Lagrangian Optimization of Two-description Scalar Quantizers

Sorina Dumitrescu, *Member, IEEE,* and Xiaolin Wu, *Senior Member, IEEE*

***Abstract*—We study the problem of optimal design of balanced two-description fixed-rate scalar quantizer (2DSQ) under the constraint of convex codecells. Using a graph-based approach to model the problem, we show that the minimum expected distortion of the 2DSQ is a convex function of the number of codecells in the side quantizers. This property allows the problem to be solved by Lagrangian minimization for which the optimal Lagrangian multiplier exists. Given a trial multiplier, we exploit a monotonicity of the objective function, and develop a simple and fast dynamic programming technique to solve the parameterized problem. To further improve the algorithm efficiency, we propose an RD-guided search strategy to find the optimal Lagrangian multiplier. In our experiments on distributions of interest for signal compression applications the proposed algorithm improves the speed of the fastest algorithm so far, by a factor of $O(K/\log K)$, where K is the number of codecells in each side quantizer.**

We also assess the impact on the optimality of the convex codecell constraint. Using a published performance analysis of 2DSQ at high rates, we show that asymptotically this constraint does not preclude optimality for L_2 distortion measure, when channels have a higher than 0.12 loss rate.

***Index Terms*—Multiple description quantization, distributed source coding, Lagrangian optimization, minimum-weight k-edge path, convexity of quantizer cells.**

I. Introduction

The problem of multiple description coding (MDC) was first posed at the 1979 IEEE Information Theory Workshop by A. D. Wyner. Early results appeared in [2], [10], [17], [21], [22], [25]. The research on MDC has intensified in recent years particularly in design algorithms, driven by the applications of networked media streaming and sensor networks, which require robust source coding.

Multiple description scalar quantization (MDSQ) is an extensively studied MDC technique. MDSQ holds the promise of being a practical solution to networked source coding for its simple, inexpensive implementations. Optimal MDSQ design, however, turns out to be nontrivial. The problem was first considered by Vaishampayan in the fixed-rate setting [18] and then for the entropy-constrained case [19]. The proposed algorithm is of generalized Lloyd-type and can guarantee only a locally optimal solution in general. Recently, Muresan and Effros proposed a graph-based approach to solve the problem in the case when the probability distribution is discrete and the side quantizers have convex codecells [15], [16]. They showed that the problem is equivalent to the minimum-weight path problem in a weighted directed acyclic graph. This finding leads to polynomial time algorithms which ensure global optimality under the imposed constraint of convex codecells. Their treatment covers the case of more than two, and unbalanced descriptions. Soon after, we proposed in [7] refinements to the graph theoretical approach for the fixed-rate case and two descriptions (2DSQ), which led to much faster design algorithms. Asymptotical analysis of multiple description scalar quantization at high rates was also provided in [20].

This paper reexamines the optimal 2DSQ design problem for the case of discrete distributions, fixed-rate and balanced side descriptions, under the constraint of convex codecells in the side quantizers. The optimization problem is to minimize the expected distortion or, equivalently, a weighted sum of the distortions of the side and central quantizers. By balanced or symmetric descriptions we mean that the two descriptions have the same rate and are weighted equally in the cost function. As proved in [7] the symmetry of the side descriptions allows for a simpler graph model for the problem. Relying on this model we prove an interesting property of the optimal fixed-rate balanced 2DSQ with convex codecells, namely, that its expected distortion is a convex function of the number of side quantizers codecells. This property enables us to solve the problem through Lagrangian minimization in conjunction with a search for the optimal Lagrangian multiplier. The appeal of this method is twofold. First, we show that the Lagrangian minimization for a given trial multiplier can be solved very efficiently by exploiting a monotonicity property of the cost function. Second, the performance analysis of 2DSQ at high rates provides us with an approximation of the optimal Lagrangian multiplier as a function of the number of codecells in the side partitions. Based on this approximation we derive an RD-guided search technique for the optimal Lagrangian multiplier. In our experiments on several distributions of interest for signal compression applications, this technique converges in at most $1.5\log_2 K$ iterations, achieving a speed improvement over the fastest existing algorithm by a factor of $K/\log_2 K$, where K is the number of codecells.

The convexity of codecells is apparently a limitation of our design approach. It was shown in [9], [16] that imposing the convexity of side quantizer codecells may result in performance loss. Using the performance analysis of 2DSQ at high rates provided by [20], we show, however, that asymptotically this constraint does not preclude optimality for channels of a failure rate higher than 0.12, in the case of L_2 distortion measure.

The paper is structured as follows. The next section presents

Supported by National Science Foundation grant CCR-0208678 as a professor at Polytechnic University, Brooklyn, NY, and by Natural Sciences and Engineering Research Council of Canada. An abstract of this work appeared in the Proceedings of 2004 IEEE Information Theory Symposium, Chicago, June-July 2004. The authors are with the Department of Electrical and Computer Engineering at McMaster University, Hamilton, Ontario, Canada L8S 4K1. E-mails: sorina/xwu@mail.ece.mcmaster.ca.

the necessary definitions, notations and the problem form lation. Also, a brief historical review of the existing 2DS design algorithms is given. In section III we present the gra representation of the problem of optimal fixed-rate balanc 2DSQ design, under the constraint of convex codecells the side quantizers. In section IV the graph-based constrain optimization problem is transformed to an unconstrained o via Lagrangian multiplier method. This leads to a desi approach of finding the minimum-weight path in a param terized graph $G(\lambda)$ in conjunction with a guided search λ, until the desired number of edges on the path is obtaine The central result of this section is that such a Lagrangi multiplier always exists. Section V develops an algorithm f the minimum-weight path problem in $G(\lambda)$, which is mo efficient than the standard solutions. The speed improveme is due to a strong monotonicity of the cost function. T following section discusses the search strategy of finding t optimal Lagrangian multiplier λ and its efficiency is assess analytically and/or empirically. In Section VII the effect of t constraint of convex codecells on the optimality of the 2DS solution is analyzed for the case of r-th power distortion, a corroborating empirical evidence is also presented. Secti VIII concludes the paper.

II. Problem Formulation and Existing Algorithm

Let X be a random variable over an alphabet $\mathcal{A} \subset \mathbf{R}$. A fixed-rate two-description scalar quantizer (2DSQ for short) is designed for communication over two channels (Fig. 1). It consists of two encoders f_1 and f_2, called side encoders, and three decoders g_1 and g_2 (the side decoders), and g_0 (the central decoder). Each source symbol x is encoded into two indices $i_1 = f_1(x)$ and $i_2 = f_2(x)$, and sent over the two side channels, one per channel. If only one channel transmits successfully, then only one index arrives at destination and can be decoded by the corresponding side decoder. When both indices i_1, i_2 arrive, they are jointly decoded by the central decoder. Formally, the side encoders are two functions $f_1 : \mathcal{A} \to \{1, \cdots, K_1\}$, $f_2 : \mathcal{A} \to \{1, \cdots, K_2\}$, for some integers $K_1, K_2 < N$. The side decoders are two one-to-one mappings $g_1 : \{1, \cdots, K_1\} \to \mathcal{C}_1$, $g_2 : \{1, \cdots, K_2\} \to \mathcal{C}_2$, where $\mathcal{C}_1, \mathcal{C}_2 \subset \mathbf{R}$ are two sets of reproduction values called codebooks. The central decoder g_0 maps each pair of indices (i_1, i_2), for which $f_1^{-1}(i_1) \cap f_2^{-1}(i_2) \neq \emptyset$, into a value in the central codebook $\mathcal{C}_0 \subset \mathbf{R}$. Let $\mathcal{I} = \{(i_1, i_2) | f_1^{-1}(i_1) \cap f_2^{-1}(i_2) \neq \emptyset\}$ and let $\mathcal{K}$ be the cardinality of $\mathcal{I}$. Then $\mathcal{C}_0$ has size $\mathcal{K}$, too. We refer to this 2DSQ as a (K_1, K_2)-level 2DSQ.

A (K_1, K_2)-level 2DSQ can also be regarded as a system of three quantizers $\mathbf{Q} = (Q_1, Q_2, Q_0)$, consisting of two side quantizers: Q_1 and Q_2, and a central quantizer Q_0. Each side quantizer Q_k, $k = 1, 2$, is specified by the encoder-decoder pair f_k, g_k. The central quantizer has the decoder g_0 and an implicit encoder $f_0 : \mathcal{A} \to \mathcal{I}$, such that $f_0(x) = (f_1(x), f_2(x))$ for any alphabet symbol x. Each encoder generates a partition of the source alphabet into codecells, a codecell being the set of all alphabet symbols mapped into the same index (or pair of indices). Thus, the three quantizers Q_1, Q_2, Q_0 have, respectively, K_1, K_2 and $\mathcal{K}$ codecells. Note that the partition of the central quantizer (central partition) is the intersection of the side partitions (i.e., the partitions of the side quantizers). Consider a suitable quantization distortion measure $d(x, Q(x)) \geq 0$. Then each quantizer Q_k, $k = 0, 1, 2$, is associated a distortion $D(Q_k)$ defined as:

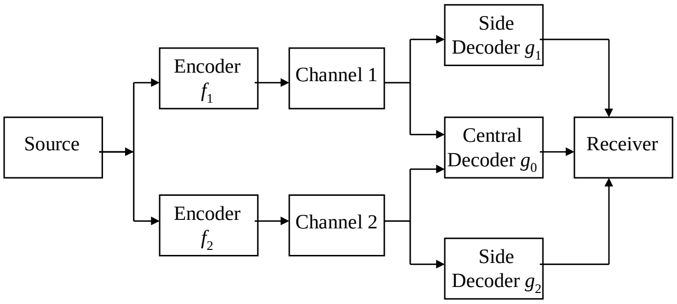


Fig. 1. Source coding scheme for communication over two channels.

$$D(Q_k) = E\{d(X, g_k(f_k(X)))\}.$$

We measure the performance of the 2DSQ by the expected distortion between the source and its reconstruction at the receiver side. Let ω_k be the probability that only the channel k transmits successfully ($k = 1, 2$), and ω_0 be the probability that both channels succeed. In case no description is received, the source is reconstructed at some high distortion D_0.Thus, the expected distortion of the 2DSQ can be expressed as

$$\bar{D}(\mathbf{Q}) = (1 - \omega_1 - \omega_2 - \omega_0)D_0 + \sum_{k=0}^{2} \omega_k D(Q_k). \quad (1)$$

If the two channels are independent with success probabilities q_1 and q_2, we have $\omega_1 = q_1(1 - q_2)$, $\omega_2 = q_2(1 - q_1)$ and $\omega_0 = q_1 q_2$.

The goal of optimal fixed-rate 2DSQ design is to construct a (K_1, K_2)-level 2DSQ $\mathbf{Q} = (Q_1, Q_2, Q_0)$ of minimal expected distortion $\bar{D}(\mathbf{Q})$.

The problem of optimal fixed-rate 2DSQ design was first addressed in [18]. The initial optimization criterion was slightly different, namely to minimize the distortion of the central quantizer subject to given upper bounds on the distortions of the side quantizers. This constrained optimization problem was solved in the classic Lagrangian form $L(\mathbf{Q}, \lambda_1, \lambda_2)$:

$$L(\mathbf{Q}, \lambda_1, \lambda_2) = D(Q_0) + \lambda_1 D(Q_1) + \lambda_2 D(Q_2) \quad (2)$$

with $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$, which is equivalent to minimizing $\bar{D}(\mathbf{Q})$, specifically when $\lambda_k = \omega_k / \omega_0$, $k = 1, 2$. Vaishampayan showed that for a continuous pdf and the squared difference distortion measure the optimal 2DSQ must have convex codecells in the central partition. He introduced the notion of index assignment as the mapping $h : \{1, 2, \cdots, \mathcal{K}\} \to \{1, 2, \cdots, K_1\} \times \{1, 2, \cdots, K_2\}$, defined by $h(l) = (i, j)$, where $f_1^{-1}(i) \cap f_2^{-1}(j)$ equals the l^{th} codecell (from left to right) of the central partition. This breaks the problem into two parts: choosing an index assignment and minimizing the Lagrangian given the index assignment. For the balanced case, where $K_1 = K_2$ and the distortions of the two side quantizers are approximately equal, it was conjectured from the experimental observations that the Langrangian multipli λ_1 and λ_2 should be equal in the optimal 2DSQ. Given t number $\mathcal{K}$ of codecells in the central partition, good inc assignments were proposed. Given the index assignment t Lagrangian was minimized by iteratively optimizing, in tu the decoder and the encoder. The algorithm can be appli to a discrete source as well, but as in the continuous case cannot guarantee the global optimum, not even with resp to a fixed index assignment.

As in [18] we also focus on the case of balanced 2DSQ. V call a 2DSQ K-level balanced if and only if $K_1 = K_2 =$ and the weights of the side distortions in $\bar{D}(\mathbf{Q})$ are equ i.e., $\omega_1 = \omega_2 = \omega$. This situation arises, for instance, when t independent channels operate at the same rate $\log_2 K$ and ea has the same success probability q. The problem of optin K-level balanced 2DSQ design is to minimize the expect distortion

$$\bar{D}(\mathbf{Q}) = (1-2\omega-\omega_0)D_0 + \omega(D(Q_1)+D(Q_2)) + \omega_0 D(Q$$

or equivalently minimize the Lagrangian $L(\mathbf{Q}, \lambda, \lambda)$, as cc sidered in [18], for $\lambda = \omega/\omega_0$.

In pursuing an efficient and globally optimal solution the problem, we consider a discrete source and restrict t solution space to 2DSQs with convex codecells in the side quantizers (we call such a 2DSQ, convex 2DSQ). This setting was first addressed by Muresan and Effros in [15], [16]. They treated the multiple description quantizer design for arbitrary number of descriptions, the descriptions not being necessarily balanced. Both the fixed-rate and entropy-constrained cases were addressed. They showed that the problem can be modeled as a minimum-weight path problem in a weighted directed acyclic graph, and hence polynomially solvable. Their algorithm requires $O(K_1 K_2 N^3)$ time and $O(K_1 K_2 N^2)$ space. In [7] the time complexity for optimal fixed-rate convex 2DSQ design is reduced to $O(K_1 K_2 N^2)$ for monotone distortion measures $d(x, Q(x))$ that satisfy the condition

$$\begin{aligned} & d(x, y_1) \le d(x, y_2), \quad \text{for all real values } \; x, y_1, y_2 \\ & \qquad \text{such that } \; x \le y_1 < y_2 \; \text{ or } \; x \ge y_1 > y_2. \end{aligned} \tag{4}$$

Note that all known distortion measures used in practice are monotone. It was also proved in [7] that for balanced descriptions the time and space complexities of the algorithm can be further reduced to $O(KN^2)$ by exploiting additional properties of the solution conferred by the symmetry of the descriptions.

This paper reexamines the above problem aiming for an improved balanced convex 2DSQ design algorithm. We prove the convexity of the optimization problem and exploit it to develop a fresh algorithmic approach to solve it. Toward presenting the new approach we start from the graph model established in [7], which is detailed in the next section.

III. Graph Representation

To develop the new 2DSQ design algorithm, we use a graph representation of the problem. This graph model is different and simpler than the graph proposed in [15], [16]. The simplification is achieved by exploiting the symmetry of the descriptions and the monotonicity of the distortion measure (4).

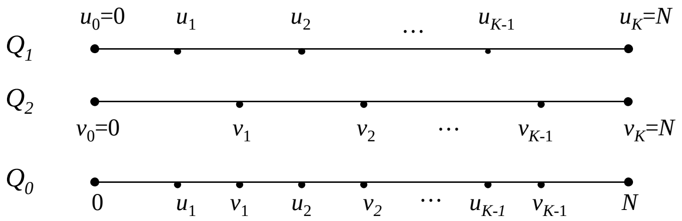


Fig. 2. K-level balanced convex 2DSQ with alternating thresholds of side quantizers.

Assume the source alphabet is finite, i.e., $\mathcal{A} = \{x_1, x_2, \cdots, x_N\} \subset \mathbf{R}$, with $x_i < x_{i+1}$, for all $1 \le i \le N-1$. Let the probability mass function (pmf) of X be $p_i = p(X = x_i)$, $1 \le i \le N$. For integers a, b such that $0 \le a \le b \le N$, denote $c(a, b] = \{x_i | a < i \le b\}$. Note that $c(a, a] = \emptyset$. A codecell is said to be convex if it is of the form $c(a, b]$. Such a codecell will be simply denoted by $(a, b]$. A scalar quantizer Q is a K-level convex scalar quantizer, if its encoder partitions the source alphabet into K codecells $(q_j, q_{j+1}]$, $0 \le j \le K-1$, for some integers q_j, $0 \le j \le K$, such that $0 = q_0 < q_1 < \cdots q_{K-1} < q_K = N$. The values q_j, $1 \le j \le K$, are called the quantizer's thresholds.

