On the Design of Optimal Noisy Channel Scalar Quantizer with Random Index Assignment

Sorina Dumitrescu, *Senior Member, IEEE*

***Abstract*—The general approach in noisy channel scalar quantizer design is an iterative descent algorithm which guarantees only a locally optimal solution. While sufficient conditions under which the local optimum becomes a global optimum are known in the noiseless channel case, such sufficient conditions were not derived for the noisy counterpart. Moreover, efficient globally optimal design techniques for general discrete distributions in the noiseless case exist, however they seem not to extend to the noisy scenario when a fixed index assignment is assumed.**

Recently, the design of noisy channel scalar quantizer with random index assignment (RIA) was proposed using a locally optimal iterative algorithm. In this work we derive sufficient conditions for the uniqueness of a local optimum, which thus guarantee the global optimality of the solution. These sufficient conditions are satisfied for a log-concave probability density function which is, additionally, symmetric around its mean. Furthermore, we show that, assuming an RIA, the globally optimal design for general discrete sources can also be carried out efficiently.

***Index Terms*—Noisy channel quantizer, random index assignment, uniqueness of a local optimum, Monge property.**

I. Introduction

A large body of literature was dedicated to the optimal design of noisy channel quantizers [1]–[7]. Joint source-channel design of multiresolution quantizer for the broadcast channel with hierarchical modulation was also addressed [8]. The design approach taken in most of the aforementioned work is an iterative procedure, which optimizes the encoder and decoder, in turn, while keeping the other component fixed. Such an approach is a generalization of Lloyd's algorithm for noiseless scalar quantizer design [9], [10], and it ensures only a locally optimal solution in general. Neither globally optimal algorithms, nor sufficient conditions for the locally optimal solutions to be globally optimal, are known.

Sufficient conditions for the uniqueness of a locally optimal solution were found in the case of noiseless channel scalar quantizer [11]–[15]. These conditions were shown to hold for any log-concave probability density function (pdf) and a wide class of distortion functions, including the convex increasing error functions. Additionally, it was shown in [16] that Trushkin's sufficient conditions also ensure the uniqueness of a local optimum in the case of multiple description and multiresolution scalar quantizers with convex cells, for convex increasing error functions.

The author is with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, Ontario. Email: sorina@mail.ece.mcmaster.ca

On the other hand, efficient globally optimal design algorithms have also been developed for general discrete distributions for noiseless channel scalar quantizers [17]–[22], multiresolution and multiple descriptions scalar quantizers with convex cells [23]–[27]. The main idea of the globally optimal approach for noiseless channel (fixed rate) scalar quantizer design can be described as follows. The reconstruction value for each quantizer bin (or cell) can be separately optimized and upon doing so, the cost function becomes the sum of the costs of individual cells. Further, the fact that each bin is a contiguous subset of the discrete source alphabet (i.e. the intersection between the alphabet and an interval of the real line) allows for the modeling of the quantizer as a path in a weighted directed acyclic graph (WDAG), where each edge corresponds to a cell. The weight of the path equals the value of the objective function. Thus, the optimization problem becomes a minimum weight path problem constrained on the number of edges, which is solvable via dynamic programming. While the works [17]–[20] do not explicitly use the graph and path terminology, they all essentially subscribe to the above framework. Additionally, it is proved in [19], [20] that the cost function satisfies a nice monotonicity property for a wide class of distortion measures, including the squared error distortion. This property is exploited to substantially accelerate the dynamic programming solution from $O(KN^2)$ to $O(KN)$ running time, where N is the size of the source alphabet and K is the number of quantizer bins. Further improvements in speed are achieved in [21], [22]. On the other hand, the modeling of the optimization problem for various types of network scalar quantizers [23]–[27] is more sophisticated due to the increased complexity of the quantizers, but stems from the same basic ideas, i.e., using dynamic programming and/or graph/path modeling.

Unfortunately, similar ideas seem not to extend to the optimal design of noisy channel scalar quantizers. The main deterrent is the fact that in the latter case the form of the objective function, which is the expected distortion given a particular index assignment, does not allow for the reconstruction value of each cell to be optimized independently of other cells.

Recently, Yu *et al.* [28] proposed the joint source-channel design of scalar and vector quantizers under the squared error distortion, assuming a random index assignment (RIA). In their framework the objective function is the average end to end distortion over all possible index assignments assuming that all index assignments are equally likely. The authors of [28] evaluate the closed form of the above mentioned cost function and propose an iterative generalized Lloyd-type algorithm, which can ensure only a locally optimal solution

in general. Because of the assumption of RIA their design algorithm does not need complete knowledge of the channel and has smaller computational complexity than previous designs. Additionally, the simulation results reported in [28] show that the proposed method is more robust to the variability in average error probability of the channel than the previous approach which assumes a fixed index assignment. Further, the RIA framework is extended in [29], [30] to the design of multiresolution quantizers for the broadcast channel with hierarchical modulation. Similarly, the proposed algorithm cannot guarantee the globally optimal solution in general.

In this work we address the design of optimal noisy channel scalar quantizer with squared error distortion and RIA. We first establish the existence of an optimal solution, which was not proved in the previous work. Next we proceed to investigating sufficient conditions for the uniqueness of a locally optimal solution, which would imply the global optimality of the existing iterative design algorithm. It is worthwhile to point out that the results on the uniqueness of the local optimum in the case of noiseless channel [11]–[15] do not directly apply and it is not straightforward if the same methodology would work in the noisy channel scenario. The main difficulty in using the techniques of the above mentioned work, is that the optimum reconstruction value $g(a,b)$ for a quantization interval (a,b), does not have the same nice properties as in the noiseless channel case. In particular, $g(a,b)$ is not necessarily increasing in its arguments and it is not even guaranteed to be contained in the interval (a,b). However, we show that a locally optimal quantizer must have the reconstruction values included in the corresponding cells. Using this property we are able to adapt the techniques of [13] to derive sufficient conditions for the uniqueness of a local optimum. These sufficient conditions are expressed in terms of the partial derivatives of the function g and we show that they are satisfied if the pdf is log-concave and, additionally, symmetric around its mean.

Furthermore, we present a globally optimal design algorithm for general discrete sources. Specifically, we show that, assuming an RIA, the general graph/path model idea can be extended from the noiseless to the noisy channel case. Additionally, we prove that a monotonicity property similar in spirit to the property identified in [19], [20], holds, allowing for a speed up of the dynamic programming algorithm, which thus achieves $O(KN)$ time complexity.

The paper is structured as follows. The following section formulates the problem of optimal design of a noisy channel K-level scalar quantizer with RIA, for sources with a continuous and positive pdf, and establishes the existence of its solution. In Section III we first present necessary conditions for a local optimum and then derive sufficient conditions for the uniqueness of a local optimum. The next section proposes an efficient globally optimal design algorithm for general discrete sources. Section V presents several experimental results to illustrate the performance of the algorithms discussed in this work. Finally, Section VI concludes the paper.

