems," *Proceedings of the IEEE*, vol.86, pp. 2210-2239, Nov. 1998.

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

[19] D. G. Luenberger, *Optimization by vector space methods.* New York: Wiley, 1969.

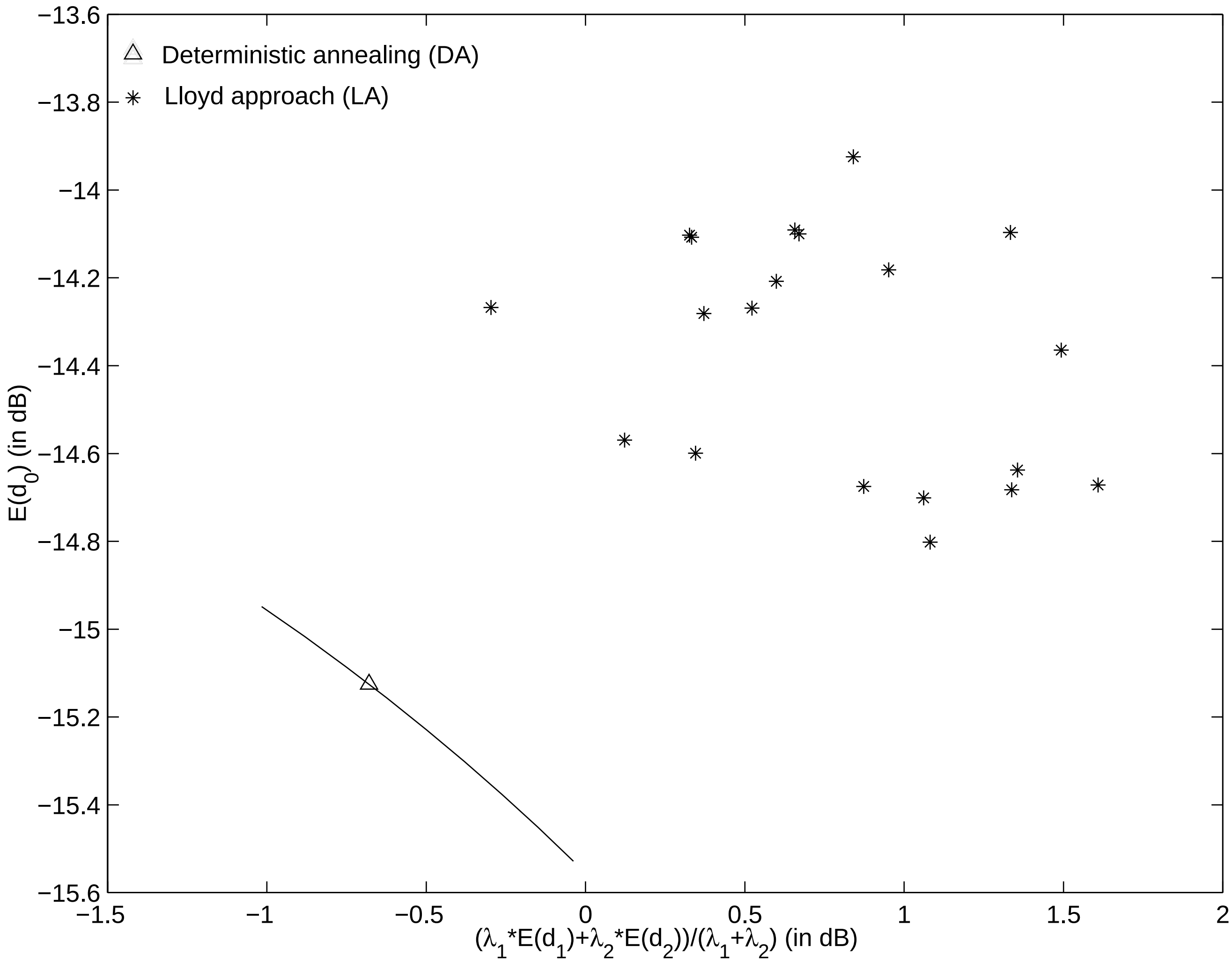

Figure 1: *2-d MDVQ for Gauss-Markov source,* $\rho = 0.9$. $R_1 = R_2 = 1.5bpss$, $\lambda_1 = \lambda_2 = 0.01$. *Minimum and maximum* D *for LA are* $-12.56dB$ *and* $-11.79dB$. D *for DA* $= -13.20dB$. *For ease of comparison, a line along which* $D = -13.20dB$ *is drawn.*