For each codecell $(a, b]$, let $\mu(a, b]$ denote its reproduction value. The contribution of codecell $(a, b]$ to the quantizer's distortion is $\sum_{i=a+1}^{b} d(x_i, \mu(a, b])p_i$. We consider only quantizers with the decoder optimized for the given encoder, therefore ideally the following relation should hold

$$\mu(a, b] = \arg\min_{y \in \mathbf{R}} \sum_{i=a+1}^{b} d(x_i, y)p_i.$$

When the distortion function is the squared distance, i.e., $d(x, y) = (x-y)^2$, there is a closed form available for the minimum in the above equation, which makes $\mu(a, b]$ computable in a finite number of operations,

$$\mu(a, b] = \frac{\sum_{i=a+1}^{b} x_i p_i}{\sum_{i=a+1}^{b} p_i}. \tag{5}$$

However, for general distortion function $d(\cdot, \cdot)$ a closed form is not known for the minimum in (5), and a continuous optimization algorithm cannot guarantee convergence in a finite number of steps. To have the algorithm terminate in finitely many steps we need to resort to some approximation, for example to stop at some level of precision or restrict the search for the minimum to a finite grid. All previous work which claims optimal quantizer design [23], [24], [15], [16] resort to such approximations. This approximation can be modeled by restricting the possible reconstruction values to a finite alphabet $\mathcal{B}$, where $\mathcal{B}$ can be a finite grid of some required precision. Then the reproduction value $\mu(a, b]$ must satisfy the relation

$$\mu(a, b] = \arg\min_{y \in \mathcal{B}} \sum_{i=a+1}^{b} d(x_i, y)p_i. \tag{6}$$

In this work we use the definition of $\mu(a,b]$ giver the case of squared distance as distortion measure, r the definition of (6) otherwise.

Further denote the distortion of the codecell $D(a,b]$. Then

$$D(a,b] = \sum_{i=a+1}^{b} d(x_i, \mu(a,b])p_i.$$

The distortion of the quantizer becomes:

$$D(Q) = \sum_{j=0}^{K-1} D(q_j, q_{j+1}].$$

We are concerned with the problem of optim balanced convex 2DSQ design, i.e., the problem of r the expected distortion (3) among all 2DSQ's with codecells in each side quantizer.

We denote by $u_0, u_1, \cdots u_K$, respectively, $v_0, v_1,$ thresholds of the first, respectively second, side q a K-level balanced convex 2DSQ. Since the centr is the intersection of the side partitions, it follov thresholds of the central partition are actually the of the two side partitions ordered in increasing order. Note that $0 = u_0 < u_1 < \cdots < u_K = N$ and $0 = v_0 < v_1 < \cdots < v_K = N$.

The following proposition was proved in [7, Proposition 2].

Proposition 1. There is an optimal K-level balanced convex 2DSQ such that the side quantizers thresholds alternate (Fig. 2):

$$\begin{aligned} u_0 \leq v_0 \leq u_1 \leq v_1 \leq u_2 \leq v_2 \leq \cdots \\ \cdots \leq u_{K-1} \leq v_{K-1} \leq u_K \leq v_K. \end{aligned} \qquad (7)$$

Proposition 1 converts the optimization problem into a graph problem as follows. Consider the weighted directed acyclic graph (WDAG) $G = (V, E)$, whose nodes (or vertices) are all ordered pairs of integers a and b such that $0 \leq a \leq b \leq N$. We denote such a pair simply by ab. The set of edges is $E = \{(ab, bc) | 0 \leq a \leq b \leq c \leq N, a < c\}$. Let 00 be the source node and NN the final node of the graph. The weight of the edge from node ab to node bc is defined as $w(ab, bc) = \omega D(a,c] + \omega_0 D(a,b]$. We can associate with any K-level balanced convex 2DSQ of alternating thresholds (7), a $2K$-edge path (i.e., a path with $2K$ edges) from the source to the final node:

$$00, 0u_1, u_1v_1, v_1u_2, u_2v_2, \cdots, v_{K-1}N, NN. \qquad (8)$$

Fig. 3 illustrates the above path.

As shown in [7] this mapping is one-to-one. Moreover, the weight of the path associated with a 2DSQ $\mathbf{Q}$ as above (i.e., the sum of weights of its edges) equals $\bar{D}(\mathbf{Q}) - (1 - 2\omega - \omega_0)D_0$.

Consequently, minimizing $\bar{D}(\mathbf{Q})$ is equivalent to the minimum-weight $2K$-edge path problem in the graph G (i.e. finding the path of minimum weight among all paths from the source to the final node, which have exactly $2K$ edges). This problem was solved in $O(KN^2)$ time [7]. Precisely, the algorithm presented in [7] runs in $2K-1$ iterations, at the k-th iteration the minimum-weight $(k+1)$-edge paths from the source to each node, being computed. In this paper we develop a Lagrangian-type algorithm for the globally optimal solution of the same problem.

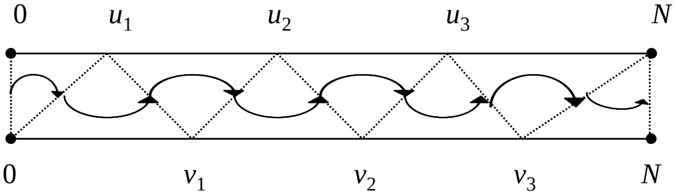


Fig. 3. Path in the WDAG G corresponding to a 4-level balanced convex 2DSQ with alternating thresholds. The nodes are represented by dotted segment lines which connect quantizer thresholds. The edges are represented by arrowed arcs. Precisely, the path illustrated is $00, 0u_1, u_1v_1, v_1u_2, u_2v_2, v_2u_3, u_3v_3, v_3N, NN$.

IV. Lagrangian Solution

The graph problem formulated above is a constrained optimization problem. The constraint is on the number of edges in the minimum weight path. A standard technique to solve constrained optimization problems is the Lagrangian method. Indeed, the Lagrangian method is ubiquitous in the literature of entropy-constrained optimal quantizer design, including the multiple description variants [6], [19], [15]. However, strictly speaking, in the entropy-constrained case this strategy can lead to solutions only for some instances of the original quantizer design problem, namely, those rate-distortion pairs on the lower convex hull of the operationally achievable rate-distortion region. The Lagrangian-based approach was also applied to fixed-rate scalar quantization [1], [12] and what is remarkable in this case is that it leads to the globally optimal solution to any instance of the constrained problem. We prove in this section that this property holds in the case of optimal fixed-rate balanced convex 2DSQ design as well.

Let $\mathcal{P}$ denote the set of all paths from the source node to the final node, in the graph G. For any path $P \in \mathcal{P}$ let $W(P)$ be its weight and $L(P)$ its length (the number of component edges). Consider the set of planar points $\mathcal{U} = \{(L(P), W(P)) | P \in \mathcal{P}\}$.

Then the problem of minimum-weight $2K$-edge path in G can be formulated as

$$\begin{aligned} &\text{minimize}_{P \in \mathcal{P}} W(P) \\ &\text{subject to } L(P) = 2K. \end{aligned} \qquad (9)$$

The underlying Lagrangian is $J(\lambda, P) = W(P) + \lambda L(P)$, over all paths $P \in \mathcal{P}$ and all real values λ. A path P^* minimizes the Lagrangian for some λ, i.e., the relation

$$J(\lambda, P^*) = \min_{P \in \mathcal{P}} J(\lambda, P) \qquad (10)$$

holds, if and only if the planar point $(L(P^*), W(P^*))$ is on the lower convex hull of $\mathcal{U}$ and the line of slope $-\lambda$ passing through this point is a support line to $\mathcal{U}$ [14], [11]. Thus, if (10) holds then the path P^* is also a minimum-weight $L(P^*)$-edge path because the lower boundary of $\mathcal{U}$ is not below its lower convex hull. Consequently, if a Lagrangian multiplier λ can be found for which there is a $2K$-edge path P^* satisfying

(10)[1], then this path is a solution of the constrained probl (9). Due to the following proposition, whose proof is giver Appendix A, such a multiplier λ is guaranteed to exist.

Proposition 2. The inequality

$$2\bar{W}(l) \leq \bar{W}(l-1) + \bar{W}(l+1) \qquad ($$

holds for all integers $l, 3 \leq l \leq 2N-1$, where $\bar{W}(l)$ is weight of the minimum-weight l-edge path from the source the final node in G.

The above proposition implies that any point $(l, \bar{W}(l))$ is the lower convex hull of $\mathcal{U}$. Let P_{2K} be a minimum wei $2K$-edge path, then the point $(L(P_{2K}), W(P_{2K}))$ coinci with $(2K, \bar{W}(2K))$, hence the following relation holds:

$$J(\lambda, P_{2K}) = \min_{P \in \mathcal{P}} J(\lambda, P) \qquad ($$

if and only if λ satisfies the relation

$$\bar{W}(2K) - \bar{W}(2K-1) \leq -\lambda \leq \bar{W}(2K+1) - \bar{W}(2K). \ ($$

This is because a line passing through $(L(P_{2K}), W(P_{2I}$ is a support line to $\mathcal{U}$ if and only if its slope is at le equal to the slope of the convex hull edge to the left (i $\bar{W}(2K) - \bar{W}(2K-1)$) and at most equal to the slope of convex hull edge to the right (i.e., $\bar{W}(2K+1) - \bar{W}(2K$ Denote by I_{opt} the range of the optimal Lagrangian multipli λ, i.e., those for which (13) holds. Consequently, I_{opt} $-$ $[\bar{W}(2K) - \bar{W}(2K+1), \bar{W}(2K-1) - \bar{W}(2K)]$. The interval I_{opt} reduces to a single value if the points $(2K-1, \bar{W}(2K-1))$, $(2K, \bar{W}(2K))$ and $(2K+1, \bar{W}(2K+1))$ are collinear.

For any λ in the interior of the interval I_{opt}, any path P^* satisfying (10) has the length $2K$. If λ equals the boundary to the left, respectively right, of I_{opt}, then there is also a path of length $2K+1$, respectively, $2K-1$, satisfying (10).

Therefore, the $2K$-edge minimum-weight path, or equivalently the globally optimal convex balanced K-level 2DSQ, can be found by solving (10) in conjunction with a search on λ until the number of the edges on the minimizing path becomes exactly $2K$. To this end we derive from G a parameterized graph $G(\lambda)$ by adding λ to the weight of each edge of G. In the resulting parameterized graph $G(\lambda)$ the minimization problem of (10) reduces to an unconstrained minimum-weight path problem. This is because $J(\lambda, P)$ equals the weight of the path P in $G(\lambda)$.

Remark 1. An immediate corollary of Proposition 2 is the convexity of the minimum expected distortion of fixed-rate balanced convex 2DSQ, as a function of the number of codecells in each side partition.

V. The Computation of the Minimum-weight Path in $G(\lambda)$

In the parameterized graph $G(\lambda)$, for any node ab other than the source, let $W_\lambda(a,b)$ denote the smallest weight of any path from 00 to the node ab. By convention, $W_\lambda(0,0) = 0$.

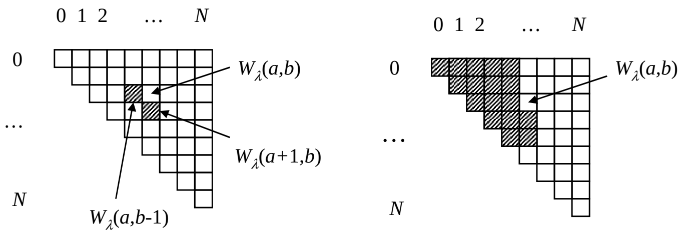


Fig. 4. Matrix $W_\lambda(\cdot,\cdot)$. Left: the shaded squares represent the locations of entries $W_\lambda(a,b-1)$ and $W_\lambda(a+1,b)$; they have to be known in order to apply the recursive formula with reduced search range for $W_\lambda(a,b)$. Right: the shaded squares represent the values already computed by the algorithm before the evaluation of $W_\lambda(a,b)$.

Recall that the weight of any edge (ab, bc) in $G(\lambda)$ equals $w(ab,bc) + \lambda$. Then for any node ab other than the source (i.e. with $b \geq 1$), we have

$$W_\lambda(a,b) = \min_{0 \leq \xi \leq a, \xi < b} \{W_\lambda(\xi, a) + w(\xi a, ab) + \lambda\}. \qquad (14)$$

Let $\xi_\lambda(a,b)$ be the value of ξ where the minimum of (14) is attained (in case of multiple points, the largest one is picked). The following proposition, which is proved in Appendix A, enables us to decrease the search range of (14).

Proposition 3. For any nodes ab and $a'b'$ other than the source, and such that $a \leq a'$ and $b \leq b'$, the following inequality holds:

$$\xi_\lambda(a,b) \leq \xi_\lambda(a',b').$$

Proposition 3 in conjunction with relation (14) immediately imply the following result.

Corollary. For all $0 \leq a < b \leq N$, we have

$$W_\lambda(a,b) = \min_{\xi_\lambda(a,b-1) \leq \xi \leq \xi_\lambda(a+1,b); \xi \leq a} \{W_\lambda(\xi,a) + w(\xi a, ab) + \lambda\}. \qquad (15)$$

In order to find a minimum-weight path in $G(\lambda)$ we compute $W_\lambda(a,b)$ for increasing values of a, b, $0 \leq a \leq b \leq N$, using the recursive relations (14) and (15) until reaching $W_\lambda(N,N)$. The path is then traced back by using the values $\xi_\lambda(a,b)$. Note that there may be several paths of minimum weight in $G(\lambda)$. These paths may even have different numbers of links (in the case when $-\lambda$ is the slope of a convex hull edge of $\mathcal{U}$). The path constructed by our algorithm is a minimum weight path with the largest number of links, which is proved in Lemma 4 in Appendix A.

The computations are organized in such a way that the entries of the upper triangular matrix $W_\lambda(\cdot,\cdot)$ are filled column by column from left to right as illustrated in Fig. 4. For each column b we first compute the entry $W_\lambda(b-1,b)$ using (14), and then proceed toward the top of the column, by applying recursion (15). Note that this recursion can be applied only if the entries immediately to the left and immediately below the current position are known. After reaching the top of the column, we finally compute $W_\lambda(b,b)$, the element at the bottom, again using (14), because this entry depends on all the other elements of the column.

In the above procedure computing all entries on the main diagonal and the superdiagonal of $W_\lambda(\cdot,\cdot)$ needs $O(N^2)$ time since each entry on these diagonals takes $O(N)$ time. But

[1] Note that a path P^* satisfying (10) is not necessarily unique. Moreover, different such paths may have different lengths.

computing all entries on any of the other $N-2$ diagonals of the upper triangular matrix of Fig. 4 collectively needs only $O(N)$ time. Indeed, let us call the j-th superdiagonal, the set of entries $W_\lambda(a,b)$ with $b=a+j$, $0 \le a \le N-j$. The entry $W_\lambda(a,a+j)$ is computed in $O(\xi_\lambda(a+1,a+j)-\xi_\lambda(a,a+j-1)+1)$ time. Hence, the total time for the j-th superdiagonal is

$$O(\textstyle\sum_{a=0}^{N-j}(\xi_\lambda(a+1,a+j)-\xi_\lambda(a,a+j-1)+1)) = \\ O(\xi_\lambda(N-j+1,N)-\xi_\lambda(0,j-1)+N-j+1)=O(N).$$

In conclusion, evaluating the whole matrix requires $O(N^2)$ time and $O(N^2)$ space.

On a second reflection, however, it is unnecessary to evaluate the entire matrix $W_\lambda(\cdot,\cdot)$ to arrive at $W_\lambda(N,N)$. The entries of a column b, $W_\lambda(a,b)$ with $b-1 \ge a \ge 0$, are only needed to compute the entries on the row b, i.e., $W_\lambda(b,c)$ with $b \le c \le N$. But, according to Proposition 3, we have $\xi_\lambda(b,c) \ge \xi_\lambda(b-1,b-1)$. Consequently, only the entries of column b up to the row $\xi_\lambda(b-1,b-1)$ are needed. The pseudo code given below describes this improved version of the algorithm.

```
**Minimum-weight path in** $G(\lambda)$.
$\xi_\lambda(0,1)=0; W_\lambda(0,1)=w(00,01)+\lambda;$
$\xi_\lambda(1,1)=0; W_\lambda(1,1)=w(00,01)+w(01,11)+2\lambda;$
for $b=2$ to $N$ do
    $a:=b-1;$
    $W_\lambda(a,b):=\min_{\xi_\lambda(b-1,b-1)\le\xi\le a}\{W_\lambda(\xi,a)+w(\xi a,ab)+\lambda\};$
    $\xi_\lambda(a,b):=\max \arg\min_{\xi_\lambda(b-1,b-1)\le\xi\le a}\{W_\lambda(\xi,a)+w(\xi a,ab)+\lambda\};$
    for $a=b-2$ down to $\xi_\lambda(b-1,b-1)$ do
        $W_\lambda(a,b):=\min_{\xi_\lambda(a,b-1)\le\xi\le\xi_\lambda(a+1,b);\xi\le a}\{W_\lambda(\xi,a)+$
                $w(\xi a,ab)+\lambda\};$
        $\xi_\lambda(a,b):=\max \arg\min_{\xi_\lambda(a,b-1)\le\xi\le\xi_\lambda(a+1,b);\xi\le a}\{W_\lambda(\xi,a)+$
                $w(\xi a,ab)+\lambda\};$
    end for
    $W_\lambda(b,b):=\min_{\xi_\lambda(b-1,b-1)\le\xi<b}\{W_\lambda(\xi,b)+w(\xi b,bb)+\lambda\};$
    $\xi_\lambda(b,b):=\max \arg\min_{\xi_\lambda(b-1,b-1)\le\xi<b}\{W_\lambda(\xi,b)+w(\xi b,bb)+\lambda\};$
end for
```

For a better comparison with the previous work it is useful to provide a more precise assessment of the computational requirement of the above algorithm. The algorithm solves a series of minimization problems. To find the minimum over n quantities, each of them has to be inspected. Assuming that all n quantities are already computed, let γ denote the average number of operations per quantity (hence $\gamma \ge 1$). Then the minimization requires γn operations. In our algorithm we need two extra operations (two additions) to evaluate each quantity inspected in the minimization. Following the analysis in the previous paragraphs we conclude that at most $2N^2$ quantities are inspected in order to solve all the minimizations (because at most $2(N-j+1)$ quantities have to be inspected to fill the j-th superdiagonal of $W_\lambda(\cdot,\cdot)$, for each $j \ge 2$, and at most N^2+N, to fill the rest). Therefore, the whole algorithm requires at most $2(\gamma+2)N^2$ operations.