II. Problem Formulation for A Continuous Source

Let X be a continuous random variable with pdf f. In this work we assume that f satisfies the following conditions.

Condition A. There is an open interval (V,W), $-\infty \leq V < W \leq \infty$, such that $f(x)$ is continuous and positive inside this interval and $f(x)=0$ outside this interval. Denote $\mathcal{I} = [V,W] \cap \mathbb{R}$.

Condition B. The random variable X has a finite second moment, i.e.,

$$\int_V^W x^2 f(x)dx < +\infty.$$

Let ν denote the mean and σ^2 the variance of the source X and let $K \geq 2$ be an integer. A K-level scalar quantizer Q is composed of an encoding function $\psi_Q : \mathcal{I} \to \{1,\cdots,K\}$ and a decoding function $\phi_Q : \{1,\cdots,K\} \to \mathbb{R}$. For each $i, 1 \leq i \leq K$, let $y_i \triangleq \phi_Q(i)$ and $\mathcal{C}_i \triangleq \psi_Q^{-1}(i)$. The sets $\mathcal{C}_i$ are referred to as quantization cells or bins. The set of reconstruction values $\{y_1,\cdots,y_K\}$ is referred to as the codebook.

The index i output by the encoder is further applied a one-to-one index assignment mapping $\pi : \{1,\cdots,K\} \to \{1,\cdots,K\}$. Finally, the index $s=\pi(i)$ is sent over the noisy channel. At the other end, upon receiving index r, the decoder first applies the inverse permutation π^{-1} followed by ϕ_Q, thus outputting the value $\phi_Q(\pi^{-1}(r))$ as the source reconstruction. Let $p(r|s)$, $1 \leq s,r \leq K$, denote the probability that the channel outputs index r, given that s was transmitted. Considering the squared error as a distortion measure, the average end to end distortion (EED) $\bar{D}(Q)$ assuming an RIA[1] is [28]

$$\bar{D}(Q) = \left(1 - \frac{Kp_{err}}{K-1}\right) D(Q) + \frac{Kp_{err}}{K-1} S(Q) + \frac{Kp_{err}}{K-1}\sigma^2, \tag{1}$$

where

$$D(Q) \triangleq \sum_{i=1}^{K} \int_{\mathcal{C}_i} (x-y_i)^2 f(x)dx,$$

$$S(Q) \triangleq \frac{1}{K}\sum_{i=1}^{K} (y_i - \nu)^2,$$

$$p_{err} \triangleq \frac{1}{K}\sum_{s=1}^{K}\sum_{r=1,r\neq s}^{K} p(r|s).$$

Notice that $D(Q)$ is the distortion due to quantization and $S(Q)$ is the so-called *scatter factor*. In order to simplify (1) let us denote

$$c_1 \triangleq 1 - \frac{Kp_{err}}{K-1}, \quad c_2 \triangleq \frac{p_{err}}{K-1}.$$

We will assume throughout this work that $0 < p_{err} < \frac{K-1}{K}$ so that $c_1 > 0$ and $c_2 > 0$.

The problem of optimal design of a noisy channel K-level scalar quantizer with RIA is formulated as [28]

$$\min_{\psi_Q,\phi_Q} \bar{D}(Q), \tag{2}$$

[1]Note that a system with RIA needs common randomness that is shared by the encoder and the decoder.

for a fixed K. The authors of [28] also derive the necessary conditions for the optimal solution to problem (2). Additionally, they point out that these conditions can be obtained from the necessary conditions for optimal standard noisy channel quantizer derived in [2] for an average symmetric channel given by $p(r|s) = p_{err}/(K-1)$ for $r \neq s$, and $p(s|s) = 1 - p_{err}$. Notice that the EED of (1) can be rewritten as

$$\bar{D}(Q) = c_1 D(Q) + c_2 \sum_{i=1}^{K} (y_i - \nu)^2 + K c_2 \sigma^2$$
$$= \sum_{i=1}^{K} \left(c_1 \int_{\mathcal{C}_i} (x - y_i)^2 f(x) dx + c_2 (y_i - \nu)^2 \right) + K c_2 \sigma^2.$$

It can be easily seen that the optimum decoder must satisfy the condition

$$y_i = \arg\min_{y \in \mathbb{R}} \left(c_1 \int_{\mathcal{C}_i} (x - y_i)^2 f(x) dx + c_2 (y_i - \nu)^2 \right),$$

leading to [28]

$$y_i = \frac{c_1 \int_{\mathcal{C}_i} x f(x) dx + c_2 \nu}{c_1 \int_{\mathcal{C}_i} f(x) dx + c_2},$$

for all $1 \leq i \leq K$. An interesting observation is that if cell $\mathcal{C}_i$ equals some arbitrary interval $(a, b]$ then the optimal reconstruction value is not guaranteed to be included in the cell $\mathcal{C}_i$, unlike the case of optimal noiseless quantizer.

It was observed in [28] that, for a fixed decoder, the encoder is optimized by assigning each sample value x to the closest reconstruction value. This can be easily verified based on the expression of the EED since the assignment of x to some cell only affects the quantization distortion $D(Q)$. It follows that we may restrict the search in problem (2) only to the set of regular quantizers, i.e. where the cells are intervals. The encoder of a regular K-level quantizer is specified by the $(K-1)$-tuple of thresholds $\mathbf{x} = (x_1, x_2, \cdots, x_{K-1})$, such that $V < x_1 < \cdots < x_{K-1} < W$, where

$$\mathcal{C}_i = (x_{i-1}, x_i], \quad 2 \leq i \leq K-1, \tag{3}$$
$$\mathcal{C}_1 = [V, x_1] \cap \mathbb{R}, \quad \mathcal{C}_K = (x_{K-1}, W] \cap \mathbb{R}.$$

Further, for any $(K-1)$-tuple $\mathbf{x}$ as above we denote $x_0 = V$ and $x_K = W$. Additionally, for every $V \leq a \leq b \leq W$ let

$$g(a, b) \triangleq \frac{c_1 \int_a^b x f(x) dx + c_2 \nu}{c_1 \int_a^b f(x) dx + c_2}. \tag{4}$$