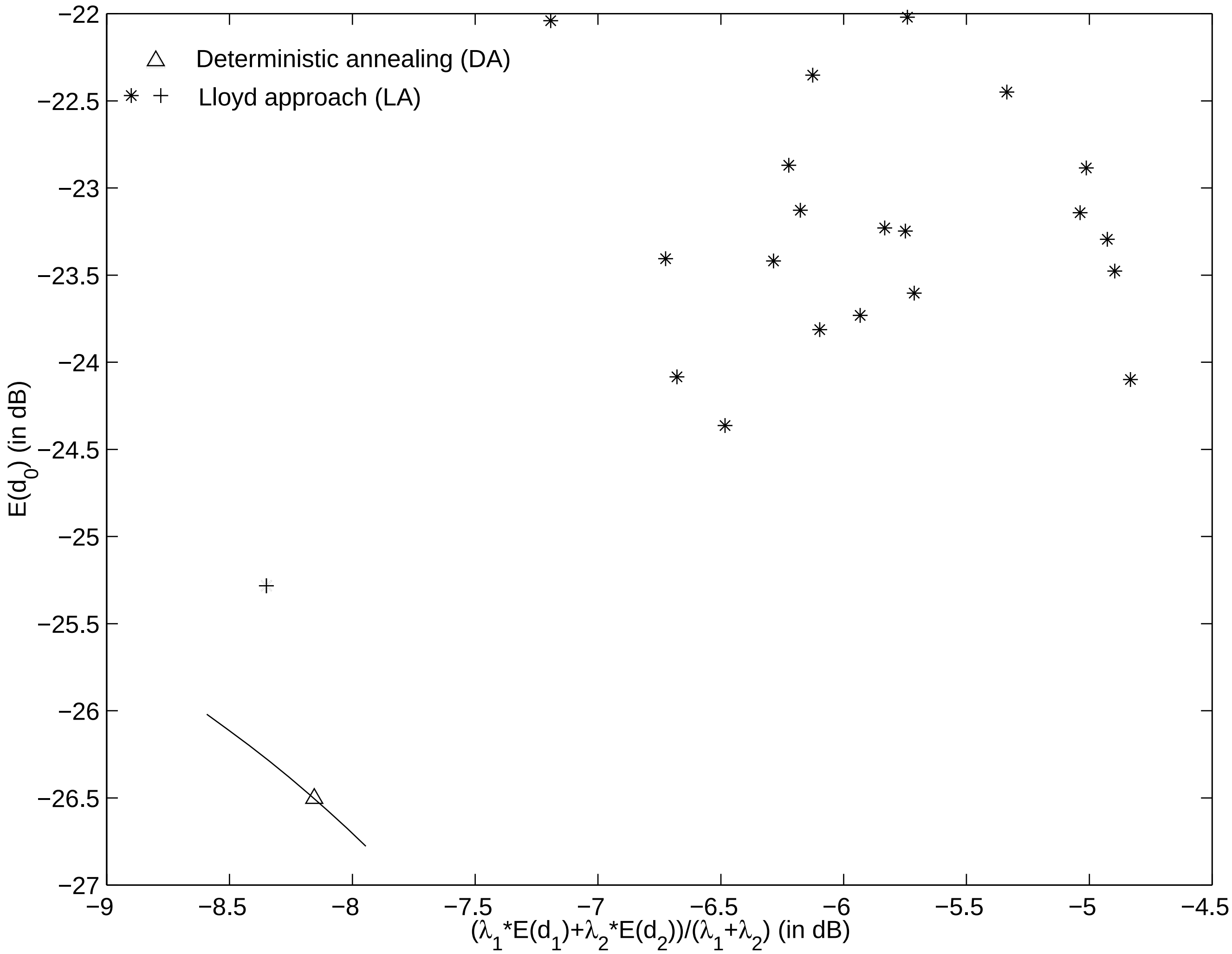

Figure 2: *Multiple description scalar quantizer for Gaussian source.* $R_1 = R_2 = 3$ *bpss,* $\lambda_1 = 0.006$, $\lambda_2 = 0.012$. *Minimum and maximum D for LA are* $-22.52dB$ *and* $-19.55dB$. *D for DA* $= -23.02dB$. *For ease of comparison, a line along which* $D = -23.02dB$ *is drawn. Design with initialization from [1] is marked by* $+$.