Next we evaluate the number of operations required by a single iteration of the algorithm of [7] for the optimal K-level balanced convex 2DSQ. Recall that the algorithm of [7] consists of $2K-1$ iterations. The k-th iteration computes the minimum $(k+1)$-edge path in the graph G from the source to each graph node (a,b) based on the minimum k-edge paths. That procedure also solves a series of minimization problems in order to fill an upper triangular matrix of the same size as our $W_\lambda(\cdot,\cdot)$. The evaluation of each quantity inspected in the minimization process needs only one operation (one addition) and the total number of quantities inspected is at least N^2 (because at least $N^2/2$ quantities have to be inspected in order to fill the main diagonal and at least $N-j+1$ for the j-th superdiagonal, for each $j \ge 1$). Therefore the total number of operations is at least $(\gamma+1)N^2$.

Since $\gamma \ge 1$ it follows that $2(\gamma+2)N^2 \le 3(\gamma+1)N^2$, which implies that the algorithm proposed in this section for the minimum-weight path in the graph $G(\lambda)$ requires at most as many computations as three iterations of the algorithm of [7].

When evaluating the time complexity of the proposed algorithm, we have assumed that each value $D(a,b]$ can be accessed in constant time. It was showed in [24], [8] that for all monotone distortion measures $d(\cdot,\cdot)$, the distortion values $D(a,b]$ over all possible intervals $(a,b]$, $0 \le a \le b \le N$, can be precomputed in $O(MN)$ time, where M is the size of the alphabet $\mathcal{B}$. Since $M=O(N)$, the required precomputation takes $O(N^2)$ time, not affecting the time complexity of the proposed algorithm. Furthermore, if the distortion measure is the ubiquitous mean-square error, the preprocessing time reduces to $O(N)$ [23]. Thus, by using a reproduction values' alphabet $\mathcal{B}$ of size $O(N^2)$, the overall complexity result still holds.

VI. RD-guided Search of the Lagrangian Multiplier

Having developed an efficient algorithm to compute the minimum-weight path in $G(\lambda)$ for a given Lagrangian multiplier λ, our attention is turned to reduce the number of iterations in finding a Lagrangian multiplier to meet the targeted number $2K$ of edges.

For each λ, denote by P_λ, the path which minimizes the Lagrangian $J(\lambda,P)$ over all $P \in \mathcal{P}$, and has the largest number of edges among all paths with this property (i.e., the path computed by the algorithm of the previous section). The length of P_λ is non-increasing in λ. Explained briefly, this is because as λ increases, the intersection of the support line of slope $-\lambda$ with the set $\mathcal{U}$, either remains the same or moves to the left.

This monotonicity can be exploited to expedite the search as follows. At any time a search interval (λ_1,λ_2) for the optimal Lagrangian multiplier is maintained, with the property that $I_{opt} \subset (\lambda_1,\lambda_2)$, i.e. $L(P_{\lambda_1}) > 2K > L(P_{\lambda_2})$. At the beginning of each iteration, a value λ_{new} is picked from the interval (λ_1,λ_2) according to a rule for updating λ. Then $P_{\lambda_{new}}$ is computed. If its length equals $2K$ then the algorithm stops, otherwise, the current search interval (λ_1,λ_2) is updated to $(\lambda_1,\lambda_{new})$ if $2K > L(P_{\lambda_{new}})$ or to $(\lambda_{new},\lambda_2)$ if $2K < L(P_{\lambda_{new}})$.

Initially, the search interval (λ_1, λ_2) is set to $(0, \gamma)$, where

$$\gamma = (2\omega + \omega_0)D(0, N]. \tag{16}$$

This ensures that $(0, \gamma)$ contains I_{opt}. The reason is the following. The planar point $(2N, \bar{W}(2N))$ is the rightmost point of intersection between the support line of slope 0 and $\mathcal{U}$ (because the function $\bar{W}(l)$ is non-increasing). Therefore, for $\lambda = 0$ we have $L(P_\lambda) = 2N$. On the other side, $\gamma = \bar{W}(2) > \bar{W}(2) - \bar{W}(3)$. Thus, the support line of slope $-\gamma$ intersects $\mathcal{U}$ at the point $(2, \bar{W}(2))$, which implies that $L(P_\gamma) = 2$. Further, since $L(P_\lambda)$ is non-increasing as λ increases, our claim follows.

In the above search framework, we propose two techniques for choosing the next trial λ value. The first technique, called RD-guided search, is derived from the fact that $-\lambda$ represents the slope of the rate-distortion function. Consider 2DSQ design for a pdf $f(x)$ defined on a compact interval under distortion metric $d(x, y) =\mid x - y \mid^r$. Let $\bar{D}^{(r)}(R)$ be the minimum expected distortion among all 2^R-level balanced convex 2DSQ's. As proved in Appendix B (Eq. (46)) with arguments along the lines of [4], [3], we have

$$\bar{D}^{(r)}(R) \approx \frac{2\omega_1 + 2^{-r}\omega_0}{2^{-r-rR}(r+1)}(\int_V^W f^{1/(r+1)}(x)dx)^{r+1} \tag{17}$$

as $R \to \infty$. Consequently, $\bar{W}(l)$ is proportional to $\frac{1}{l^r}$ as l becomes very large, and its derivative is proportional to $\frac{1}{l^{r+1}}$. Based on this property of $\bar{W}(l)$ we use the following interpolation technique to update λ in the Lagrangian optimization. First we find the real values α and β such that

$$\lambda_i = \frac{\alpha}{L(P_{\lambda_i})^{r+1}} + \beta, \qquad i = 1, 2.$$

Then update the λ value to

$$\lambda_{new} = \frac{\alpha}{(2K)^{r+1}} + \beta.$$

Clearly, the number of iterations required to find the optimal Lagrangian multiplier by the RD-guided search depends on the quality of the approximation (17). If (17) held with equality the number of iterations would be 1. The better the approximation the smaller the number of iterations. Bounding the error in (21) seems very difficult. But we do have empirical evidence to support the high efficiency of the RD-guided search.

We tested the RD-guided search on several source distributions: Gaussian, Laplacian, and the mixture of two Gaussians. These distributions are widely used to model real data in signal compression applications. We have also included in our test set a real p.m.f. of DPCM residuals obtained from an audio signal (Fig. 5). Most data to be quantized in practice, such as transform coefficients (wavelet, DCT, etc.) and DPCM residuals, have a p.m.f. like Fig. 5, obeying Laplacian or generalized Gaussian distribution. Experiments are conducted for three values of N: $500, 1000$ and 2000, various channel success probabilities $q = 0.5, 0.6, 0.7, 0.8, 0.9$, and all K values ranging from 1 to 49. The distortion measure used is the squared distance. The number of iterations in relation to K, N, and q are presented in Figs. 6-12. Before running the algorithm a continuous p.d.f. is first discretized via uniform prequantization.

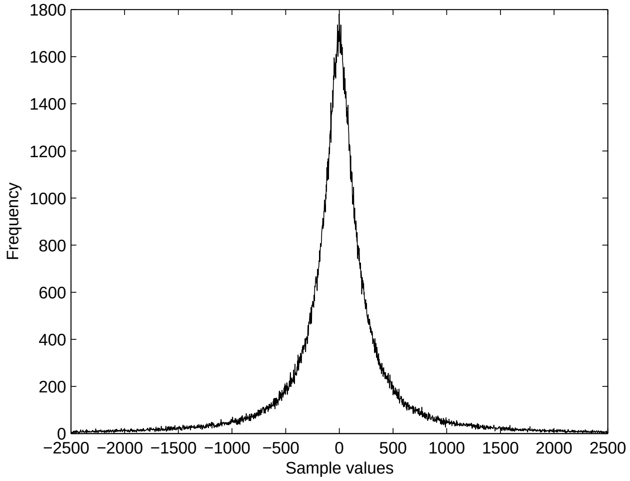


Fig. 5. Histogram of DPCM residuals to be quantized.

Figs. 6-9 plot the average number of iterations (averaged over q), versus K, for the zero-mean unit-variance Gaussian and Laplacian distributions, and the following Gaussian mixtures

$$f_1(x) = 1/2g(-1, 1) + 1/2g(1, 4), \tag{18}$$

$$f_2(x) = 3/4g(-1, 1) + 1/4g(1, 4), \tag{19}$$

where $g(\mu, \sigma^2)$ is the normal pdf of mean μ and variance σ^2. Each figure contains plots for different N values. Fig. 10 plots the results for the real sample p.m.f.

One can observe in Figs. 6- 10 that the average number of iterations is not monotonic in K, but it has a general tendency of increase with K at a growth rate close to $O(\log K)$. To quantify this we include in the figures 6-10 a plot of the function $\alpha \log_2 K$, with an α chosen for each case approximately as the smallest positive value such that $\alpha \log_2 K$ is an upper bound for the average number of iterations for all $K \geq 6$. Note that $1.5 \log_2 K$ is an absolute upper bound for all the five cases, for all $K \geq 2$. Other interesting observations are: the number of iterations has a tendency to decrease as q increases (see Figs. 11 and 12), and very importantly, it does not exhibit a dependency with N (i.e., independent of the precision of quantizer thresholds).

Recall from the previous section that the number of operations required by one iteration of the Lagrangian-based 2DSQ design algorithm is at most as three iterations of the algorithm of [7]. Since the latter runs in $2K - 1$ iterations, we conclude that for the tested pmf's the RD-guided search is faster than the algorithm of [7] by a factor of $\frac{4K-2}{9\log_2 K}$.

Next we discuss another search strategy, the so-called secant search. This search technique also allows us to deal with the pathological case when I_{opt} consists of a single value.

In the secant search λ is updated as follows: $\lambda_{new} = (W(P_{\lambda_2}) - W(P_{\lambda_1}))/(L(P_{\lambda_1}) - L(P_{\lambda_2}))$. Note that $-\lambda_{new}$ is the slope of the line passing through the planar points $(L(P_{\lambda_1}), W(P_{\lambda_1}))$ and $(L(P_{\lambda_2}), W(P_{\lambda_2}))$. If this line is not a support line of the set $\mathcal{U}$, i.e., it does not include a convex hull edge, then the support line of the same slope

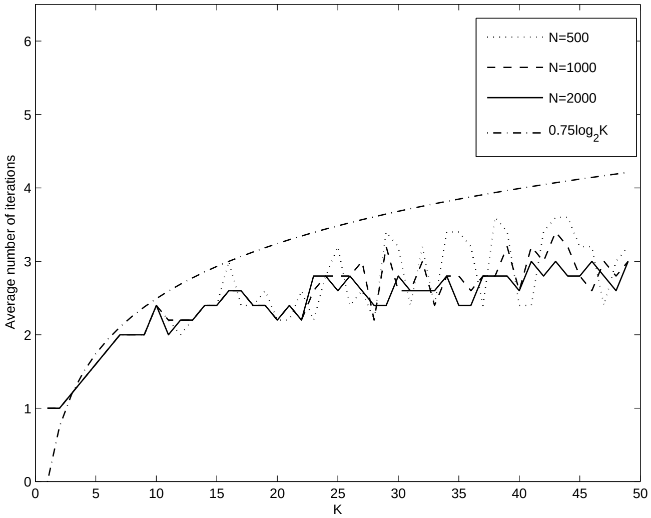


Fig. 6. The number of iterations (average over $q = 0.5, 0.6, \cdots 0.9$) as a function of K, for three values of N, in the case of a discretized zero mean unit variance Gaussian distribution.

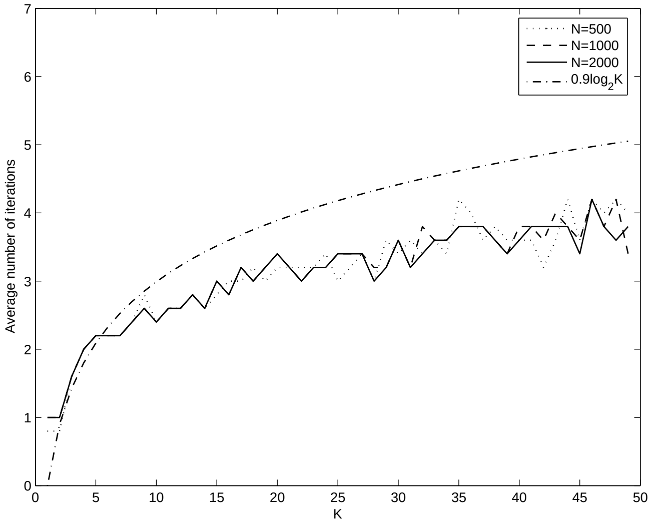


Fig. 7. The number of iterations (average over $q = 0.5, 0.6, \cdots 0.9$) as a function of K, for three values of N, in the case of a discretized zero mean unit variance Laplacian distribution.

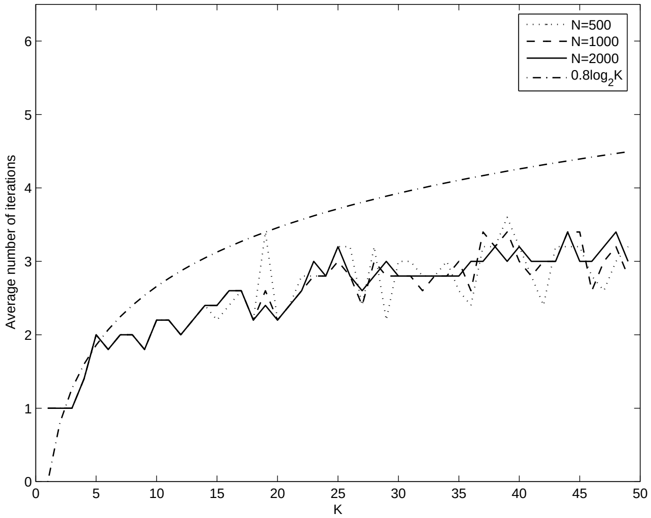


Fig. 8. The number of iterations (average over $q = 0.5, 0.6, \cdots 0.9$) as a function of K, for three values of N, in the case of the discretized mixed Gaussian distribution of pdf of (18).

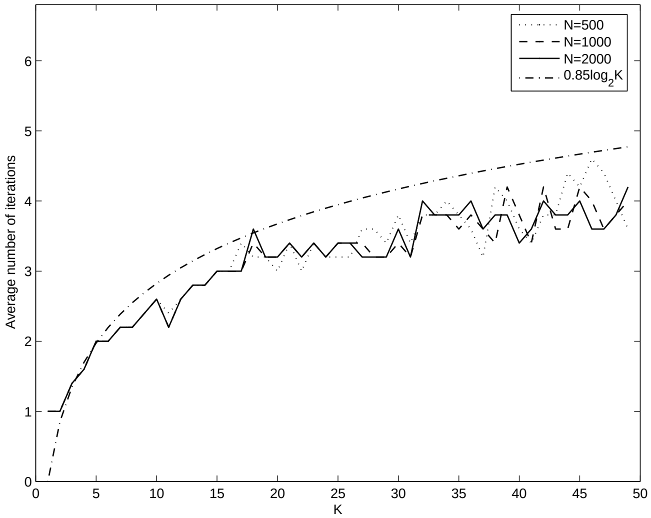


Fig. 9. The number of iterations (average over $q = 0.5, 0.6, \cdots 0.9$) as a function of K, for three values of N, in the case of the discretized mixed Gaussian distribution of pdf of (19).

intersects the set $\mathcal{U}$ at some different point (because two parallel lines may not have points in common). Then the planar point $(L(P_{\lambda_{new}}), W(P_{\lambda_{new}}))$ is different from both $(L(P_{\lambda_1}), W(P_{\lambda_1}))$ and $(L(P_{\lambda_2}), W(P_{\lambda_2}))$. Therefore, it is guaranteed that $L(P_{\lambda_1}) > L(P_{\lambda_{new}}) > L(P_{\lambda_2})$ with both inequalities strict. This implies that the length of $P_{\lambda_{new}}$ is different from all path lengths obtained previously.