Then the optimal decoder corresponding to encoder $\mathbf{x}$ must satisfy

$$y_i = g(x_{i-1}, x_i), \quad 1 \leq i \leq K. \tag{5}$$

Let us denote $\mathcal{O}_K \triangleq \{(x_1, x_2, \cdots, x_{K-1}) | V < x_1, x_{K-1} < W, x_i < x_{i+1}, 1 \leq i \leq K-2\}$. Further, for any $\mathbf{x} \in \mathcal{O}_K$ let us denote by $Q_{\mathbf{x}}$ the quantizer satisfying relations (3) and (5). The above considerations imply that, in order to solve problem (2) it is sufficient to solve the following problem

$$\min_{\mathbf{x} \in \mathcal{O}_K} \bar{D}(Q_{\mathbf{x}}). \tag{6}$$

While problem (6) was considered in [28] it was not actually proved that its solution exists. It is clear that $\inf_{\mathbf{x} \in \mathcal{O}_K} \bar{D}(Q_{\mathbf{x}})$ is finite because $\bar{D}(Q_{\mathbf{x}}) \geq 0$ for every $\mathbf{x} \in \mathcal{O}_K$, but it is not clear whether or not this infimum is achieved for some $\mathbf{x} \in \mathcal{O}_K$. It is relevant to mention here the following difference between the noiseless and the noisy channel quantizers. By splitting a cell into two the distortion due to quantization will always decrease[2], therefore an optimal noiseless channel quantizer may not have empty cells. On the other hand, splitting a cell may lead to an increase of the scatter factor and ultimately to an increase of the EED (in particular if the cell is situated far away from ν). Therefore, it is not clear if the optimum noisy channel scalar quantizer with at most K levels (if it exists) will have K nonempty cells. The following lemma will help us prove that the latter is true, and further establish the existence of the minimum in problem (6). The lemma clarifies that there are situations where it is possible to split a cell into two and simultaneously lower the EED. In order to state the lemma we need the following notation. For every $V \leq a \leq b \leq W$ let

$$cost(a, b) = c_1 \int_a^b (t - g(a, b))^2 f(t) dt + c_2 (g(a, b) - \nu)^2.$$

In other words, if $(a, b]$ is a cell in quantizer Q with optimized decoder, then $cost(a, b)$ represents its contribution to the EED $\bar{D}(Q)$. Notice that $cost(a, a) = 0$ for all $V \leq a \leq W$.

Lemma 1. *Let $V \leq a \leq \nu \leq b \leq W$, with $a < b$. Then there is some $x, a < x < b$, such that*

$$cost(a, x) + cost(x, b) < cost(a, b).$$

This lemma is proved in Appendix A. Next we prove the main result of this section, namely, that the minimum in (6) exists.

Theorem 1. *There exists $\hat{\mathbf{x}} \in \mathcal{O}_K$ such that*

$$\bar{D}(Q_{\hat{\mathbf{x}}}) = \min_{\mathbf{x} \in \mathcal{O}_K} \bar{D}(Q_{\mathbf{x}}).$$

Proof: Let us denote by $\bar{\mathcal{O}}_K$ the closure of $\mathcal{O}_K$, i.e., $\bar{\mathcal{O}}_K \triangleq \{(x_1, \cdots, x_{K-1}) | V \leq x_1 \leq \cdots \leq x_{K-1} \leq W\}$. To each $\mathbf{x} \in \bar{\mathcal{O}}_K \setminus \mathcal{O}_K$ we can still associate a quantizer $Q_{\mathbf{x}}$ satisfying (3) and (5). This quantizer has some of the cells empty or consisting of one element. According to (4) and (5), if $\mathcal{C}_i$ is empty or contains only one value, then $y_i = \nu$. Further, notice that $\bar{D}(Q_{\mathbf{x}})$ is a continuous function of $\mathbf{x}$ over $\bar{\mathcal{O}}_K$, a fact which will be used in the sequel.

Our proof is organized as follows. Part 1 proves that $\inf_{\mathbf{x} \in \bar{\mathcal{O}}_K} \bar{D}(Q_{\mathbf{x}})$ is achieved by some $\bar{\mathbf{x}} \in \bar{\mathcal{O}}_K$. Part 2 shows that $\bar{\mathbf{x}} \in \mathcal{O}_K$.

Part 1. First we show that there exist some finite values V_0 and W_0, $V \leq V_0 < W_0 \leq W$, such that for any $\mathbf{x} \in \bar{\mathcal{O}}_K$ with $x_1 < V_0$ or $x_{K-1} > W_0$, there exists $\mathbf{x}' \in \bar{\mathcal{O}}_K$ with $x'_1 \geq V_0$ and $x'_{K-1} \leq W_0$ such that $\bar{D}(Q_{\mathbf{x}'}) \leq \bar{D}(Q_{\mathbf{x}})$. This fact implies that $\inf_{\mathbf{x} \in \bar{\mathcal{O}}_K} \bar{D}(Q_{\mathbf{x}}) = \inf_{\mathbf{x} \in \mathcal{U}} \bar{D}(Q_{\mathbf{x}})$, where $\mathcal{U} \triangleq \{\mathbf{x} \in \bar{\mathcal{O}}_K | x_1 \geq V_0, x_{K-1} \leq W_0\}$. Since $\mathcal{U}$ is a compact set and a continuous function achieves its infimum on a compact

[2]Recall our assumption that the pdf is positive on (V, W).

set it follows that there is $\bar{\mathbf{x}} \in \mathcal{U}$ such that

$$\bar{D}(Q_{\bar{\mathbf{x}}}) = \inf_{\mathbf{x}\in\bar{\mathcal{O}}_K} \bar{D}(Q_{\mathbf{x}}). \tag{7}$$

In order to prove the claim let Q^* be an optimal K-level noiseless channel quantizer, i.e., $D(Q^*) = \min_{\mathbf{x}\in\mathcal{O}_K} D(Q_{\mathbf{x}}) = \min_{\mathbf{x}\in\bar{\mathcal{O}}_K} D(Q_{\mathbf{x}})$. Let $\mathbf{x}^* \in \mathcal{O}_K$ be the $(K-1)$-tuple describing the partition of Q^* and let y_i^*, $1 \le i \le K$, denote its reconstruction values. Further, let

$$\Delta \triangleq \sqrt{\sum_{i=1}^{K}(y_i^* - \nu)^2}.$$

It follows that $S(Q^*) = \Delta^2/K$. Notice that the claim we want to prove is trivially satisfied if V and W are finite. Therefore, let us assume that $V = -\infty$ and $W = \infty$[3]. Now let $V_0 \triangleq \min\{\nu-\Delta, x_1^*\}$ and $W_0 \triangleq \max\{\nu+\Delta, x_{K-1}^*\}$. Now consider $\mathbf{x} \in \bar{\mathcal{O}}_K$ with $x_1 < V_0$ or $x_{K-1} > W_0$. Let y_{min} denote the smallest value in the codebook of $Q_{\mathbf{x}}$ and y_{max} denote the largest value. We next distinguish between two cases.

Case 1. $y_{min} < V_0$ or $y_{max} > W_0$. Then $S(Q_{\mathbf{x}}) \ge \frac{1}{K}\max\{(y_{min}-\nu)^2, (y_{max}-\nu)^2\} > \frac{1}{K}\Delta^2 = S(Q^*)$, while $\bar{D}(Q_{\mathbf{x}}) \ge D(Q^*)$ by the choice of Q^*. Then $\bar{D}(Q^*) < \bar{D}(Q_{\mathbf{x}})$ and the claim follows by letting $\mathbf{x}' = \mathbf{x}^*$.