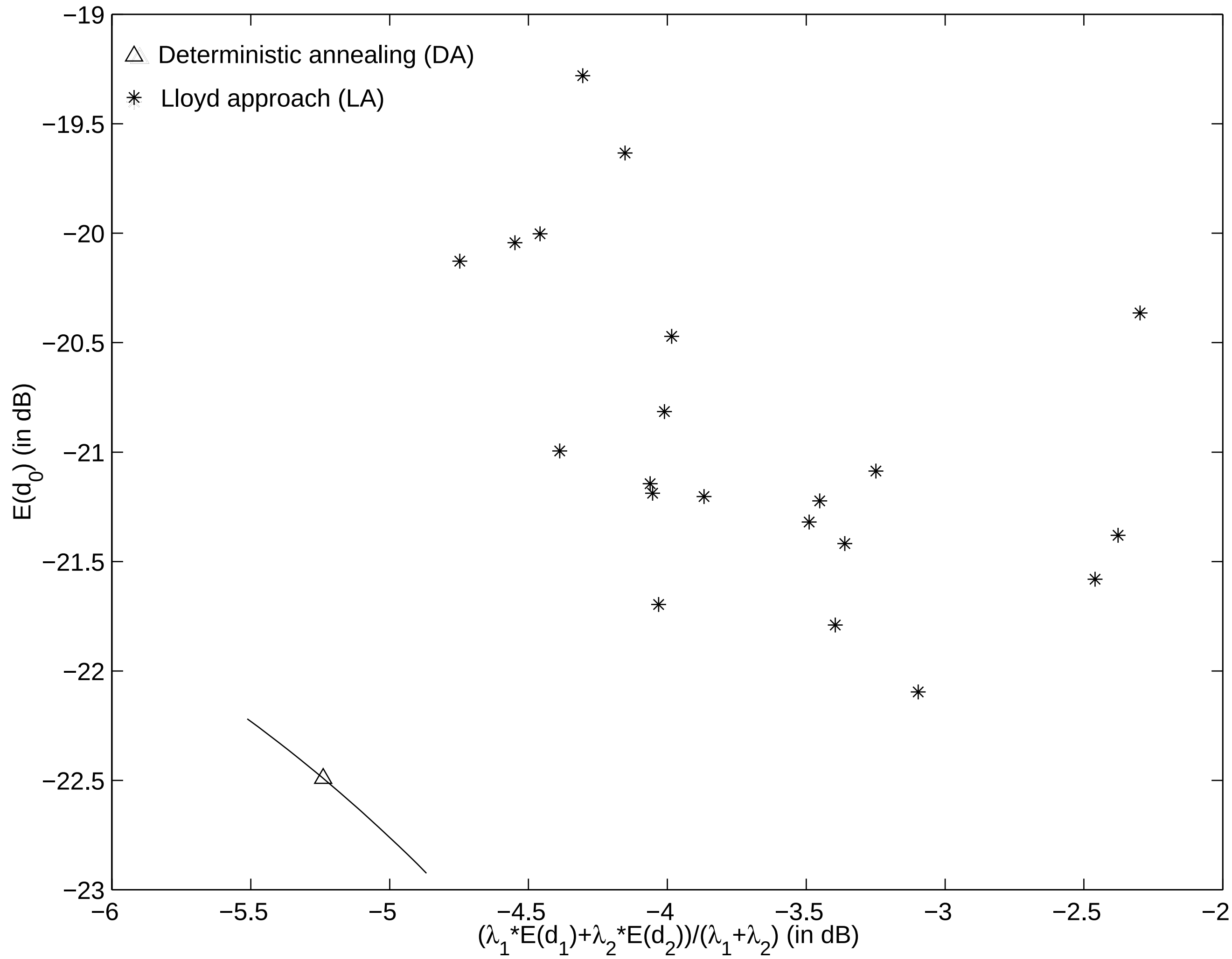

Figure 3: *Multiple description scalar quantizer for Gaussian source.* $R_1 = 3$ *bpss,* $R_2 = 2$ *bpss,* $\lambda_1 = \lambda_2 = 0.01$*. Minimum and maximum* D *for LA are* $-18.34dB$ *and* -16.78 *dB.* D *for* $DA = -19.35dB$*. For ease of comparison, a line along which* $D = -19.35dB$ *is drawn.*