If the line passing through the points $(L(P_{\lambda_1}), W(P_{\lambda_1}))$ and $(L(P_{\lambda_2}), W(P_{\lambda_2}))$ is a support line to the set $\mathcal{U}$, then the point $(2K, \bar{W}(2K))$ is on this line as well, hence $I_{opt} = \{\lambda_{new}\}$ according to (13). Thus, the path $P_{\lambda_{new}}$ output by our algorithm will have the length equal to $L(P_{\lambda_1})$, not $2K$. Further, in order to construct the desired $2K$-edge path we use Lemma 2 which is stated and proved in Appendix A. Specifically, we apply iteratively this lemma s times to obtain from the paths P_{λ_2} and P_{λ_1}, two other paths P_1' of $2K$ edges and P_2' of $L(P_{\lambda_1}) + L(P_{\lambda_2}) - 2K$ edges such that the sum of the weights of P_1' and P_2' in $G_{\lambda_{new}}$ is at most equal to the sum of weights of P_{λ_1} and P_{λ_2}. Since both P_{λ_1} and P_{λ_2} are minimum weight paths in $G_{\lambda_{new}}$, P_1' and P_2' are minimum weight paths in $G_{\lambda_{new}}$, too. Then P_1' is our desired path and the algorithm stops. The number s equals the smallest of $|L(P_{\lambda_1}) - 2K|$ and $|2K - L(P_{\lambda_2})|$. This additional step required to construct the optimal path takes at most $O(K^2)$ time. Consequently, it does not change the $O(N^2)$ time complexity per iteration.

To bound the number of iterations required by the secant search, note that the length of the paths $P_{\lambda_{new}}$ is different for different iterations (possibly except the last one). Since in total there are only $2N - 1$ possible path lengths, the number of iterations cannot be larger than $2N$. But this bound is too loose without taking into account the specifics of our optimization problem. We prove as Proposition 4 in Appendix C that for

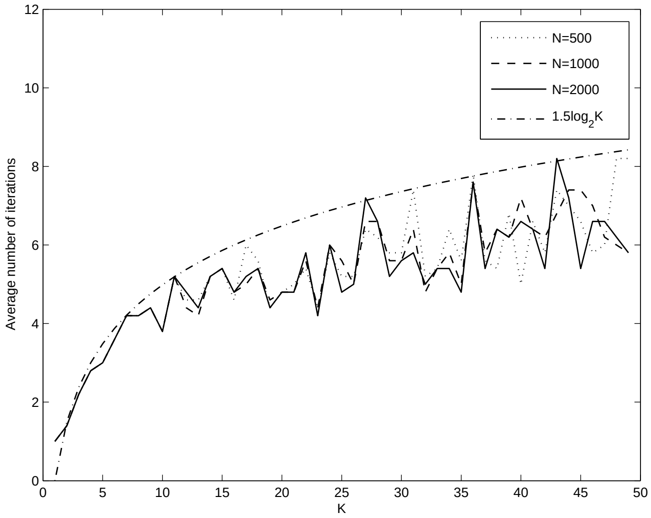


Fig. 10. The number of iterations (average over $q = 0.5, 0.6, \cdots 0.9$) as a function of K, for three values of N, for the real data of the histogram in Fig. 5.

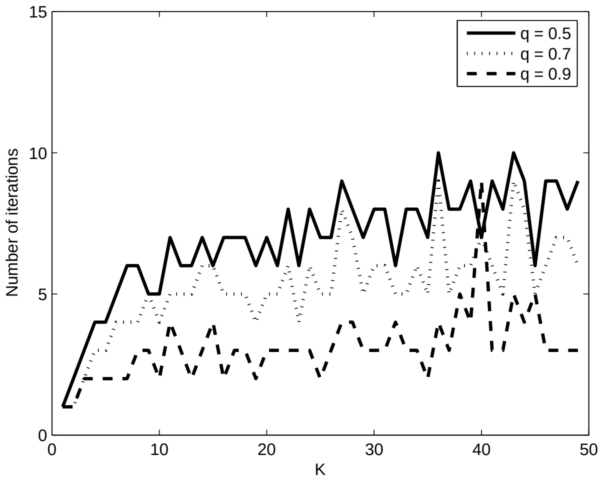


Fig. 12. Comparison of the number of iterations for various channel success probabilities ($q = 0.5, 0.7, 0.9$), for the real data of histogram in Fig. 5, and $N = 2000$.

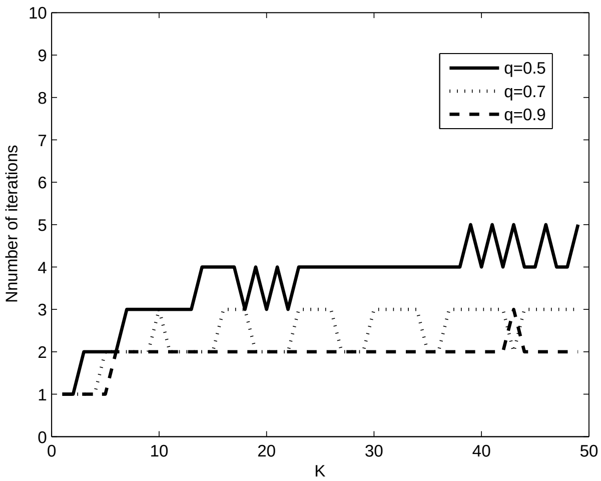


Fig. 11. Comparison of the number of iterations for various channel success probabilities ($q = 0.5, 0.7, 0.9$), for a discretized zero mean unit variance Gaussian distribution and $N = 2000$.

discretizations of continuous distributions and L_2 distortion measure, the number of iterations of the secant search is at most $8K + \lceil \log_{3/2} N \rceil$ for $K << N$.

In practice one can adopt a hybrid method that uses the two search techniques in combination: start with the RD-guided search and then switch to the secant search, only if the desired path length is still not found after some number of iterations (e.g., after $2 \log_2 K$ iterations). According to the results mentioned above, for pmf's obtained by discretizing continuous pdf's, and L_2 distortion measure, the running time of this hybrid method is $O((K + \log N)N^2)$ in the worst case if $K << N$ (as is the case in practice). This suggests an improvement in speed by a factor of $O(\frac{K^2 N}{K + \log N})$ over the algorithm of [15], [16].

In addition, the new algorithm is more efficient in use of memory than the previous ones. Its space complexity is $O(N^2)$ while the algorithms of [15], [16], respectively of [7] have space complexities of $O(K^2 N^2)$, respectively $O(KN^2)$. For modern computers, such a drastic reduction in working space of the algorithm will greatly reduce the probability of cache misses and hence reduce the algorithm running time in practice.

The worst-case time complexity of the proposed Lagrangian-based algorithm seems higher than that of our earlier algorithm [7]. But in all our experiments such worst case behaviour never occurs. Instead, the new algorithm is faster by a factor of $\frac{4K-2}{9 \log_2 K}$.

VII. Codecell Convexity and Optimality

The new 2DSQ design algorithm developed in this paper assumes the convexity of codecells. We now assess the impact of this constraint on the optimality. It is known that the optimal fixed-rate single description quantizer must have convex codecells [9]. On the other hand, this condition may preclude optimality for 2DSQ's [9].

Assume a continuous probability distribution with pdf $f(x)$ defined on a compact interval $[V, W]$ and let the distortion measure be $d(x, y) =\mid x - y \mid^r$. We will use the performance analysis of 2DSQ at high rates, provided in [20] for a family of index assignments with increasing number of diagonals (i.e., increasing number of codecells in the central partition). The analysis of [20] is based on modeling the central quantizer as a compander. Consider 2^R to be the number of codecells in each side partition, let k denote the number of diagonals of the index assignment matrix, and let a be a number in $(0, 1)$ such that $k = 2^{Ra}$. Let $D_0^{(r)}(a)$, respectively $D_1^{(r)}(a)$, $D_2^{(r)}(a)$, denote the central, respectively side 1 and side 2 distortions under the r-th power difference distortion measure, for a 2DSQ with an index assignment matrix with 2^{Ra} diagonals and optimal companding function. According to [20, Eq.

r	$\min \frac{\omega}{\omega_0}$	$\max q$
1	0.25	0.800
2	0.125	0.888
3	0.0625	0.941
4	0.03125	0.969

1	2						
3	4	6					
	5	7	8				
		9	10	11			
				12	14		
				13	15	16	
					17	18	20
						19	21

1	3	6					
2	5	7	9				
4	8	10	12	14			
	11	13	15	17	20		
		16		19	22	25	
			18	21	24	27	28
				23	26	29	31
					30	32	33

Fig. 13. Modified linear index assignment considered by Vaishampayan [18] for two values of the parameter k ($2k$ is the number of diagonals used, other than the main diagonal). Left: $k = 1$. Right: $k = 2$.

(18),(19)], when $R \to \infty$, the following approximations hold

$$D_0^{(r)}(a) \approx \frac{2^{-2r}}{r+1} 2^{-rR(1+a)} (\int_V^W f^{1/(r+1)}(x)dx)^{r+1}. \quad (20)$$

$$D_1^{(r)}(a) = D_2^{(r)}(a) \approx \beta_r 2^{-r-rR(1-a)} (\int_V^W f^{1/(r+1)}(x)dx)^{r+1},$$

for $a \in (0,1)$, where β_r was defined as

$$\beta_r = \lim_{k\to\infty} \frac{\sum_{i=1}^k i^r}{k^{r+1}}.$$

Note that

$$\beta_r = \lim_{k\to\infty} \sum_{i=1}^k \frac{1}{k} (\frac{i}{k})^r = \int_0^1 x^r dx = \frac{1}{r+1}.$$

Thus, relation (21) becomes

$$D_1^{(r)}(a) = D_2^{(r)}(a) \approx \frac{2^{-rR(1-a)}}{2^r(r+1)} (\int_V^W f^{1/(r+1)}(x)dx)^{r+1} \quad (2$$

Using (20) and (22), we obtain the expected distortion of t balanced 2DSQ at high rates

$$\bar{D}^{(r)}(a) \approx \frac{2\omega 2^{rRa} + \omega_0 2^{-r} 2^{-rRa}}{2^{r+rR}(r+1)} (\int_V^W f^{1/(r+1)}(x)dx)^{r+} \quad (2$$

for $a \in (0,1)$. The analysis provided in [20] does r directly apply to the case with convex codecells. Howev as proved in Appendix B (Eq. (46)) with arguments along t lines of [4], [3], relation (23) holds in the case of conv side quantizers, too, with the corresponding value $a =$ Consequently, the convex 2DSQ is optimum at high ra if and only if $\min_{a\in[0,1)} \bar{D}^{(r)}(a) = \bar{D}^{(r)}(0)$. Minimizi $\bar{D}^{(r)}(a)$ is equivalent to minimizing the function $F(a)$

$$F(a) = 2\omega 2^{rRa} + \omega_0 2^{-r} 2^{-rRa},$$

on the interval $[0,1)$. $F(a)$ is a convex function $[0,1)$, and its unique minimum is achieved at the po... $a_0 = \max\{0, \frac{1}{2rR} \log_2 \frac{\omega_0}{2^{r+1}\omega}\}$. It follows that the necessary and sufficient condition for a_0 to be 0 is $\frac{\omega_0}{2^{r+1}\omega} \leq 1$, or $\omega/\omega_0 \geq 1/2^{r+1}$. In the case when the 2DSQ is designed for communication over two independent channels, each with probability of success q , we have $\omega = q(1-q)$ and $\omega_0 = q^2$, hence the above inequality is equivalent to $q \leq \frac{2^{r+1}}{2^{r+1}+1}$.

In conclusion, the above arguments show that asymptotically in R, the convex-codecell condition does not preclude optimality when the channel probability of success is at most $\frac{2^{r+1}}{2^{r+1}+1}$, for continuous distributions and r-th power distortion measure. Table 1 lists the value of this maximum bound for several values of r. For $r = 2$ the codecell convexity will not preclude optimality if the channel has a failure rate of 12% or higher. The larger the value of r, the more relaxed the condition for the side quantizers of optimal 2DSQ to be convex.

Next we present experimental evidence to the intuition that for poor channels codecell convexity does not prevent optimality. We applied the algorithm of [18] to a memoryless, unit-variance Gaussian source to optimize the 8-level balanced 2DSQ. As a measure of codecell convexity, index assignments of different spreads were tested. The spread of the index assignment is defined in [18] as the largest number of central codecells situated between the extreme points of a side codecell. We have considered the modified linear index assignment proposed by Vaishampayan [18], with parameter $k = 1, 2, \cdots 6$, where $2k$ is the number of diagonals other than the main diagonal used in the index assignment (Fig. 13). As k increases the spread of the index assignment increases, or further deviates from convexity.

For each q we minimized $\bar{D}(\mathbf{Q})$ for each k and then took the minimum over all k. Our results showed that for $q \in (0, 0.925]$ the minimal expected distortion was always achieved for $k = 1$. Note that the algorithm of [18] can change the initial index assignment in the iterative design process. It may start with a certain assignment and end up in a different one, because some of the index pairs may be allocated empty codecells in the central partition (also some of index pairs may change their order). We observed that for $q \in (0, 0.925]$, when starting with the modified index assignment of spread $k = 1$, which is not convex as shown in Fig. 13, the algorithm consistently converged to an assignment yielding a convex 2DSQ.

We have also applied the proposed algorithm to a discretized version of a memoryless, unit-variance Gaussian source. For $q \in (0, 0.925]$ the minimum values of $\bar{D}(\mathbf{Q})$ match those obtained by applying the iterative algorithm of Vaishampayan [18] for a series of index assignments of various spreads and taking the minimum over all these index assignments (Fig. 14). But for $q \in (0.925, 1]$, assignments of higher spreads lead to lower expected distortion (see Fig. 15). As expected, the codecell convexity compromises the optimality when channel conditions are very good.

However, there are also cases where the proposed algorithm outperforms the locally optimal algorithm of [18]. Let us consider three examples of mixed Gaussian distribution. The

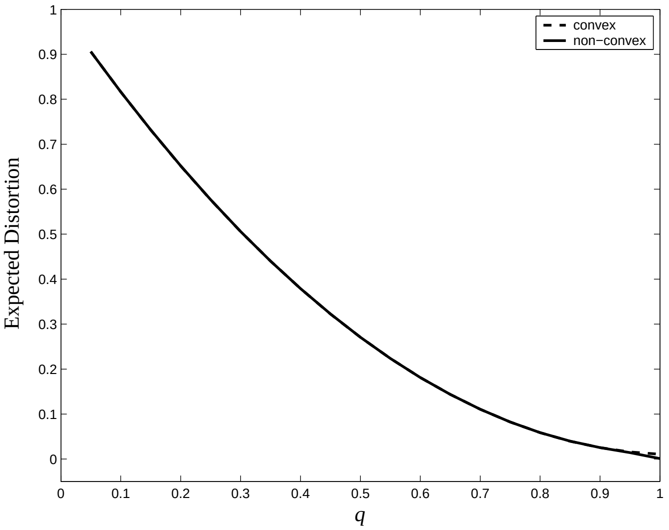


Fig. 14. Minimal expected distortion for memoryless Gaussian source at side rate $R = 3$ ($K = 8$), as a function of $q \in (0, 1]$. The dotted curve is the performance of the proposed algorithm; the solid curve is the performance of the algorithm of [18] (the best result over a large set of index assignments is plotted). For $q \in (0, 0.925]$ the two curves are identical.

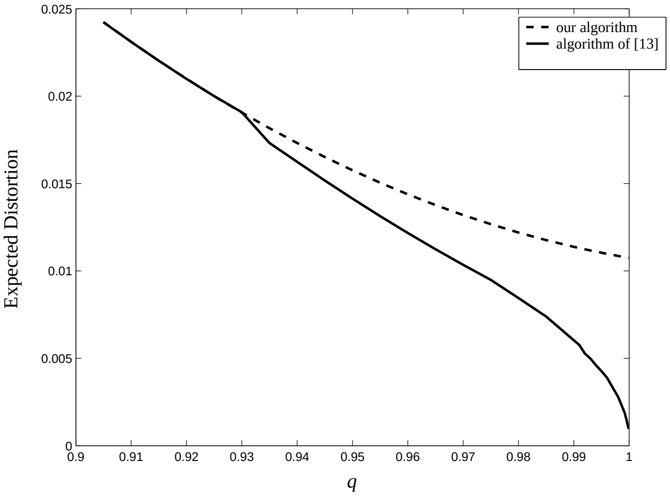


Fig. 15. The magnification of Fig. 14 in the range $q \in [0.9, 1]$.

pdf's of the three distributions are

$$f_1(x) = 1/2g(0, 1/16) + 1/2g(6, 1), \tag{24}$$

$$f_2(x) = 1/2g(0, 1/4) + 1/2g(6, 1), \tag{25}$$

$$f_3(x) = 1/4g(0, 1/16) + 3/4g(6, 1), \tag{26}$$

where $g(\mu, \sigma^2)$ denotes the pdf of the normal distribution of mean μ and variance σ^2. We applied the algorithm of [18] to minimize the expected distortion for $K = 4$, using the modified linear index assignment of spread $k = 1$. We also applied the proposed algorithm to a discretized version of each of the three mixtures. The results obtained for $q = 0.9$, respectively $q = 0.5$, are recorded in Table 2, respectively Table 3. To visualize an example we plot in Fig. 16 the histogram of the mixed Gaussian distribution given by (26), and the thresholds of the central partition obtained by the two algorithms.

distribution	D [18]	D proposed algorithm	relative difference
f_1	0.2149	0.1813	18.53%
f_2	0.2338	0.2224	5.12%
f_3	0.1900	0.1684	12.82%

TABLE II
PERFORMANCE COMPARISON BETWEEN THE ALGORITHM OF [18] AND OURS ON THE MIXED GAUSSIAN DISTRIBUTIONS GIVEN BY (24)-(26), FOR $K = 4$ AND $q = 0.9$.

distribution	D [18]	D proposed algorithm	relative difference
f_1	2.7178	2.5855	5.11%
f_2	2.7349	2.6397	3.6%
f_3	2.1282	2.0423	4.2%

TABLE III
PERFORMANCE COMPARISON BETWEEN THE ALGORITHM OF [18] AND OURS ON THE MIXED GAUSSIAN DISTRIBUTIONS GIVEN BY (24)-(26), FOR $K = 4$ AND $q = 0.5$.