Case 2. $y_{min} \ge V_0$ and $y_{max} \le W_0$. We will construct $\mathbf{x}'$ by optimizing the encoder. Let K_0 denote the number of distinct reconstruction values, $1 \le K_0 \le K$. We may assume that $K_0 > 1$ since otherwise we can obtain a quantizer with smaller EED by splitting the only cell into two, in view of Lemma 1. We first relabel the reconstruction values, if necessary, such that $y_i < y_{i+1}$ for all $1 \le i \le K_0-1$[4]. An optimal encoder for this set of reconstruction values is given by $\mathbf{x}' \in \bar{\mathcal{O}}_K$ where

$$\begin{aligned} x_i' &= \tfrac{y_i+y_{i+1}}{2}, \quad 1 \le i \le K_0-1, \\ x_i' &= x_{K_0-1}', \quad K_0 \le i \le K-1. \end{aligned} \tag{8}$$

It follows that $\mathbf{x}' \in \mathcal{U}$ and $\bar{D}(Q_{\mathbf{x}'}) \le \bar{D}(Q_{\mathbf{x}})$. Thus, the claim follows. With this observation, the first part of the proof is completed.

Part 2. Let $\bar{\mathbf{x}} \in \bar{\mathcal{O}}_K$ such that (7) holds. We need to show that $\bar{\mathbf{x}} \in \mathcal{O}_K$. Let us assume that $\bar{\mathbf{x}} \in \bar{\mathcal{O}}_K \setminus \mathcal{O}_K$. This implies that $Q_{\bar{\mathbf{x}}}$ has less than K nonempty cells. Let $i, 1 \le i \le K$, be such that $x_{i-1} \le \nu \le x_i$ and $x_{i-1} < x_i$. Then, in view of Lemma 1, it is possible to split the cell $(x_{i-1}, x_i]$ into two nonempty cells and strictly decrease the EED. Since the new encoder still corresponds to some $\mathbf{x} \in \bar{\mathcal{O}}_K$, this contradicts the optimality of $\bar{\mathbf{x}}$. Thus, the assumption that $\bar{\mathbf{x}}$ is not in $\mathcal{O}_K$ is false. Now the proof of the theorem is complete. ■

III. Sufficient Conditions for the Uniqueness of a Local Optimum

The iterative algorithm proposed in [28] for problem (6) can only guarantee a locally optimal solution in general. The main goal of this section is to establish sufficient conditions under which the local optimum is unique, thus becoming a global optimum. For this we need to introduce a few notations first and make some observations regarding the function g. For every $V \le a < b \le W$ denote

$$\begin{aligned} \rho(a,b) &\triangleq \int_a^b f(x)dx, \\ \mu(a,b) &\triangleq \frac{\int_a^b x f(x)dx}{\int_a^b f(x)dx}. \end{aligned}$$

The fact that $f(x) > 0$ for all $x \in (a,b)$ implies that

$$a < \mu(a,b) < b,$$

for all $V \le a < b \le W$. Notice, additionally, that for every $V \le a < b \le W$, $g(a,b)$ can be expressed as

$$g(a,b) = \frac{c_1\rho(a,b)\mu(a,b) + c_2\nu}{c_1\rho(a,b) + c_2}, \tag{9}$$

which further implies that $g(a,b)$ is a convex combination of $\mu(a,b)$ and ν. Therefore, $g(a,b)$ is situated between $\mu(a,b)$ and ν. Further, it can be easily seen that the function $g(a,b)$ is continuous and differentiable on $\{(a,b)|V < a < b < W\}$. By applying the rules of differentiation we obtain the following relations, for all $V < a < b < W$,

$$g_1(a,b) \triangleq \tfrac{\partial g}{\partial a}(a,b) = \tfrac{c_1 f(a)(g(a,b)-a)}{c_1\rho(a,b)+c_2}, \tag{10}$$

$$g_2(a,b) \triangleq \tfrac{\partial g}{\partial b}(a,b) = \tfrac{c_1 f(b)(b-g(a,b))}{c_1\rho(a,b)+c_2}. \tag{11}$$

Additionally, relation (10) also holds for $V < a < b \le W$ and (11) also holds for $V \le a < b < W$. Next we formally state some necessary conditions for the optimal noisy channel quantizer solving problem (6).

Proposition 1. *Let $\mathbf{x} \in \mathcal{O}_K$ be a solution to problem (6). Then the following relations hold*

$$x_{i-1} < g(x_{i-1}, x_i) < x_i, \quad 1 \le i \le K, \tag{12}$$

$$x_i - g(x_{i-1}, x_i) = g(x_i, x_{i+1}) - x_i, 1 \le i \le K-1. \tag{13}$$

Proof: Let $K_0, 1 \le K_0 \le K$, denote the number of distinct values in the codebook. Relabel the reconstruction values, if necessary, such that $y_i < y_{i+1}$ for all $1 \le i \le K_0-1$. Construct $\mathbf{x}' \in \bar{\mathcal{O}}_K$ satisfying (8). Then $\bar{D}(Q_{\mathbf{x}'}) \le \bar{D}(Q_{\mathbf{x}})$, implying that $\mathbf{x}'$ achieves the infimum in (7). Further, according to Part 2 of the proof of Theorem 1, $\mathbf{x}' \in \mathcal{O}_K$, leading to $K_0 = K$. Then $\mathbf{x}'$ is the unique sequence of thresholds optimizing the encoder for the given codebook. Therefore, one has that $\mathbf{x} = \mathbf{x}'$ and $y_i = g(x_{i-1}, x_i)$ for $1 \le i \le K$, a fact which, together with (8), leads to (13).

Note further that $\mu(V, x_1) > V$, $\mu(x_{K-1}, W) < W$ and $V < \nu < W$. These inequalities imply that $g(V, x_1) > V$ and $g(x_{K-1}, W) < W$, which means that $y_1 > V$ and $y_K < W$. Corroborating with the fact that $x_{i-1} = \frac{y_{i-1}+y_i}{2} < y_i$ for $2 \le i \le K$ and $y_i < \frac{y_i+y_{i+1}}{2} = x_i$ for $1 \le i \le K-1$, relations (12) follow. With this observation the proof is complete. ■

In [28] the following controlled iterative algorithm was proposed. Start with some $\mathbf{x}^{(0)} \in \mathcal{O}_K$. Then, at each iteration

[3]The case when only one of them is infite can be handled similarly.

[4]Unlike the noiseless channel case we cannot assume that the optimum decoder for a fixed encoder either has all K reconstruction values distinct or that they obey the relations $y_i < y_{i+1}$ for all i.

$j, j \leq 1$, construct $\mathbf{x}^{(j)}$ according to

$$x_i^{(j)} = \frac{\bar{g}(x_{i-1}^{(j-1)}, x_i^{(j-1)}) + \bar{g}(x_i^{(j-1)}, x_{i+1}^{(j-1)})}{2}, \quad 1 \leq i \leq K-1, \tag{14}$$

where, for $V \leq a \leq b \leq W$,

$$\bar{g}(a,b) \triangleq \begin{cases} g(a,b), & \text{if } a \leq g(a,b) \leq b \\ a, & \text{if } g(a,b) < a \\ b, & \text{if } b < g(a,b) \end{cases}.$$

The algorithm stops when the decrease in EED falls below some threshold. Notice that

$$\bar{g}(a,b) = \arg \min_{y \in [a,b]} \left(c_1 \int_a^b (t-y)^2 f(t) dt + c_2 (y-\nu)^2 \right),$$

and, if $a < \bar{g}(a,b) < b$ then $\bar{g}(a,b) = g(a,b)$. Additionally, $\bar{g}(a,a) = a$ for $V \leq a \leq W$.

It was proved in [28] that the algorithm converges if $V > -\infty$ and $W < \infty$, by invoking the argument of [33]. It was also argued that the limit point satisfies conditions (13). We additionally show that the limit point also fulfills relations (12), a property which will prove essential for our development.

Proposition 2. *Any limit point $\mathbf{x}$ of the controlled iterative algorithm described by (14) is in $\mathcal{O}_K$ and satisfies relations (12) and (13).*

Remark 1. *It is important to note that relations (12) are not necessarily satisfied for arbitrary points in $\mathcal{O}_K$. The fact that they are satisfied by the local optimum will be crucial in establishing sufficient conditions for the uniqueness of the local optimum.*

The following lemma proves a simple property that will be extensively used in the proof of Proposition 2.

Lemma 2. *Let $V \leq a < b \leq W$. If $\bar{g}(a,b) = a$ then $a > \nu$, while if $\bar{g}(a,b) = b$ then $b < \nu$.*

Proof: Assume that $\bar{g}(a,b) = a$. Then the definition of $\bar{g}$ implies that $g(a,b) \leq a$. According to (9) $g(a,b)$ is a strictly convex combination of $\mu(a,b)$ and ν and therefore it is situated between the two values. Since $\mu(a,b) > a$ it follows that $\nu < a$. The remaining claim follows similarly. ∎

Proof of Proposition 2: Let $\mathbf{x}$ be a limit point of the controlled iterative algorithm with update equations (14). Then $\mathbf{x} \in \bar{\mathcal{O}}_K$ and the following equations clearly hold

$$x_{i-1} \leq \bar{g}(x_{i-1}, x_i) \leq x_i, \quad 1 \leq i \leq K, \tag{15}$$

$$x_i = \frac{\bar{g}(x_{i-1}, x_i) + \bar{g}(x_i, x_{i+1})}{2}, \quad 1 \leq i \leq K-1. \tag{16}$$

We will first prove that $\mathbf{x} \in \mathcal{O}_K$. Let us assume for contradiction that $x_{i-1} = x_i$ for some $1 \leq i \leq K$. Then at least one of the following statements is true: S1) there exists some $1 \leq j \leq K-i$ such that $x_i < x_{i+j}$; S2) there exists some $1 \leq k \leq i-1$ such that $x_{i-1-k} < x_{i-1}$. Assume that S1 holds (the case when S2 holds can be treated similarly) and let j be the smallest integer with the specified property. Then one has $x_{i-1} = x_i = \cdots = x_{i+j-1} < x_{i+j}$. Further, aided by (16), one obtains that $\bar{g}(x_{i+j-2}, x_{i+j-1}) = x_{i+j-1} = \bar{g}(x_{i+j-1}, x_{i+j})$. In view of Lemma 2 the latter equality implies that $x_i = x_{i+j-1} > \nu$. It further follows that statement S2 must hold as well. Therefore, let k be the smallest integer with the property described in S2. We then have $x_{i-1-k} < x_{i-k} = \cdots = x_{i-1} = x_i$ and further that $\bar{g}(x_{i-k}, x_{i-k+1}) = x_{i-k} = \bar{g}(x_{i-1-k}, x_{i-k})$, based on (16). Applying again Lemma 2 leads to $x_i = x_{i-k} < \nu$, which contradicts the previous conclusion that $x_i > \nu$. Therefore, the proof that $\mathbf{x} \in \mathcal{O}_K$ is completed.

Let us prove now that relations (15) hold with strict inequality. Let us assume for contradiction that $\bar{g}(x_{i-1}, x_i) = x_i$ for some i. Lemma 2 implies that $x_i < \nu$, consequently, $i < K$. Using further (16) one obtains that $\bar{g}(x_i, x_{i+1}) = x_i$, which leads to $x_i > \nu$ according to Lemma 2. Thus, we have reached a contradiction. A similar contradiction is obtained by assuming that $x_{i-1} = \bar{g}(x_{i-1}, x_i)$. It follows that inequalities (15) are strict. This further implies that $\bar{g}(x_{i-1}, x_i) = g(x_{i-1}, x_i)$ for all $1 \leq i \leq K$, and, further, that (12) and (13) hold, thus concluding the proof. ∎

We are now ready to present the main result of this section, which establishes sufficient conditions for the uniqueness of a local optimum.

Theorem 2. *Assume that the following relations hold for all $V < a < b < W$,*

$$g_1(a, W) < 1, \quad g_2(V, b) < 1, \tag{17}$$

$$g_1(a,b) + g_2(a,b) < 1. \tag{18}$$

Then there is at most one $\mathbf{x} \in \mathcal{O}_K$ satisfying relations (12) and (13).

Remark 2. *It is relevant to mention that in the case of noiseless channel quantizer the conditions for the local optimum have the same form as (13), while (12) is satisfied by default. Therefore, it is natural to ask the question whether the results established in [11]–[15] could be applied directly to derive sufficient conditions for the uniqueness of a local optimum in the noisy channel case. Unfortunately, this is not possible. One of the main reasons is that the proofs in the above mentioned work rely heavily on the fact that the counterpart of the function $g(a,b)$ in the noiseless case is strictly increasing in both a and b, and thus its partial derivatives are positive. In our case $g(a,b)$ is not necessarily included in (a,b) and therefore, according to (10) and (11) its partial derivatives may take negative values.*

The proof of the theorem hinges on the following lemma, which is proved in Appendix B.

Lemma 3. *Assume that inequalities (17) and (18) hold for all $V < a < b < W$. Then the following statements are valid.*

T1) For any $V < b_0 < b_1 < W$ one has $b_0 - g(V, b_0) < b_1 - g(V, b_1)$.

T2) For any $V < a < b_0 < b_1 < W$ such that $a < g(a, b_0) < b_0$ and $a < g(a, b_1) < b_1$, one has $b_0 - g(a, b_0) < b_1 - g(a, b_1)$.

T3) For any $V < a_0 < a_1 < W$ one has $g(a_0, W) - a_0 > g(a_1, W) - a_1$.