VIII. CONCLUSION

We show that optimal balanced fixed-rate two-description scalar quantizer design can be treated as a Lagrangian-type optimization problem, if convexity of side quantizer codecells is assumed. It turns out that for a very large class of distortion measures and for any given target rate, the Lagrangian multiplier exists for the globally optimal solution, under the above specified constraint. By exploiting a monotonicity of the objective function we develop a fast dynamic programming technique to solve the parameterized problem given a trial Lagrangian multiplier. Furthermore, an RD-guided search technique is also proposed. It makes the Lagrangian optimization process to converge in a small number of iterations in our experiments. The relationship between codecell convexity and optimality is also discussed.

Appendix A. Proofs of Propositions 2 and 3

Our development hinges on the Monge propriety satisfied by the function $D(a, b]$, namely

$$D(a, b] + D(a', b'] \le D(a, b'] + D(a', b], \quad \text{for all } a \le a' \le b \le b'. \tag{27}$$

It was proved in [24, Lemma 4] that the above relation holds for monotone distortion functions as defined by (4). We will also use the following lemma, which was established in [7, Lemma 1]. The proof of this lemma can be done simply by replacing the weights of the edges and then applying (27).

Lemma 1. Let νa, ab, $\nu' a'$ and $a'b'$ be nodes in the graph G, such that $\nu \le \nu'$, $a \le a'$, $b \le b'$, $\nu < b$ and $\nu' < b'$. Then the following assertions hold:

i) if $\nu' \le a$ and $\nu' < b$, then $w(\nu a, ab) + w(\nu' a', a'b') \le w(\nu' a, ab) + w(\nu a', a'b')$;
ii) if $a' \le b$ and $\nu' < b$, then $w(\nu a, ab) + w(\nu' a', a'b') \le w(\nu a, ab') + w(\nu' a', a'b)$;
iii) if $\nu' \le a$ and $a' \le b$, then $w(\nu a, ab) + w(\nu' a', a'b') \le w(\nu' a, ab') + w(\nu a', a'b)$.

In order to prove Proposition 2, we additionally need the following lemma.

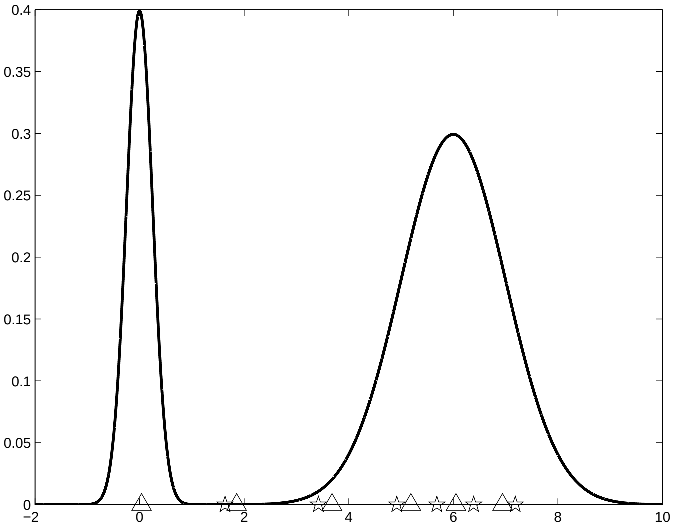
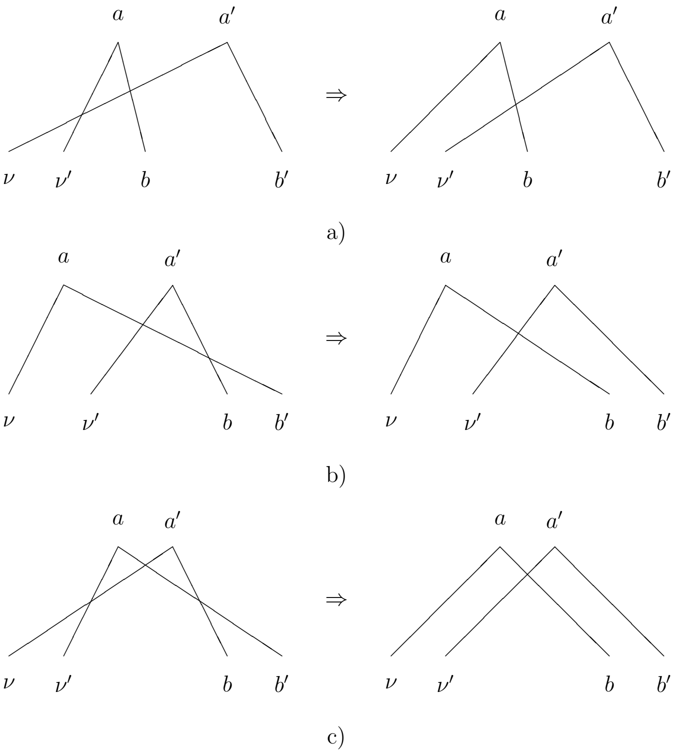


Fig. 16. The mixed Gaussian distribution (26) and the thresholds of the central partition obtained by the algorithm of [18] (triangles) and our algorithm (stars). In both cases $K = 4$.



a)

b)

c)

Fig. 17. Illustration of Lemma 1. By replacing the two edges to the left by the two edges to the right, the total weight decreases. a) Lemma 1 i); b) Lemma 1 ii); c) Lemma 1 iii).

Lemma 2. For any integers l, l' such that $0 \le l-1 < l'+1$, and any two paths in G, P_1 with $l-1$ edges, from 00 to some vertex $a_{l-1}a_l$, and P_2 with $l'+1$ edges from 00 to $b_{l'+1}b_{l'+2}$, $a_{l-1} \ge b_{l'+1}$, $a_l \ge b_{l'+2}$, there are two other paths P_1' with l edges from 00 to $b_{l'+1}b_{l'+2}$, and P_2' with l' edges from 00 to $a_{l-1}a_l$, such that

$$W(P_1') + W(P_2') \le W(P_1) + W(P_2). \tag{28}$$

Proof. Let the paths P_1 and P_2 be

$$\begin{aligned} &P_1 : a_0a_1, a_1a_2, a_2a_3, \cdots, a_{l-2}a_{l-1}, a_{l-1}a_l,\\ &P_2 : b_0b_1, b_1b_2, b_2b_3, \cdots, b_{l'}b_{l'+1}, b_{l'+1}b_{l'+2}, \end{aligned}$$

where $0 = a_0 = a_1 < a_2 < \cdots < a_{l-1} \le a_l \le N$ and $0 = b_0 = b_1 < b_2 < \cdots < b_{l'+1} \le b_{l'+2} \le N$.

Let i be the largest integer between 0 and $l-1$ for which both inequalities $a_i \le b_{i+l'-l+1}, a_{i+1} \le b_{i+l'-l+2}$ hold. Such an integer exists because the inequalities are satisfied for $i = 0$. Clearly, $i < l-2$. Let also j be the smallest integer between 0 and $l-1-i$ for which both inequalities $a_{i+j} \ge b_{i+j+l'-l+1}, a_{i+j+1} \ge b_{i+j+l'-l+2}$ hold. Obviously, such an integer exists because for $j = l-1-i$ the inequalities are satisfied. We distinguish between three cases: $j = 0$, $j = 1$ and $j \ge 2$. We start with the most general case: $j \ge 2$.

The definitions of i and j imply, on one hand, that $a_i \le b_{i+l'-l+1}, a_{i+1} < b_{i+l'-l+2}, a_{i+j} > b_{i+j+l'-l+1}, a_{i+j+1} \ge b_{i+j+l'-l+2}$, and on the other hand, that $a_{i+1} < b_{i+l'-l+2}$, $a_{i+2} > b_{i+l'-l+3}$, $a_{i+3} < b_{i+l'-l+4}$, $a_{i+4} > b_{i+l'-l+5}$, $\cdots, a_{i+j-1} < b_{i+j+l'-l}$. In other words, $a_{i+1} < b_{i+l'-l+2} \le b_{i+l'-l+3} \le a_{i+2} \le a_{i+3} < b_{i+l'-l+4} < b_{i+l'-l+5} < a_{i+4} < \cdots < b_{i+j+l'-l-1} < a_{i+j-2} < a_{i+j-1} < b_{i+j+l'-l} < b_{i+j+l'-l+1} < a_{i+j}$. Clearly, j must be an even integer.

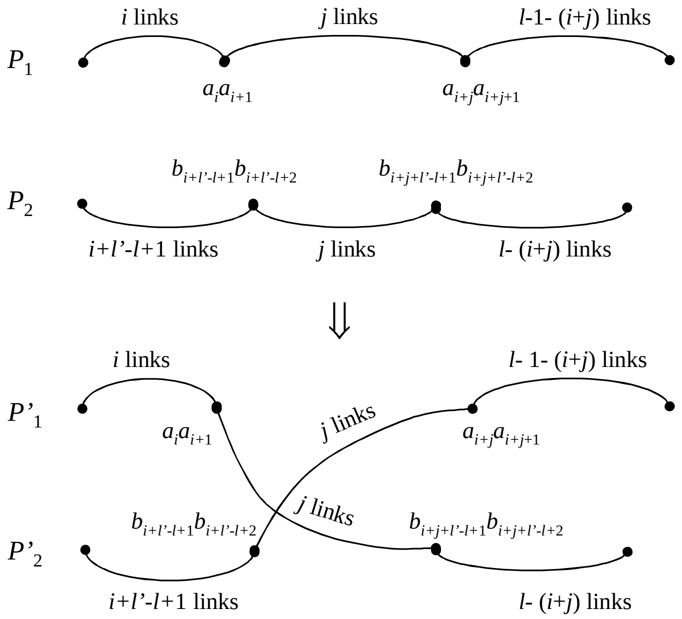
We construct the new paths P_1' of l edges, and P_2' of l' edges, in the following way. P_1' connects the source with $b_{l'+1}b_{l'+2}$ via the nodes a_ia_{i+1} and $b_{i+j+l'-l+1}b_{i+j+l'-l+2}$. The edges up to the node a_ia_{i+1} are the first i edges of P_1, and the edges from the node $b_{i+j+l'-l+1}b_{i+j+l'-l+2}$ are the last $l-i-j$ edges of P_2. The vertices a_ia_{i+1} and $b_{i+j+l'-l+1}b_{i+j+l'-l+2}$ are connected by the following j-edge path:

$$\begin{gathered} a_ia_{i+1}, a_{i+1}b_{i+l'-l+3}, b_{i+l'-l+3}a_{i+3}, a_{i+3}b_{i+l'-l+5},\\ b_{i+l'-l+5}a_{i+5}, \cdots, b_{i+j+l'-l-1}a_{i+j-1},\\ a_{i+j-1}b_{i+j+l'-l+1}, b_{i+j+l'-l+1}b_{i+j+l'-l+2}. \end{gathered} \tag{29}$$

P_2' connects the source with $a_{l-1}a_l$ via the nodes $b_{i+l'-l+1}b_{i+l'-l+2}$ and $a_{i+j}a_{i+j+1}$. The edges up to the node $b_{i+l'-l+1}b_{i+l'-l+2}$ are the first $i+l'-l+1$ edges of P_2, and the edges from the node $a_{i+j}a_{i+j+1}$ are the last $l-i-1-j$ edges of P_1. The vertices $b_{i+l'-l+1}b_{i+l'-l+2}$ and $a_{i+j}a_{i+j+1}$ are connected by the following j-edge path:

$$\begin{gathered} b_{i+l'-l+1}b_{i+l'-l+2}, b_{i+l'-l+2}a_{i+2}, a_{i+2}b_{i+l'-l+4},\\ b_{i+l'-l+4}a_{i+4}, \cdots, a_{i+j-2}b_{i+j+l'-l},\\ b_{i+j+l'-l}a_{i+j}, a_{i+j}a_{i+j+1}. \end{gathered} \tag{30}$$

Note that the new paths P_1' and P_2' are obtained from the old ones by interchanging a_{i+2k} and $b_{i+2k+l'-l+1}$, for all $k, 1 \le k \le j/2$, and by interchanging a_{i+j+1} and $b_{i+j+l'-l+2}$. In order to establish (28), it is sufficient to show that the sum of the weights of the two paths (29) and (30) is at most equal to the sum of weights of the edges of P_1 between a_ia_{i+1} and $a_{i+j}a_{i+j+1}$, and of the edges of P_2 between $b_{i+l'-l+1}b_{i+l'-l+2}$ and $b_{i+j+l'-l+1}b_{i+j+l'-l+2}$. Further, for



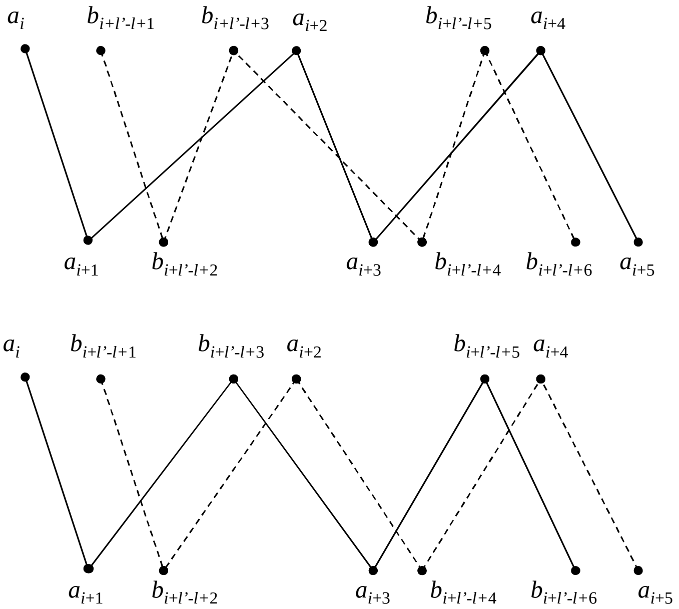


Fig. 19. Change of j-edge subpaths - case $j = 4$. Up: Old subpaths. Down: New subpaths.

this it is enough to show that the following inequalities hold:

$$\begin{gathered} w(a_i a_{i+1}, a_{i+1} b_{i+l'-l+3}) + \\ w(b_{i+l'-l+1} b_{i+l'-l+2}, b_{i+l'-l+2} a_{i+2}) \leq \\ w(a_i a_{i+1}, a_{i+1} a_{i+2}) + \\ w(b_{i+l'-l+1} b_{i+l'-l+2}, b_{i+l'-l+2} b_{i+l'-l+3}); \qquad (31) \\ w(b_{i+j+l'-l} a_{i+j}, a_{i+j} a_{i+j+1}) + \\ w(a_{i+j-1} b_{i+j+l'-l+1}, b_{i+j+l'-l+1} b_{i+j+l'-l+2}) \leq \\ w(a_{i+j-1} a_{i+j}, a_{i+j} a_{i+j+1}) + \\ w(b_{i+j+l'-l} b_{i+j+l'-l+1}, b_{i+j+l'-l+1} b_{i+j+l'-l+2}); \qquad (32) \\ w(a_{i+k-1} b_{i+k+1}, b_{i+k+1} a_{i+k+1}) + \\ w(b_{i+k+l'-l} a_{i+k}, a_{i+k} b_{i+k+l'-l+2}) \leq \end{gathered}$$

for all $k, 2 \leq k \leq j-1$. The relation (31) follows from Lemma 1 ii), the relation (32) follows from Lemma 1 i), and the relation (33) follows from Lemma 1 iii). Now the proof of case $j \geq 2$ is complete.

Let us consider now the case $j = 1$. This case implies that $a_{i+1} = b_{i+l'-l+2}$. Moreover, $a_i \leq b_{i+l'-l+1} < a_{i+1} = b_{i+l'-l+2} < b_{i+l'-l+3} \leq a_{i+2}$. The new paths P_1' and P_2' are constructed in a similar way. The only difference consists in that the nodes $a_i a_{i+1}$ and $b_{i+j+l'-l+1} b_{i+j+l'-l+2}$ are connected in P_1' by a single edge (this is possible since $a_{i+1} = b_{i+j+l'-l+1}$ and $a_i < b_{i+j+l'-l+2}$) and the nodes $b_{i+l'-l+1} b_{i+l'-l+2}$ and $a_{i+j} a_{i+j+1}$ are connected in P_2' again by a single edge (possible because $b_{i+l'-l+2} = a_{i+j}$ and $b_{i+l'-l+1} < a_{i+j+1}$). In order to establish relation (28) it is sufficient to show the inequality

$$\begin{gathered} w(a_i a_{i+1}, a_{i+1} b_{i+l'-l+3}) + w(b_{i+l'-l+1} a_{i+1}, a_{i+1} a_{i+2}) \leq \\ w(a_i a_{i+1}, a_{i+1} a_{i+2}) + w(b_{i+l'-l+1} a_{i+1}, a_{i+1} b_{i+l'-l+3}), \end{gathered}$$

which follows from Lemma 1, i) (it corresponds to the case when $a = a'$).