T4) For any $V < a_0 < b_0 < W$, $V < a_1 < b_1 < W$, such

that $a_0 < a_1$, $a_0 < g(a_0, b_0) < b_0$ and $a_1 < g(a_1, b_1) < b_1$, the following holds

$$g(a_0, b_0) - a_0 < g(a_1, b_1) - a_1 \Rightarrow$$
$$b_0 - g(a_0, b_0) < b_1 - g(a_1, b_1).$$

Now we are ready to prove Theorem 2.

Proof of Theorem 2: Let us assume for contradiction that there are two distinct $(K-1)$-tuples $\mathbf{x}'$ and $\mathbf{x}''$ in $\mathcal{O}_K$ satisfying the conditions $x'_i < g(x'_i, x'_{i+1}) < x'_{i+1}$, $x''_i < g(x''_i, x''_{i+1}) < x''_{i+1}$ for all $0 \leq i \leq K-1$ and

$$x'_i - g(x'_{i-1}, x'_i) = g(x'_i, x'_{i+1}) - x'_i, \quad (19)$$
$$x''_i - g(x''_{i-1}, x''_i) = g(x''_i, x''_{i+1}) - x''_i, \quad (20)$$

for all $1 \leq i \leq K-1$. Since $\mathbf{x}' \neq \mathbf{x}''$ it follows that there is $k_0, 1 \leq k_0 < K$, such that $x'_i = x''_i$ for all $0 \leq i < k_0$, and $x'_{k_0} \neq x''_{k_0}$. Let us assume without loss of generality (wlg) that $x'_{k_0} < x''_{k_0}$. Then statements $T1$ and $T2$ of Lemma 3 imply that $x'_{k_0} - g(x'_{k_0-1}, x'_{k_0}) < x''_{k_0} - g(x''_{k_0-1}, x''_{k_0})$. Using further (19) and (20) it follows that $g(x'_{k_0}, x'_{k_0+1}) - x'_{k_0} < g(x''_{k_0}, x''_{k_0+1}) - x''_{k_0}$.

Next we will use mathematical induction to prove the following assertion.

Assertion. For all $k_0 \leq j \leq K-1$ the following relations hold

$$x'_j < x''_j, \quad g(x'_j, x'_{j+1}) - x'_j < g(x''_j, x''_{j+1}) - x''_j. \quad (21)$$

We have already shown that (21) hold for $j = k_0$. Assume now that (21) hold for some $j, k_0 \leq j < K-1$. Then condition $T4$ of Lemma 3 implies that

$$x'_{j+1} - g(x'_j, x'_{j+1}) < x''_{j+1} - g(x''_j, x''_{j+1}). \quad (22)$$

By summing all three inequalities in (21) and (22) side by side, we obtain that $x'_{j+1} < x''_{j+1}$. Using further (22), (19) and (20) for $i = j+1$, one obtains that the second inequality in (21) is also satisfied for $j+1$. With this, the assertion is proved.

It follows that relations (21) are fulfilled for $j = K-1$. On the other hand, condition $T3$ of Lemma 3 implies that $g(x'_{K-1}, x'_K) - x'_{K-1} > g(x''_{K-1}, x''_K) - x''_{K-1}$, thus leading to a contradiction. This observation concludes the proof of the theorem. ■

Corollary 1. *Assume that the pdf f is log-concave and symmetric around ν (i.e., $f(x) = f(2\nu - x)$ for all $x \in \mathbb{R}$). Then there is at most one $\mathbf{x} \in \mathcal{O}_K$ satisfying relations (12) and (13).*

Proof: It is sufficient to prove that if f satisfies the conditions in the hypothesis then relations (17) and (18) hold. For this we will use the properties of the partial derivatives of μ for log-concave densities, proved in prior work. Namely, let $\mu_1(a,b) \triangleq \frac{\partial \mu}{\partial a}(a,b)$ and $\mu_2(a,b) \triangleq \frac{\partial \mu}{\partial b}(a,b)$ for $V < a < b < W$. It was proved in [13] that, if f is log-concave then for all $V < a < b < W$ one has

$$\mu_1(a, W) \leq 1, \quad \mu_2(V, b) \leq 1, \quad (23)$$
$$\mu_1(a,b) + \mu_2(a,b) \leq 1. \quad (24)$$

We will first prove inequality (18). The following equalities can be easily derived based on (9)

$$g_1(a,b) = \frac{c_1\rho(a,b)\mu_1(a,b)}{c_1\rho(a,b)+c_2} + \frac{c_1c_2f(a)(\nu-\mu(a,b))}{c_1\rho(a,b)+c_2},$$
$$g_2(a,b) = \frac{c_1\rho(a,b)\mu_2(a,b)}{c_1\rho(a,b)+c_2} + \frac{c_1c_2f(b)(\mu(a,b)-\nu)}{c_1\rho(a,b)+c_2}.$$

The above relations imply that

$$g_1(a,b) + g_2(a,b) = \frac{c_1\rho(a,b)(\mu_1(a,b)+\mu_2(a,b))}{c_1\rho(a,b)+c_2} + \frac{c_1c_2(f(a)-f(b))(\nu-\mu(a,b))}{c_1\rho(a,b)+c_2}.$$

The log-concavity of f implies that the first term in the above expression is strictly smaller than 1, in light of (24) and of the fact that $\frac{c_1\rho(a,b)}{c_1\rho(a,b)+c_2} < 1$. We will next show that the second term is non-positive when f is additionally symmetric around ν. First notice that the log-concavity and the symmetry of f around ν imply that f is non-decreasing on $(V, \nu]$ and non-increasing on $[\nu, W)$. If $b \leq \nu$ then $f(a) \leq f(b)$, while $\mu(a,b) < b \leq \nu$, thus the claim follows. When $a \geq \nu$, we have $f(a) \geq f(b)$ and $\mu(a,b) > a \geq \nu$, which again imply the claim. Assume now that $a < \nu < b$. If $a \leq 2\nu - b$ then one has $f(a) \leq f(b)$ and $\mu(a,b) \leq \nu$. If $a > 2\nu - b$ then one has $f(a) > f(b)$ and $\mu(a,b) > \nu$, thus proving the claim.

Let us prove now inequalities (17). One has

$$g_2(V,b) = \frac{c_1\rho(V,b)\mu_2(V,b)}{c_1\rho(V,b)+c_2} + \frac{c_1c_2f(b)(\mu(V,b)-\nu)}{c_1\rho(V,b)+c_2} < 1 + \frac{c_1c_2f(b)(\mu(V,b)-\nu)}{c_1\rho(V,b)+c_2} < 1,$$

where the first inequality follows from (23) and $0 < \frac{c_1\rho(V,b)}{c_1\rho(V,b)+c_2} < 1$, while the last one follows since $\mu(V,b) < \mu(V,W) = \nu$. On the other hand, one obtains

$$g_1(a,W) = \frac{c_1\rho(a,W)\mu_1(a,W)}{c_1\rho(a,W)+c_2} + \frac{c_1c_2f(a)(\nu-\mu(a,W))}{c_1\rho(a,W)+c_2} < 1$$

using the facts that $\mu_1(a,W) \leq 1$ according to (23), $0 < \frac{c_1\rho(a,W)}{c_1\rho(a,W)+c_2} < 1$ and $\mu(a,W) > \mu(V,W) = \nu$. Thus, the proof is complete. ■

IV. Globally Optimal Algorithm for Discrete Sources

In this section we consider a finite source alphabet and present a globally optimal design algorithm for the noisy channel scalar quantizer with RIA. The case of a finite source is relevant in practical situations where the design is performed based on a set of training samples.