The case $j = 0$ is the simplest one. In this case the nodes $a_i a_{i+1}$ and $b_{i+l'-l+1} b_{i+l'-l+2}$ coincide. Hence the paths P_1 and P_2 have this node in common. This common node partitions each of the two paths in two subpaths (a prefix and a suffix). The new paths P_1' and P_2' are obtained starting from the old ones and interchanging the two suffixes. Relation (28) is trivially satisfied with equality. □

Proposition 2. The inequality

$$2\bar{W}(l) \leq \bar{W}(l-1) + \bar{W}(l+1)$$

holds for all $l, 3 \leq l \leq 2N-1$, where $\bar{W}(l)$ is the weight of the minimum-weight l-edge path from the source to the final node in G.

Proof. Let P_1 and P_2 be the minimum weight $(l-1)$-link path and the minimum weight $(l+1)$-link path, respectively, from the source to the final node in G. According to Lemma 2, there are two paths P_1' and P_2' both of l-links, from the source to the final node in G, such that relation (28) holds. Then the conclusion of Proposition 2 trivially follows. □

In order to establish Proposition 3, we need the following lemma.

Lemma 3. The function $W_\lambda(a, b)$ satisfies the Monge condition, i.e.,

$$\begin{gathered} W_\lambda(a, b) + W_\lambda(a', b') \leq W_\lambda(a, b') + W_\lambda(a', b) \\ \text{for all } 0 \leq a \leq a' \leq b \leq b' \leq N. \qquad (34) \end{gathered}$$

Proof. The nontrivial case is when $0 \leq a < a' \leq b < b'$. The proof proceeds by induction on b'. The base case is $b' = 2$. Then $a = 0$ and $a' = b = 1$, and (34) becomes

$$W_\lambda(0,1) + W_\lambda(1,2) \leq W_\lambda(0,2) + W_\lambda(1,1) \qquad (35)$$

For each of the nodes 01, 02 and 11 there is only one path ending in that node, hence that is the minimum-weight path. These paths are, respectively, $00, 01$; $00, 02$; $00, 01, 11$. It is sufficient if we prove the inequality in which $W_\lambda(1,2)$ is replaced by the weight of some path ending in the node 12 (not necessarily the minimum-weight one), for instance, the path: $00, 01, 12$. Hence, a sufficient condition for (35) is the following

$$w(00,01)+w(00,01)+w(01,12)+3\lambda \leq \\ w(00,02)+w(00,01)+w(01,11)+3\lambda,$$

which is obviously satisfied with equality.

Now we show the inductive step $b'-1 \to b'$. Denote by ξ, the value $\xi_\lambda(a,b')$ and by ξ', the value $\xi_\lambda(a',b)$. Applying the definition of $\xi_\lambda(a,b')$ and $\xi_\lambda(a',b)$, we obtain that

$$W_\lambda(a,b') = W_\lambda(\xi,a)+w(\xi a, ab')+\lambda, \tag{36}$$

$$W_\lambda(a',b) = W_\lambda(\xi',a')+w(\xi' a', a'b)+\lambda. \tag{37}$$

Further we need to distinguish between the cases $\xi \geq \xi'$ and $\xi < \xi'$.

Case $\xi \geq \xi'$. Since $\xi \leq a < a'$, it follows that $\xi' \leq a$ and $\xi < a'$. The definitions of $W_\lambda(a,b)$ and $W_\lambda(a',b')$ imply, respectively, that

$$W_\lambda(a,b) \leq W_\lambda(\xi',a)+w(\xi' a, ab)+\lambda, \tag{38}$$

$$W_\lambda(a',b') \leq W_\lambda(\xi,a')+w(\xi a', a'b')+\lambda. \tag{39}$$

Note that $a' \leq b < b'$, hence $a' \leq b'-1$. Consequently, since $\xi' \leq \xi \leq a < a'$, the inequality

$$W_\lambda(\xi',a)+W_\lambda(\xi,a') \leq W_\lambda(\xi,a)+W_\lambda(\xi',a')$$

is valid according to the inductive hypothesis. Also Lemma 1, iii) can be applied for $\nu = \xi'$ and $\nu' = \xi$. Combining these two results, we obtain that the sum of the righthand sides of inequalities (38) and (39) is smaller or equal than the sum of the righthand sides of equalities (36) and (37). This implies that (34) is satisfied, too.

Case $\xi < \xi'$. The proof follows the same idea. In (38) and (39), ξ and ξ' are interchanged. Then Lemma 1, ii) is applied to reach the conclusion.□

Proposition 3. For any nodes ab and $a'b'$ other than the source, and such that $a \leq a'$ and $b \leq b'$, the following relation holds:

$$\xi_\lambda(a,b) \leq \xi_\lambda(a',b').$$

Proof. Assume that

$$\xi_\lambda(a,b) > \xi_\lambda(a',b').$$

We will show that this assumption leads to a contradiction. Let $\nu = \xi_\lambda(a',b')$ and $\nu' = \xi_\lambda(a,b)$. Then

$$W_\lambda(a,b) = W_\lambda(\nu',a)+w(\nu' a, ab)+\lambda, \tag{40}$$

$$W_\lambda(a',b') = W_\lambda(\nu,a')+w(\nu a', a'b')+\lambda. \tag{41}$$

Note that $\nu \leq a'$ and $\nu < b'$. Also $\nu' \leq a$ and $\nu' < b$. Using the inequality $\nu < \nu'$, we obtain that $\nu \leq a$ and $\nu < b$. Furthermore, $\nu' \leq a'$ and $\nu < b'$. These imply that

$$W_\lambda(a,b) \leq W_\lambda(\nu,a)+w(\nu a, ab)+\lambda, \tag{42}$$

$$W_\lambda(a',b') \leq W_\lambda(\nu',a')+w(\nu' a', a'b')+\lambda. \tag{43}$$

Since the function $W_\lambda(\cdot,\cdot)$ satisfies the Monge condition (Lemma 3), it follows that

$$W_\lambda(\nu,a)+W_\lambda(\nu',a') \leq W_\lambda(\nu',a)+W_\lambda(\nu,a').$$

Moreover, Lemma 1, i) can be applied, and combining these two observations, yields that the sum of the righthand sides of inequalities (42) and (43), which we denote by $\mathcal{A}$, is smaller or equal than the sum of the righthand sides of equalities (40) and (41), which we denote by $\mathcal{B}$. Further, we obtain that

$$W_\lambda(a,b)+W_\lambda(a',b') \leq \mathcal{A} \leq \mathcal{B} \leq W_\lambda(a,b)+W_\lambda(a',b').$$

The conclusion is that both relations (42) and (43) are satisfied with equality. But equality in (43) contradicts the definition of $\xi_\lambda(a',b')$. □

Lemma 4. The path computed by the algorithm of Section 5 is a minimum-weight path in $G(\lambda)$, with the largest number of edges.

Proof. Let

$$P : a_0a_1, a_1a_2, a_2a_3, \cdots, a_{l-1}a_l, a_la_{l+1}$$

be the path computed by the algorithm of Section 5. It has the property that $a_l = a_{l+1} = N$, $a_0 = a_1 = 0$ and $a_i = \xi_\lambda(a_{i+1}, a_{i+2})$ for all $0 \leq i \leq l-1$. P clearly has l edges. Assume now that there is another path P' from the source to the final node, which is also a minimum-weight path in $G(\lambda)$, and it has l' edges with $l' > l$. Let P' be

$$P' : b_0b_1, b_1b_2, b_2b_3, \cdots, b_{l'-1}b_{l'}, b_{l'}b_{l'+1},$$

where $b_0 = b_1 = 0$ and $b_{l'} = b_{l'+1} = N$.

We show first by using an inductive argument, that $b_{l'+1-j} \leq a_{l+1-j}$ for all $j, 0 \leq j \leq l+1$.

Base Step. For $j=0$ and $j=1$ the inductive hypothesis is trivially satisfied with equality.

Inductive Step. Let j be an arbitrary integer such that $2 \leq j < l+1$. Assume that the inductive hypothesis is satisfied for all integers from 0 to $j-1$ inclusively. Since $b_{l'+1-(j-2)} \leq a_{l+1-(j-2)}$ and $b_{l'+1-(j-1)} \leq a_{l+1-(j-1)}$, it follows by Proposition 3 that $\xi_\lambda(b_{l'+1-(j-1)}, b_{l'+1-(j-2)}) \leq \xi_\lambda(a_{l+1-(j-1)}a_{l+1-(j-2)}) = a_{l+1-j}$. The last equality is due to the definition of the path P. Further, since the prefix of the path P' up to the node $b_{l'+1-(j-1)}b_{l'+1-(j-2)}$ is a minimum weight path in $G(\lambda)$ from the source to that node, it follows by the definition of $\xi_\lambda(b_{l'+1-(j-1)}, b_{l'+1-(j-2)})$ that $b_{l'+1-j} \leq \xi_\lambda(b_{l'+1-(j-1)}, b_{l'+1-(j-2)})$. Now the inequality $b_{l'+1-j} \leq a_{l+1-j}$ follows and the inductive proof is over.

For $j = l$ we have thus, $b_{l'+1-l} \leq a_1 = 0$. On the other hand, $l' > l$ implies $b_{l'+1-l} \geq b_2 > 0$, yielding a contradiction. □

Appendix B. Asymptotic Analysis at High Resolution

Assume a probability distribution with continuous pdf $f(x)$ defined on the compact interval $[V,W]$. Any K-level balanced convex 2DSQ satisfying the relation stated by Proposition 1 is completely specified by the central partition. Following the approach of [20] we model the central quantizer as a compander. In other words, the central quantizer Q_0 is the composition of three functions, $Q_0 = G^{-1} \circ Q^u_{2K-1} \circ G$, where Q^u_{2K-1} is the $2K-1$ level uniform quantizer on the interval $[0,1]$, G is an invertible function, $G : [V,W] \rightarrow [0,1]$, and h is its inverse, $h = G^{-1}$, $h : [0,1] \rightarrow [V,W]$. Function G is moreover assumed to be strictly increasing and differentiable, hence its derivative $g = G'$ is positive. Furthermore g is assumed to be continuous. Therefore, the balanced 2-DSQ of rate R is defined by the central partition with thresholds $t_0 = V < t_1 < t_2 < \cdots < t_{2K-2} < t_{2K-1} = W$, where $K = 2^R$ and $t_i = h(i/(2K-1))$ for all $0 \le i \le 2K-1$.

Then the thresholds of Q_1 are $t_0, t_1, t_3, t_5, \cdots, t_{2K-3}, t_{2K-1}$, and those of Q_2 are $t_0, t_2, t_4, \cdots, t_{2K-4}, t_{2K-2}, t_{2K-1}$. Consider the r-th power difference as distortion measure, and denote by $D_r(Q)$ the distortion of a quantizer Q, and by $D_r(a,b]$ the distortion of a codecell $(a,b]$. According to [4, Eq. (1.6)] and [3, Eq. (8)], the following approximation holds as $R \rightarrow \infty$

$$D^{(r)}(Q_0) \approx \frac{1}{(2K-1)^r 2^r (r+1)} \int_V^W \frac{f(x)}{g^r(x)} dx.$$

Further, since $\lim_{K\rightarrow\infty} \frac{(2K-1)^r}{K^r} = 2^r$ and $K = 2^R$ we approximate $(2K-1)^r$ by $2^{Rr}2^r$, which leads to

$$D^{(r)}(Q_0) \approx \frac{1}{2^{Rr} 2^{2r} (r+1)} \int_V^W \frac{f(x)}{g^r(x)} dx. \tag{44}$$

Similar approximations for the side quantizers distortions do not follow directly from [4], [3] because the side quantizers are not companders. Precisely, the function G maps each side quantizer Q_k to a quantizer over the interval $[0,1]$ which has all codecells except one, of equal length. But such approximations can be derived very easily following the same ideas as in [4], [3]. We illustrate this for the side quantizer Q_1. Note that

$$D^{(r)}(Q_1) = D^{(r)}[t_0, t_1] + \sum_{i=1}^{K-1} D^{(r)}(t_{2i-1}, t_{2i+1}].$$

Assuming that for high K the distribution is approximately uniform over each codecell, it follows that the representation point can be approximated by the midpoint and obtain

$$\begin{gathered} D^{(r)}(t_{2i-1}, t_{2i+1}] \approx \\ f(t_{2i-1}) \textstyle\int_{t_{2i-1}}^{t_{2i+1}} |x - (t_{2i-1} + t_{2i+1})/2|^r dx = \\ f(t_{2i-1}) \tfrac{(t_{2i+1} - t_{2i-1})^{r+1}}{2^r (r+1)}. \end{gathered}$$

Further, using the mean value theorem and making another approximation, we obtain

$$\begin{gathered} t_{2i+1} - t_{2i-1} = h(\tfrac{2i+1}{2K-1}) - h(\tfrac{2i-1}{2K-1}) \approx \\ h'(\tfrac{2i-1}{2K-1}) \tfrac{2}{2K-1} = \tfrac{2}{2K-1} \cdot \tfrac{1}{g(t_{2i-1})}. \end{gathered}$$

Thus, we get

$$D^{(r)}(t_{2i-1}, t_{2i+1}] \approx \frac{t_{2i+1} - t_{2i-1}}{(r+1)(2K-1)^r} \frac{f(t_{2i-1})}{g^r(t_{2i-1})}.$$

Similarly, it follows that

$$D^{(r)}[t_0, t_1] \approx \frac{1}{2^r (r+1)(2K-1)^r} (t_1 - t_0) \frac{f(t_0)}{g^r(t_0)}.$$

Then we have that

$$\begin{gathered} D^{(r)}(Q_1) \approx \tfrac{1}{(r+1)(2K-1)^r} ((t_1 - t_0) \tfrac{f(t_0)}{g^r(t_0)} + \\ \textstyle\sum_{i=1}^{K-1} (t_{2i+1} - t_{2i-1}) \tfrac{f(t_{2i-1})}{g^r(t_{2i-1})}) - \\ \tfrac{2^r - 1}{2^r (r+1)(2K-1)^r} (t_1 - t_0) \tfrac{f(t_0)}{g^r(t_0)}. \end{gathered}$$

By multiplying with K^r we obtain that

$$\begin{gathered} K^r D^{(r)}(Q_1) \approx \tfrac{K^r}{(r+1)(2K-1)^r} [(t_1 - t_0) \tfrac{f(t_0)}{g^r(t_0)} + \\ \textstyle\sum_{i=1}^{K-1} (t_{2i+1} - t_{2i-1}) \tfrac{f(t_{2i-1})}{g^r(t_{2i-1})}] - \\ \tfrac{(2^r - 1)K^r}{2^r (r+1)(2K-1)^r} (t_1 - t_0) \tfrac{f(t_0)}{g^r(t_0)}. \end{gathered}$$

Note that, as $K \rightarrow \infty$, the last term approaches 0 because $t_1 - t_0 \rightarrow 0$, while the expression inside the square brackets approaches $\int_V^W \frac{f(x)}{g^r(x)} dx$, thus,

$$K^r D^{(r)}(Q_1) \rightarrow \frac{1}{(r+1)2^r} \int_V^W \frac{f(x)}{g^r(x)} dx. \tag{45}$$

This yields the approximation

$$D^{(r)}(Q_1) \approx \frac{1}{2^{Rr}(r+1)2^r} \int_V^W \frac{f(x)}{g^r(x)} dx.$$

Similarly, $D^{(r)}(Q_2)$ can be approximated by the same quantity as above. Using these approximations together with (44) and the fact that the minimum value for the integral is achieved when $g(x) = \frac{f^{1/(r+1)}(x)}{\int_V^W f^{1/(r+1)}(x) dx}$ [3], the following approximation of the minimal expected distortion is obtained:

$$\bar{D}^{(r)} \approx \frac{2\omega + 2^{-r}\omega_0}{2^{Rr} 2^r (r+1)} (\int_V^W f^{1/(r+1)}(x) dx)^{r+1}. \tag{46}$$

The above result can be put in the following form, which will be used in Appendix C,

$$\lim_{K\rightarrow\infty} K^R \bar{D}^{(r)} = \frac{2\omega + 2^{-r}\omega_0}{2^r (r+1)} (\int_V^W f^{1/(r+1)}(x) dx)^{r+1}. \tag{47}$$

Relation (47) follows immediately from (45), together with the equalities

$$\lim_{K\rightarrow\infty} (2K-1)^r D^{(r)}(Q_0) = \frac{1}{(r+1)2^r} \int_V^W \frac{f(x)}{g^r(x)} dx,$$

which was proved in [5, Theorem 1], and $g(x) = \frac{f^{1/(r+1)}(x)}{\int_V^W f^{1/(r+1)}(x) dx}$.

Appendix C. Number of Iterations of Secant Search

We assume here the squared error as distortion measure and a probability mass function obtained by applying a fine pre-quantizer to a continuous strictly positive pdf $f(x)$ defined on a compact interval $[V,W]$. The pre-quantizer partitions the total interval $[V,W]$ in N equal sub-intervals and selects the optimum representation value for each sub-interval. In other words, we have

$$\begin{aligned} p_i &= \int_{V+(i-1)(W-V)/N}^{V+i(W-V)/N} f(x) dx, \\ x_i &= \frac{\int_{V+(i-1)(W-V)/N}^{V+i(W-V)/N} x f(x) dx}{p_i}, \end{aligned} \tag{48}$$

for all $1 \leq i \leq N$. We will denote by $\bar{D}_N(l)$, respectively $\bar{D}_l$, the smallest expected distortion of an l-level balanced convex 2DSQ for the pmf defined above, respectively for the pdf $f(x)$.

Proposition 4. For a pmf as above and N and K large enough, the number of iterations in the secant search is upper bounded by $8K + \lceil \log_{3/2} N \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function[2].

In order to prove the above proposition we need the following three lemmas.

Lemma 5. There is a positive constant c_3 such that

$$|\bar{D}_l - \bar{D}_N(l)| \leq \frac{c_3}{N}, \tag{49}$$

for all $N >> l$.

Lemma 6. There is some positive integer l_0, such that for all $l, l_0 \leq l << N$, we have

$$\bar{D}_N(l) > 3\bar{D}_N(2l). \tag{50}$$

Lemma 7. Assume that $2K << N$ and $L(P_{\lambda_2}) < 2K$. Then, if $L(P_{\lambda_1}) - L(P_{\lambda_2}) \geq \max 3\{2l_0, L(P_{\lambda_2})\}$, it follows that $L(P_{\lambda_{new}}) - L(P_{\lambda_2}) < \frac{2}{3}(L(P_{\lambda_1}) - L(P_{\lambda_2}))$, where λ_{new} is obtained according to the secant search strategy, i.e.,

$$\lambda_{new} = (W(P_{\lambda_2}) - W(P_{\lambda_1}))/(L(P_{\lambda_1}) - L(P_{\lambda_2})),$$

and l_0 is the constant defined in Lemma 6.

We first prove Proposition 4, then each of the three lemmas.

Proof of Proposition 4. We can think of the secant search as a search in the interval of integers $[L(P_{\lambda_2}), L(P_{\lambda_1})]$. After each iteration the search interval is reduced as follows:

Case 1: If $L(P_{\lambda_{new}}) < 2K$ then the new search interval is $[L(P_{\lambda_{new}}), L(P_{\lambda_1})]$.

Case 2: If $L(P_{\lambda_{new}}) > 2K$ then the new search interval is $[L(P_{\lambda_2}), L(P_{\lambda_{new}})]$.

The search stops when $L(P_{\lambda_{new}}) = 2K$. Note that after each iteration the search interval is reduced by at least one unit since $L(P_{\lambda_2}) < L(P_{\lambda_{new}}) < L(P_{\lambda_1})$ as justified in Section 6.

The total number of iterations is the sum of three quantities q_1, q_2 and q_3 defined as follows. q_1 is the total number of iterations when Case 1 happens. q_2 is the total number of iterations when the size of the current search interval is smaller than $6K$ and Case 2 happens. Finally, q_3 is the total number of iterations when the size of the current search interval is at least $6K$ and Case 2 happens.

Since the values $L(P_{\lambda_{new}})$ corresponding to distinct iterations are different, and since there are at most $2K-1$ different positive integers smaller than $2K$ it follows that $q_1 \leq 2K-1$. By a similar reasoning we obtain that $q_2 \leq 6K - 1$.

In order to provide an upper bound for q_3 we assume that $K \geq l_0$. Also assume that $L(P_{\lambda_1}) - L(P_{\lambda_2}) \geq 6K$. Clearly, $L(P_{\lambda_1}) - L(P_{\lambda_2}) \geq 6l_0$. Also, since $2K > L(P_{\lambda_2})$, we have $L(P_{\lambda_1}) - L(P_{\lambda_2}) \geq 3L(P_{\lambda_2})$. By Lemma 7 it follows that $L(P_{\lambda_{new}}) - L(P_{\lambda_2}) < \frac{2}{3}(L(P_{\lambda_1}) - L(P_{\lambda_2}))$. Consequently, if the length of the current search interval is at least $6K$ and situation 2 happens then the length of the new search interval is reduced by at least $\frac{2}{3}$. Starting from the initial search interval $[1, 2N]$, the total number of times this reduction can be applied until the size of the interval becomes smaller than $6K$ is at most $\lceil \log_{3/2}(2N/(6K)) \rceil$. Therefore $q_3 \leq \lceil \log_{3/2}(2N/(6K)) \rceil < \lceil \log_{3/2} N \rceil$. □

Proof of Lemma 5.

According to equation (47) in Appendix B, we have

$$\lim_{l \to \infty} l^2 \bar{D}_l = c_0, \tag{51}$$

where

$$c_0 = \frac{1}{12}(2\omega + 2^{-2}\omega_0)(\int_V^W f^{1/(3)}(x)dx)^3. \tag{52}$$

The proof of this lemma hinges on relation (51), but we moreover need a way to relate $\bar{D}_N(l)$ to $\bar{D}_l$. Let us first establish a mapping between the l-level balanced convex 2DSQ's for the pdf $f(x)$ and those of the pmf $p_i, 1 \leq i \leq N$. Let an l-level balanced convex 2DSQ $\mathbf{Q}_f$ of the pdf $f(x)$ have the following thresholds in the central partition: $t_0 = V < t_1 < t_2 < \cdots < t_{2l-2} < t_{2l-1} = W$. Define now an l-level balanced convex 2DSQ $\mathbf{Q}_p$ of the pmf $p_i, 1 \leq i \leq N$, with the following thresholds $0 = q_0 < q_1 < \cdots < q_{2l-2} < q_{2l-1} = N$, where:

$$q_i = \lceil \frac{N(t_i - V)}{W - V} \rceil \tag{53}$$

for all $0 \leq i \leq 2l-1$. The 2DSQ $\mathbf{Q}_p$ corresponds to a convex 2DSQ $\mathbf{Q}'_f$ of the pdf $f(x)$ with the thresholds

$$t'_i = V + \frac{q_i(W-V)}{N} \tag{54}$$

for $0 \leq i \leq 2l-1$. It is easy to see that the expected distortions of $\mathbf{Q}_p$ and of $\mathbf{Q}'_f$ differ only by the distortion of the pre-quantizer applied to $f(x)$ in order to obtain the pmf $p_i, 1 \leq i \leq N$. We will denote this quantity by Δ_N. Thus,

$$\bar{D}(\mathbf{Q}'_f) = \bar{D}(\mathbf{Q}_p) + \Delta_N. \tag{55}$$

Next we provide an upper bound for $|\bar{D}(\mathbf{Q}'_f) - \bar{D}(\mathbf{Q}_f)|$. Note first that relation (53) implies that

$$\frac{N(t_i - V)}{W - V} \leq q_i < \frac{N(t_i - V)}{W - V} + 1. \tag{56}$$

Using further the equality (54), after some algebraical manipulation we obtain

$$0 \leq t'_i - t_i < \frac{W - V}{N}. \tag{57}$$

We assume that the pdf is smooth enough and $N >> l$ so that the optimum l-level balanced convex 2DSQ of the pdf $f(x)$ has the distance between any consecutive thresholds larger than $\frac{W-V}{N}$. Therefore we will consider only 2DSQ's with this property. This property together with the above inequality imply that

$$V = t_0 = t'_0 < t_1 \leq t'_1 < \cdots t_i \leq t'_i < t_{i+1} \leq t'_{i+1} < \\ \cdots < t_{2l-2} \leq t'_{2l-2} < t_{2l-1} = t'_{2l-1} = W. \tag{58}$$

[2] For any real number x, $\lceil x \rceil$ is the smallest integer larger or equal to x.

Let Q_0, respectively Q'_0, denote the central quantizer of $\mathbf{Q}_f$, respectively $\mathbf{Q}'_f$. For any $u < u' \in [V, W]$ we denote

$$\mu(u, u') = \frac{\int_u^{u'} xf(x)dx}{\int_u^{u'} f(x)dx}. \tag{59}$$

Then

$$\begin{gathered} D(Q'_0) - D(Q_0) = \\ \textstyle\sum_{i=1}^{2l-1} (\int_{t'_{i-1}}^{t'_i} (x - \mu(t'_{i-1}, t'_i))^2 f(x)dx - \\ \textstyle\int_{t_{i-1}}^{t_i} (x - \mu(t_{i-1}, t_i))^2 f(x)dx). \end{gathered} \tag{60}$$

Applying the inequalities (58) we further obtain

$$\begin{gathered} D(Q'_0) - D(Q_0) = \\ \textstyle\sum_{i=1}^{2l-1} \int_{t'_{i-1}}^{t_i} ((x - \mu(t'_{i-1}, t'_i))^2 - \\ (x - \mu(t_{i-1}, t_i))^2) f(x)dx + \\ \textstyle\sum_{i=1}^{2l-2} \int_{t_i}^{t'_i} ((x - \mu(t'_{i-1}, t'_i))^2 - \\ (x - \mu(t_i, t_{i+1}))^2) f(x)dx = \\ \textstyle\sum_{i=1}^{2l-1} \int_{t'_{i-1}}^{t_i} 2(\mu(t_{i-1}, t_i) - \mu(t'_{i-1}, t'_i)) \cdot \\ (x - \tfrac{\mu(t'_{i-1}, t'_i) + \mu(t_{i-1}, t_i)}{2}) f(x)dx + \\ \textstyle\sum_{i=1}^{2l-2} \int_{t_i}^{t'_i} 2(\mu(t_i, t_{i+1}) - \mu(t'_{i-1}, t'_i)) \cdot \\ (x - \tfrac{\mu(t'_{i-1}, t'_i) + \mu(t_i, t_{i+1})}{2}) f(x)dx. \end{gathered} \tag{61}$$

Using the fact that the absolute value of a sum is smaller or equal than the sum of absolute values and the fact that $f(x)$ is positive we further obtain that

$$\begin{gathered} |D(Q'_0) - D(Q_0)| \leq \cdot \\ 2 \textstyle\sum_{i=1}^{2l-1} |\mu(t'_{i-1}, t'_i) - \mu(t_{i-1}, t_i)| \cdot \\ \textstyle\int_{t'_{i-1}}^{t_i} |x - \tfrac{\mu(t'_{i-1}, t'_i) + \mu(t_{i-1}, t_i)}{2}| f(x)dx + \\ 2 \textstyle\sum_{i=1}^{2l-2} |\mu(t_i, t_{i+1}) - \mu(t'_{i-1}, t'_i)| \cdot \\ \textstyle\int_{t_i}^{t'_i} |x - \tfrac{\mu(t'_{i-1}, t'_i) + \mu(t_i, t_{i+1})}{2}| f(x)dx. \end{gathered} \tag{62}$$

Using the following property of the centroid $\mu(u, u') \in [u, u']$, it follows that $\frac{\mu(t'_{i-1}, t'_i) + \mu(t_{i-1}, t_i)}{2} \in [t_{i-1}, t'_i]$. Then for $x \in [t'_{i-1}, t_i]$ we have $|x - \frac{\mu(t'_{i-1}, t'_i) + \mu(t_{i-1}, t_i)}{2}| \leq t'_i - t_{i-1}$. Likewise $\frac{\mu(t'_{i-1}, t'_i) + \mu(t_i, t_{i+1})}{2} \in [t'_{i-1}, t_{i+1}]$. Since $[t_i, t'_i] \subset [t'_{i-1}, t_{i+1}]$, for $x \in [t_i, t'_i]$ we have $|x - \frac{\mu(t'_{i-1}, t'_i) + \mu(t_i, t_{i+1})}{2}| \leq t_{i+1} - t'_{i-1} \leq t_{i+1} - t_{i-1}$. By applying these observations to (62) we obtain:

$$\begin{gathered} |D(Q'_0) - D(Q_0)| \leq \\ 2 \textstyle\sum_{i=1}^{2l-1} |\mu(t'_{i-1}, t'_i) - \mu(t_{i-1}, t_i)| (t'_i - t_{i-1}) \cdot \\ \textstyle\int_{t'_{i-1}}^{t_i} f(x)dx + \\ 2 \textstyle\sum_{i=1}^{2l-2} |\mu(t_i, t_{i+1}) - \mu(t'_{i-1}, t'_i)| (t_{i+1} - t_{i-1}) \cdot \\ \textstyle\int_{t_i}^{t'_i} f(x)dx. \end{gathered} \tag{63}$$

Next we will treat each sum separately. In order to deal with the second sum notice that $\int_{t_i}^{t'_i} f(x)dx \leq M_0(t'_i - t_i) < M_0(W - V)/N$, where M_0 denotes the maximum value of $f(x)$ over $[V, W]$ (this value is finite due to the continuity of the pdf). Also we have $|\mu(t_i, t_{i+1}) - \mu(t'_{i-1}, t'_i)| \leq W - V$. Then the following inequalities hold

$$\begin{gathered} \textstyle\sum_{i=1}^{2l-2} |\mu(t_i, t_{i+1}) - \mu(t'_{i-1}, t'_i)| (t_{i+1} - t_{i-1}) \int_{t_i}^{t'_i} f(x)dx \leq \\ \tfrac{M_0(W-V)^2}{N} \textstyle\sum_{i=1}^{2l-2} (t_{i+1} - t_{i-1}) = \\ \tfrac{M_0(W-V)^2}{N} (t_{2l-2} - t_0 + t_{2l-1} - t_1) \leq \tfrac{2M_0(W-V)^3}{N}. \end{gathered} \tag{64}$$

In order to find an upper bound for the first sum in relation (63) notice that

$$\begin{gathered} |\mu(t'_{i-1}, t'_i) - \mu(t_{i-1}, t_i)| \leq \\ |\mu(t'_{i-1}, t'_i) - \mu(t_{i-1}, t'_i)| + |\mu(t_{i-1}, t'_i) - \mu(t_{i-1}, t_i)|. \end{gathered} \tag{65}$$

Using the definition of $\mu(\cdot, \cdot)$ in (59) and the inequalities $t_{i-1} \leq t'_{i-1} < t'_i$ from (58), we obtain the following sequence of relations

$$\begin{gathered} \mu(t_{i-1}, t'_i) \textstyle\int_{t_{i-1}}^{t'_i} f(x)dx = \int_{t_{i-1}}^{t'_i} xf(x)dx = \\ \textstyle\int_{t_{i-1}}^{t'_{i-1}} xf(x)dx + \int_{t'_{i-1}}^{t'_i} xf(x)dx = \\ \mu(t_{i-1}, t'_{i-1}) \textstyle\int_{t_{i-1}}^{t'_{i-1}} f(x)dx + \mu(t'_{i-1}, t'_i) \int_{t'_{i-1}}^{t'_i} f(x)dx, \end{gathered}$$

which further imply the equality

$$\mu(t_{i-1}, t'_i) = \mu(t_{i-1}, t'_{i-1}) \frac{\int_{t_{i-1}}^{t'_{i-1}} f(x)dx}{\int_{t_{i-1}}^{t'_i} f(x)dx} + \mu(t'_{i-1}, t'_i) \frac{\int_{t'_{i-1}}^{t'_i} f(x)dx}{\int_{t_{i-1}}^{t'_i} f(x)dx}.$$

The above relation leads to

$$\begin{gathered} |\mu(t'_{i-1}, t'_i) - \mu(t_{i-1}, t'_i)| = \\ |\mu(t'_{i-1}, t'_i) - \mu(t_{i-1}, t'_{i-1})| \tfrac{\int_{t_{i-1}}^{t'_{i-1}} f(x)dx}{\int_{t_{i-1}}^{t'_i} f(x)dx} \leq \\ \tfrac{(t'_i - t_{i-1}) M_0 (W-V)}{N \int_{t_{i-1}}^{t'_i} f(x)dx}. \end{gathered} \tag{66}$$

Moreover, using the definition of $\mu(\cdot, \cdot)$ in (59) and the inequalities $t_{i-1} < t_i \leq t'_i$ from (58), it follows that

$$\begin{gathered} \mu(t_{i-1}, t'_i) \textstyle\int_{t_{i-1}}^{t'_i} f(x)dx = \int_{t_{i-1}}^{t'_i} xf(x)dx = \\ \textstyle\int_{t_{i-1}}^{t_i} xf(x)dx + \int_{t_i}^{t'_i} xf(x)dx = \\ \mu(t_{i-1}, t_i) \textstyle\int_{t_{i-1}}^{t_i} f(x)dx + \mu(t_i, t'_i) \int_{t_i}^{t'_i} f(x)dx, \end{gathered}$$

further leading to

$$\mu(t_{i-1}, t'_i) = \mu(t_{i-1}, t_i) \frac{\int_{t_{i-1}}^{t_i} f(x)dx}{\int_{t_{i-1}}^{t'_i} f(x)dx} + \mu(t_i, t'_i) \frac{\int_{t_i}^{t'_i} f(x)dx}{\int_{t_{i-1}}^{t'_i} f(x)dx}.$$

Using the above relation we obtain

$$\begin{gathered} |\mu(t_{i-1}, t'_i) - \mu(t_{i-1}, t_i)| = \\ |\mu(t_i, t'_i) - \mu(t_{i-1}, t_i)| \tfrac{\int_{t_i}^{t'_i} f(x)dx}{\int_{t_{i-1}}^{t'_i} f(x)dx} \leq \\ \tfrac{(t'_i - t_{i-1}) M_0 (W-V)}{N \int_{t_{i-1}}^{t'_i} f(x)dx}. \end{gathered} \tag{67}$$