Let $\mathcal{A} = \{a_1, a_2, \cdots, a_N\} \subseteq \mathbb{R}$ be the source alphabet, where $N > 0$ and $a_n < a_{n+1}$, $1 \leq n \leq N-1$. Let $f(a_n)$ denote the probability of symbol a_n. We will assume that $f(a_n) > 0$ for all n. We preserve the notation ν and σ^2 for the mean, respectively the variance of the source. Let ψ'_Q denote the encoding function of a K-level quantizer Q for this discrete alphabet, i.e., $\psi'_Q : \mathcal{A} \to \{1, \cdots, K\}$, and let ϕ'_Q denote the decoding function, i.e., $\phi'_Q : \{1, \cdots, K\} \to \mathcal{B}$. For each $i, 1 \leq i \leq K$, let $y_i \triangleq \phi'_Q(i)$ and $\mathcal{C}_i \triangleq \psi'^{-1}_Q(i)$. Further, the setup of the transmission system remains as in Section II.

We make the following assumptions on the reconstruction alphabet $\mathcal{B}$. Either $\mathcal{B}=\mathbb{R}$ or $\mathcal{B}$ is a discrete set satisfying the following conditions:

C1) $\mathcal{B}=\{b_0+\Delta i | i\in\mathbb{Z}\}$ for some $b_0\in\mathbb{R}$ and $\Delta>0$.

C2) $\mathcal{B}\cap[a_u,a_{u+2}]\neq\emptyset$ for all $0\leq u\leq N-1$, where $a_0=a_1-1$ and $a_{N+1}=a_N+1$ by convention.

The reason for the above assumptions will become clear shortly.

Since the source alphabet is discrete, the average EED $\bar{D}(Q)$ assuming an RIA is given by

$$\bar{D}(Q)=\sum_{i=1}^{K}\left(c_1\sum_{a\in\mathcal{C}_i}(a-y_i)^2f(a)+c_2(y_i-\nu)^2\right)+Kc_2\sigma^2. \tag{25}$$

The problem of optimal design of a noisy channel K-level scalar quantizer with RIA becomes

$$\min_{\psi'_Q,\phi'_Q}\bar{D}(Q), \tag{26}$$

for a fixed K.

The optimum decoder condition is now

$$y_i=\arg\min_{y\in\mathcal{B}}\left(c_1\sum_{a\in\mathcal{C}_i}(a-y)^2f(a)+c_2(y-\nu)^2\right). \tag{27}$$

Since the expression to be minimized in (27) is quadratic in y, it easily follows that its minimizer over the set $\mathcal{B}$ is $\hat{y}(\mathcal{C}_i)$ defined as

$$\hat{y}(\mathcal{C}_i)\triangleq round_{\mathcal{B}}\left(\frac{c_1\sum_{a\in\mathcal{C}_i}af(a)+c_2\nu}{c_1\sum_{a\in\mathcal{C}_i}f(a)+c_2}\right), \tag{28}$$

where $round_{\mathcal{B}}(z)$ denotes the element in $\mathcal{B}$ that is closest to z. We make the convention that in case of a tie the smallest value is considered.

As in the case of a continuous source, the optimum encoder given a fixed decoder has to assign each source sample to the closest reconstruction value. Therefore, we will impose the condition that each quantizer cell $\mathcal{C}_i$ is a contiguous subset of $\mathcal{A}$, i.e., a set of the form $\mathcal{C}(u,v]=\{x\in\mathcal{A}:a_u<x\leq a_v\}$, for some integers u,v with $0\leq u\leq v\leq N$. Additionally, we may restrict the search in problem (26) only to quantizers that are decoder optimized. Therefore, the quantizer is completely specified by the sequence of partition thresholds. However, there is a notable difference versus the continuous case. Namely, the solution to problem (26) is no longer guaranteed to have all K cells nonempty. We will illustrate this observation with an example.

Example: Consider a uniform source over the alphabet $\{-12,-10,10,12\}$. Let $\mathcal{B}=\mathbb{R}$. Clearly, $\nu=0$ and $\sigma^2=122$. Let $K=4$ and assume that the channel is a binary symmetric channel with bit error rate $\epsilon=0.1$. Then $p_{err}=0.19$, $c_1=0.7467$ and $c_2=0.0633$. The only quantizer with four nonempty cells is Q_1 with $\mathcal{C}_1=\{-12\}$, $\mathcal{C}_2=\{-10\}$, $\mathcal{C}_3=\{10\}$ and $\mathcal{C}_4=\{12\}$. The optimal reconstructions are $y_1=-8.9600$, $y_2=-7.4667$, $y_3=7.4667$ and $y_4=8.9600$. Then the quantization distortion and the scatter factor are $D(Q_1)=7.8297$ and $S(Q_1)=68.0164$ leading to $\bar{D}(Q_1)=53.9836$. Consider now quantizer Q_2 with $\mathcal{C}_1=\{-12,-10\}$, $\mathcal{C}_2=\{10,12\}$ and $\mathcal{C}_3=\mathcal{C}_4=\emptyset$. The optimal reconstructions are $y_1=-9.4046$, $y_2=9.4046$ and $y_3=y_4=0$. Then one has $D(Q_2)=3.5454$ and $S(Q_2)=44.2231$ leading to $\bar{D}(Q_2)=44.7570$. It follows that $\bar{D}(Q_2)<\bar{D}(Q_1)$, which implies that the quantizer with all four cells nonempty is not optimal.

According to the above considerations in order to solve problem (26) we need to consider quantizers with k nonempty contiguous cells for all $1\leq k\leq K$. Such a quantizer is completely specified by the $(k+1)$-tuple of integer thresholds $\mathbf{t}^k=(0=t_0,t_1,\cdots,t_k=N)$ with $t_{i-1}<t_i$, $1\leq i\leq k$. Its cells are $\mathcal{C}_i=\mathcal{C}(t_{i-1},t_i]$, for $1\leq i\leq k$, and $\mathcal{C}_j=\emptyset$ for $k+1\leq j\leq K$. Let us, additionally, denote $\mathcal{T}_k\triangleq\{\mathbf{t}^k\in\mathbb{N}^{k+1}:t_{i-1}<t_i,1\leq i\leq K,t_0=0,t_k=N\}$.