By replacing (66) and (67) in (65), it follows that

$$\begin{gathered} |\mu(t'_{i-1}, t'_i) - \mu(t_{i-1}, t_i)| \leq \\ \tfrac{2M_0(W-V)(t'_i - t_{i-1})}{N \int_{t_{i-1}}^{t'_i} f(x)dx} \leq \tfrac{2M_0(W-V)^2}{N \int_{t_{i-1}}^{t'_i} f(x)dx}, \end{gathered}$$

which further implies the following inequality

$$\sum_{i=1}^{2l-1} |\mu(t'_{i-1}, t'_i) - \mu(t_{i-1}, t_i)|(t'_i - t_{i-1}) \int_{t'_{i-1}}^{t_i} f(x)dx \leq \frac{2M_0(W-V)^2}{N} \sum_{i=1}^{2l-1} (t'_i - t_{i-1}) \leq \frac{4M_0(W-V)^3}{N}. \quad (68)$$

Further, by replacing (64) and (68) in (63) it follows that

$$|D(Q'_0) - D(Q_0)| \quad \leq \quad \frac{12M_0(W-V)^3}{N}. \quad (69)$$

According to the above relation there is some positive constant c_1 (i.e., which does not depend on N) such that

$$|D(Q'_0) - D(Q_0)| \quad \leq \quad \frac{c_1}{N}.$$

A similar result can be obtained for the difference of distortions of corresponding side quantizers of $\mathbf{Q}'_f$ and $\mathbf{Q}_f$. Therefore we conclude that there is some constant $c_2 > 0$ such that

$$|\bar{D}(\mathbf{Q}'_f) - \bar{D}(\mathbf{Q}_f)| \quad \leq \quad \frac{c_2}{N}. \quad (70)$$

From relation (55) we obtain that

$$|\bar{D}(\mathbf{Q}'_f) - \bar{D}(\mathbf{Q}_p)| = \Delta_N = \sum_{i=1}^{N} \int_{V+(i-1)(W-V)/N}^{V+i(W-V)/N} (x - x_i)^2 f(x)dx \leq \frac{(W-V)^2}{N^2} \sum_{i=1}^{N} \int_{V+(i-1)(W-V)/N}^{V+i(W-V)/N} f(x)dx = \frac{(W-V)^2}{N^2} \leq \frac{c'_2}{N},$$

for some suitable positive constant c'_2. Further,

$$|\bar{D}(\mathbf{Q}_f) - \bar{D}(\mathbf{Q}_p)| \leq |\bar{D}(\mathbf{Q}_f) - \bar{D}(\mathbf{Q}'_f)| + |\bar{D}(\mathbf{Q}'_f) - \bar{D}(\mathbf{Q}_p)| \leq \frac{c_2+c'_2}{N}.$$

Let $c_3 = c_2 + c'_2$, then the above relation implies that

$$\bar{D}(\mathbf{Q}_p) - \frac{c_3}{N} \leq \bar{D}(\mathbf{Q}_f) \leq \bar{D}(\mathbf{Q}_p) + \frac{c_3}{N}.$$

The above sequence of inequalities remains valid if we apply infimum over all possible $\mathbf{Q}_f$, i.e.,

$$\inf_{\mathbf{Q}_f} \bar{D}(\mathbf{Q}_p) - \frac{c_3}{N} \leq \inf_{\mathbf{Q}_f} \bar{D}(\mathbf{Q}_f) \leq \inf_{\mathbf{Q}_f} \bar{D}(\mathbf{Q}_p) + \frac{c_3}{N},$$

which implies

$$|\inf_{\mathbf{Q}_f} \bar{D}(\mathbf{Q}_f) - \inf_{\mathbf{Q}_f} \bar{D}(\mathbf{Q}_p)| \leq \frac{c_3}{N}.$$

Note that $\mathbf{Q}_p$ is a function of $\mathbf{Q}_f$ and as $\mathbf{Q}_f$ varies over the whole set of l-level balanced convex 2DSQ's of $f(x)$, $\mathbf{Q}_p$ varies over the whole set of l-level balanced convex 2DSQ's of the pmf $p_i, 1 \leq i \leq N$. Therefore, $\inf_{\mathbf{Q}_f} \bar{D}(\mathbf{Q}_p) = \bar{D}_N(l)$, which further implies our claim. □

Proof of Lemma 6.

According to equation (47) in Appendix B, we have

$$\lim_{l \to \infty} l^2 \bar{D}_l = c_0, \quad (71)$$

where c_0 is given by (52). Relation (71) implies that there is some integer l_0 such that for all integers $l \geq l_0$ the following inequality holds

$$|l^2 \bar{D}_l - c_0| \leq c_0/14, \quad (72)$$

which further implies that

$$|\bar{D}_l - \frac{c_0}{l^2}| \leq \frac{c_0}{14l^2}. \quad (73)$$

According to Lemma 5 there is a positive constant c_3 such that the following inequality is valid for all $l << N$,

$$|\bar{D}_l - \bar{D}_N(l)| \leq \frac{c_3}{N}. \quad (74)$$

Relations (74) and (73) further imply that

$$|\bar{D}_N(l) - \frac{c_0}{l^2}| \leq |\bar{D}_N(l) - \bar{D}_l| + |\bar{D}_l - \frac{c_0}{l^2}| \leq \frac{c_3}{N} + \frac{c_0}{14l^2},$$

which lead to

$$\frac{13c_0}{14l^2} - \frac{c_3}{N} \leq \bar{D}_N(l) \leq \frac{15c_0}{14l^2} + \frac{c_3}{N}, \quad (75)$$

for all l, N with $l \geq l_0$ and $N >> l$. When $l \geq l_0$, it follows that $2l \geq l_0$, consequently, relation (75) also holds if l is replaced by $2l$, i.e.,

$$\frac{13c_0}{14 \cdot 4l^2} - \frac{c_3}{N} \leq \bar{D}_N(2l) \leq \frac{15c_0}{14 \cdot 4l^2} + \frac{c_3}{N}, \quad (76)$$

Let $c = \sqrt{\frac{c_0}{32c_3}}$. Let $l_0 \leq l << N$. Assume that the condition $l << N$ implies that $l < c\sqrt{N}$. Then we have $4\frac{c_3}{N} \leq \frac{c_0}{8l^2}$, which further leads to

$$3(\frac{15c_0}{14 \cdot 4l^2} + \frac{c_3}{N}) \leq \frac{13c_0}{14l^2} - \frac{c_3}{N}. \quad (77)$$

Finally, from (75), (76) and (77) it follows that

$$3\bar{D}_N(2l) \leq \bar{D}_N(l).$$

□

Proof of Lemma 7.

Let ℓ denote $L(P_{\lambda_2})$, k denote $L(P_{\lambda_1}) - L(P_{\lambda_2})$, and denote ℓ' denote $L(P_{\lambda_{new}})$.

Consider the points A, B, C, D and E on the lower convex hull of $\mathcal{U}$, corresponding to the following abscisa respectively: $\ell, \lfloor \ell + \frac{1}{3}k + 1 \rfloor, \lfloor \ell + \frac{2}{3}k \rfloor, \ell + k, \ell'$, where $\lfloor \cdot \rfloor$ denotes the floor function[3]. Hence, $A = (\ell, \bar{W}(\ell))$, $B = (\lfloor \ell + \frac{1}{3}k + 1 \rfloor, \bar{W}(\lfloor \ell + \frac{1}{3}k + 1 \rfloor)$, $C = (\lfloor \ell + \frac{2}{3}k \rfloor, \bar{W}(\lfloor \ell + \frac{2}{3}k \rfloor)$, $D = (\ell + k, \bar{W}(\ell + k))$ and $E = (\ell', \bar{W}(\ell'))$.

Our development hinges on the following relation

$$slope(BC) > slope(AD), \quad (78)$$

where $slope(BC)$ and $slope(AD)$ denote the slope of the line BC, respectively AD. After proving the above relation our argument proceeds as follows. Any support line to $\mathcal{U}$ passing through a point to the right of C has the slope at least equal to $slope(BC)$, therefore strictly larger than $slope(AD)$. Since there is a support line passing through E of slope equal to $-\lambda_{new} = slope(AD)$, it follows that E has to be situated to the left of C. This implies that $\ell' < \ell + \frac{2}{3}k$, which proves our claim.

We proceed now to prove (78). By the definition of a slope, we have

$$slope(BC) = \frac{\bar{W}(\lfloor \ell + \frac{2}{3}k \rfloor) - \bar{W}(\lfloor \ell + \frac{1}{3}k + 1 \rfloor)}{\lfloor \ell + \frac{2}{3}k \rfloor - \lfloor \ell + \frac{1}{3}k + 1 \rfloor} = \frac{\bar{D}_N(\lfloor \ell + \frac{2}{3}k \rfloor) - \bar{D}_N(\lfloor \ell + \frac{1}{3}k + 1 \rfloor)}{\lfloor \ell + \frac{2}{3}k \rfloor - \lfloor \ell + \frac{1}{3}k + 1 \rfloor}. \quad (79)$$

[3] For any real value x, $\lfloor x \rfloor$ is the largest integer smaller or equal to x.

We show next that the denominator of the above ratio is smaller than $k/3$. According to the definition of the floor function ($\lfloor\cdot\rfloor$) we have:

$$\ell+\frac{2}{3}k-1<\lfloor\ell+\frac{2}{3}k\rfloor\le\ell+\frac{2}{3}k \tag{80}$$

and

$$\ell+\frac{1}{3}k<\lfloor\ell+\frac{1}{3}k+1\rfloor\le\ell+\frac{1}{3}k+1 \tag{81}$$

By multiplying by -1 the sequence of inequalities (81) and then adding it to (80), we obtain

$$\frac{1}{3}k-2<\lfloor\ell+\frac{2}{3}k\rfloor-\lfloor\ell+\frac{1}{3}k+1\rfloor<\frac{1}{3}k. \tag{82}$$

Relations (79) and (82) imply that

$$slope(BC)>\frac{\bar{D}_N(\lfloor\ell+\frac{2}{3}k\rfloor)-\bar{D}_N(\lfloor\ell+\frac{1}{3}k+1\rfloor)}{\frac{1}{3}k}. \tag{83}$$

Since

$$slope(AD)=\frac{\bar{D}_N(\ell+k)-\bar{D}_N(\ell)}{k}, \tag{84}$$

in order to prove (78) it is sufficient to show that

$$\frac{\bar{D}_N(\lfloor\ell+\frac{2}{3}k\rfloor)-\bar{D}_N(\lfloor\ell+\frac{1}{3}k+1\rfloor)}{\frac{1}{3}k}>\frac{\bar{D}_N(\ell+k)-\bar{D}_N(\ell)}{k},$$

which after some algebra becomes equivalent to

$$\bar{D}_N(\ell)+3\bar{D}_N(\lfloor\ell+\frac{2}{3}k\rfloor)>3\bar{D}_N(\lfloor\ell+\frac{1}{3}k+1\rfloor)+\bar{D}_N(\ell+k). \tag{85}$$

In order to prove the above inequality we first establish that

$$\bar{D}_N(\ell)>3\bar{D}_N(\lfloor\ell+\frac{1}{3}k+1\rfloor). \tag{86}$$

Since $\bar{D}_N(\cdot)$ is decreasing it follows that

$$\bar{D}_N(\ell)\ge\bar{D}_N(\max\{\ell,l_0\}). \tag{87}$$

Since $\ell<2K$ and $2K<<N$, it follows that $\ell<<N$. Consequently, we can apply Lemma 6, which implies that

$$\bar{D}_N(\max\{\ell,l_0\})\ge3\bar{D}_N(2\max\{\ell,l_0\}). \tag{88}$$

From $k\ge3\max\{2l_0,\ell\}$ we obtain $\ell+\frac{1}{3}k\ge\max\{2l_0+\ell,2\ell\}\ge\max\{2l_0,2\ell\}$. Since $\lfloor\ell+\frac{1}{3}k+1\rfloor>\ell+\frac{1}{3}k$, we further obtain that $\lfloor\ell+\frac{1}{3}k+1\rfloor>\max\{2l_0,2\ell\}$. Using again the decreasing monotonicity of $\bar{D}_N(\cdot)$ it follows that

$$\bar{D}_N(\max\{2\ell,2l_0\})\ge\bar{D}_N(\lfloor\ell+\frac{1}{3}k+1\rfloor). \tag{89}$$

Relations (87), (88) and (89) further imply inequality (86).

Since $\lfloor\ell+\frac{2}{3}k\rfloor<\ell+k$ we obtain that $\bar{D}_N(\lfloor\ell+\frac{2}{3}k\rfloor)\ge\bar{D}_N(\ell+k)$, which yields $3\bar{D}_N(\lfloor\ell+\frac{2}{3}k\rfloor)\ge\bar{D}_N(\ell+k)$. The above relation together with (86) lead to (85). □

References

[1] A. Aggarwal, B. Schieber, and T. Tokuyama, ”Finding a minimum-weight k-link path in graphs with the concave monge property and applications”, *Discrete & Computational Geometry* , vol. 12, pp. 263-280, 1994.

[2] R. Ahlswede, ”The rate distortion region for multiple descriptions without excess rate”, *IEEE Trans. Inform. Th.* , vol. IT-31(6), pp. 721-726, Nov. 1985.

[3] V. R. Algazi, ”Useful approximations to optimal quantization”, *IEEE Trans. Commun. Technol.*, vol. COM-14, pp. 297-301, June 1966.

[4] W. R. Bennett, ”Spectra of Quantized Signals”, Bell Syst. Tech. J., vol. 27, pp. 446-472, July 1948.

[5] J. A. Bucklew, and G. L. Wise, ”Multidimensional Asymptotic Quantization Theory with rth Power Distortion Measures”, *IEEE Trans. Inform. Th.* , vol. IT-28(2), pp. 239-247, Mar. 1982.

[6] P. A. Chou, T. Lookabaugh, and R. M. Gray, ”Entropy-constrained vector quantization”, *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, pp. 31-42, Jan. 1989.

[7] S. Dumitrescu and X. Wu, ”Fast algorithms for optimal two-description scalar quantizer design”, *Algorithmica*, vol. 41, no. 4, pp. 269-287, Feb. 2005.

[8] S. Dumitrescu and X. Wu, ”Algorithms for optimal multi-resolution quantization”, *J. Algorithms*, 50(2004), pp. 1-22.

[9] M. Effros and D. Muresan, ”Codecell Contiguity in Optimal Fixed-Rate and Entropy-Constrained Network Scalar Quantizers”, *Proc. DCC'2002*, pp. 312-321, April 2002.

[10] 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.

[11] H. Everett III, ”Generalized Lagrange multiplier method for solving problems of optimum allocation of resources”, *Operations Res.*, 11, pp. 399-417, 1963.

[12] R. M. Gray, ”A Lagrangian formulation of fixed-rate quantization”, *Proc. DCC'2005*, pp. 261-269, March 2005.

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

[14] D. G. Luenberger, *Optimization by vector space methods*, John Wiley & Sons, 1969.

[15] D. Muresan and M. Effros, ”Quantization as histogram segmentation: globally optimal scalar quantizer design in network systems”, *Proc. DCC'2002*, pp. 302-311, April 2002.

[16] D. Muresan and M. Effros, ”Quantization as histogram segmentation: globally optimal scalar quantizer design in network systems”, in preparation.

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

[18] V. A. Vaishampayan, ”Design of multiple-description scalar quantizers”, *IEEE Trans. Inform. Th.*, vol. 39, no. 3, pp. 821-834, May 1993.

[19] V. A. Vaishampayan, J. Domaszewicz, ”Design of entropy-constrained multiple-description scalar quantizers”, *IEEE Trans. Inform. Th.*, vol. 40, no. 1, pp. 245-250, Jan. 1994.

[20] V. A. Vaishampayan and J.-C. Batllo, ”Asymptotic analysis of multiple description scalar quantizers”, *IEEE Trans. Inform. Th.*, vol. 44, no. 1, pp. 278-284, Jan. 1998.

[21] H. S. Witsenhausen and A. D. Wyner, ”Source coding for multiple descriptions II: A binary source”, *Bell Syst. Tech. J.*, vol. 60, pp. 2281-2292, 1981.

[22] J. K. Wolf, A. D. Wyner, and J. Ziv, ”Source coding for multiple descriptions”, *Bell Syst. Tech. J.*, vol. 59, pp. 1417-1426, 1980.

[23] X. Wu, ”Optimal quantization by matrix searching”, *J. Algorithms*, 12(1991), pp. 663-673.

[24] X. Wu and K. Zhang, ”Quantizer monotonicities and globally optimal scalar quantizer design”, *IEEE Trans. Inform. Theory*, vol. 39, pp. 1049-1053, May 1993.

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

Authors' Biographies

Sorina Dumitrescu received the B.Sc. degree in 1990, and the Ph.D. degree in 1997, both in mathematics, from University of Bucharest, Romania.

She is currently an Assistant Professor at the Department of Electrical and Computer Engineering, McMaster University, Ontario, Canada. Her research interests include multimedia computing and communications, joint source-channel coding, signal quantization, steganalysis. She currently holds an NSERC University Faculty Award.

Xiaolin Wu received his B.Sc. from Wuhan University, China in 1982, and Ph.D. degree from University of Calgary, Canada in 1988, both in computer science.

He is currently a Professor at the Department of Electrical and Computer Engineering, McMaster University, Ontario, Canada, and holds the NSERC-Dalsa Research Chair in digital cinema. Dr. Wu's research interests include network-aware multimedia communication, joint source-channel coding, signal quantization and compression, and image processing. He has published over one hundred seventy research papers and holds two patents in these fields. He is an associate editor of IEEE Transactions on Multimedia.