Now let us assign a cost $\omega(\mathcal{C}_i)$ to each cell $\mathcal{C}_i$, as follows,

$$\omega(\mathcal{C}_i)=c_1\sum_{a\in\mathcal{C}_i}(a-\hat{y}(\mathcal{C}_i))^2f(a)+c_1(\hat{y}(\mathcal{C}_i)-\nu)^2.$$

Then it becomes clear that the EED $\bar{D}(Q)$ of (25) equals the sum of the costs of all cells plus a term that does not depend on Q, i.e., $\bar{D}(Q)=\sum_{i=1}^{K}\omega(\mathcal{C}_i)+Kc_2\sigma^2$. We will assume for simplicity that $\nu\in\mathcal{B}$. Then ν is the optimal reconstruction corresponding to each empty cell, leading to $\omega(\emptyset)=0$. It follows that problem (26) can be recast as

$$\min_{1\leq k\leq K}\min_{\mathbf{t}^k\in\mathcal{T}_k}\sum_{i=1}^{k}\omega(\mathcal{C}(t_{i-1},t_i]). \tag{29}$$

Consider now the WDAG $G=(V,E,w)$, where the vertex set is $V=\{0,1,\cdots,N\}$ and the edge set is $E=\{(u,v)\in V^2:0\leq u<v\leq N\}$. Let the weight of edge (u,v) be $w(u,v)\triangleq\omega(C(u,v])$. The source node in G is 0 and the final node is N. A path from some node u to another node v is a sequence of connected edges, and the length of the path equals the number of edges in the path. If the beginning and the end of the path are not specified we will understand that the path starts at the source node and ends at the final node.

It is easy to see that for every $k,1\leq k\leq K$, there is a one-to-one correspondence between the set $\mathcal{T}_k$ of quantizer thresholds and the set of k-edge paths in G. Additionally, the weight of the path corresponding to some $\mathbf{t}^k$, i.e., the sum of the weights of its component edges, equals the cost in (29). Therefore, problem (29) can be solved by finding the minimum weight k-edge path in G for every $k,1\leq k\leq K$, followed by solving the outer minimization over all values of k. If the weight of each edge can be computed in constant time, then the solution can be found via dynamic programming in $O(KN^2)$ running time. In order to see how this can be done, let us denote by $\bar{W}_k(v)$ the weight of the minimum weight k-edge path from the source to vertex v, for each $0\leq k\leq K$, and each vertex $v\in V,v\geq k$. Then the following recurrence relation holds, for $1\leq k\leq K$, and $v\in V,v\geq k$,

$$\bar{W}_k(v)=\min_{k-1\leq u<v}\left(\bar{W}_{k-1}(u)+w(u,v)\right). \tag{30}$$

The dynamic programming algorithm computes the values $\bar{W}_k(v)$ for all vertices $v\geq k$, in increasing order of k from 1 to K. After that the solution is found by minimizing $\bar{W}_k(N)$ over all values of k. Assuming that $w(u,v)$ can be determined in

$O(1)$ time for each edge (u,v), solving (30) takes $O(N)$ time. Therefore, computing all values $\bar{W}_k(v)$ requires $O(KN^2)$ operations. Since the last minimization takes only $O(K)$ time, it follows that the overall time complexity is $O(KN^2)$, proving the claim.

To ensure that the computation of the cost of each edge takes only $O(1)$ time we compute and store, as in [19], the following values during a preprocessing step: $P(0,n] \triangleq \sum_{i=1}^{n} f(a)$, $m_1(0,n] \triangleq \sum_{i=1}^{n} af(a)$ and $m_2(0,n] \triangleq \sum_{i=1}^{n} a^2 f(a)$, for all $1 \leq n \leq N$. Then the computation of the expression in (28) takes a constant amount of operations when $\mathcal{B} = \mathbb{R}$ or $\mathcal{B}$ is a discrete set with the regular structure imposed by condition $C1$. Note that the preprocessing step takes only $O(N)$ time, therefore it does not increase the asymptotic running time of the solution algorithm.

In the case of optimal design of noiseless channel scalar quantizer it was shown that the cost function has a nice monotonicity property which enables a speed up of the dynamic programming algorithm by a factor of $O(N)$ [19], [20]. The key factor which makes this property hold is the fact that the edge weights preserve the so-called Monge property [31]. A natural question is whether the latter property still holds in our case as well. We point out that the main difference in the graph model between our case and the noiseless case resides in the definition of the weight associated to each edge[5]. In the noiseless case, the weight assigned to edge (u,v) is

$$\min_{y \in \mathcal{B}} \sum_{a \in \mathcal{C}(u,v]} (a-y)^2 f(a) = \sum_{a \in \mathcal{C}(u,v]} (a - round_{\mathcal{B}}(\nu(u,v)))^2 f(a),$$

where

$$\nu(u,v) = \frac{\sum_{a \in \mathcal{C}(u,v]} af(a)}{\sum_{a \in \mathcal{C}(u,v]} f(a)}.$$

The attempt to apply the technique of [20] to show that the edge weights of G obey the Monge property is hindered by the fact that the optimal reconstruction $\hat{y}(\mathcal{C}(u,v])$ defined in (28), is not necessarily within the boundaries of the set $\mathcal{C}(u,v]$. Fortunately, we are able to get around this difficulty since, as in the case of a continuous source, the optimal solution to problem (26) is a nearest neighbour quantizer. This fact implies that the optimal solution contains only cells $\mathcal{C}(u,v]$ for which the inequalities $a_u < \hat{y}(\mathcal{C}(u,v]) < a_{v+1}$ hold. Therefore, we can safely modify the cost of the edges so that to impose the latter constraint, without sacrificing the optimality of the solution. This claim is proved next.

Lemma 4. *There is an optimal solution to problem (29) specified by a $(k+1)$-tuple of thresholds $\mathbf{t}^k \in \mathcal{T}_k$, for some $1 \leq k \leq K$, satisfying the following properties.*

1) *For every $i, 1 \leq i \leq k$, one has $b_{low} \leq \hat{y}(\mathcal{C}(t_{i-1}, t_i]) \leq b_{high}$, where $b_{low} \triangleq min\,(\mathcal{B} \cap [a_1, a_N])$ and $b_{high} \triangleq max\,(\mathcal{B} \cap [a_1, a_N])$.*
2) *For every $i, 2 \leq i \leq k$, one has $a_{t_{i-1}} < \hat{y}(\mathcal{C}(t_{i-1}, t_i])$ and for every $i, 1 \leq i \leq k-1$, one has $\hat{y}(\mathcal{C}(t_{i-1}, t_i]) < a_{t_i+1}$.*

Remark 3. *Notice that, unlike the continuous case, in the discrete case we cannot conclude that the optimal reconstruction is within the boundaries of a cell even in an optimal quantizer. An illustration of this observation is the example on the previous page. However, the property stated in Lemma 4 is powerful enough for our purpose of proving the Monge property.*

Proof of Lemma 4: To prove the first claim note first that $a_1 \leq \nu \leq a_N$. It follows that the cost function in (27) is decreasing for $y \in (-\infty, a_1]$ and increasing for $y \in [a_N, \infty)$, thus the minimum is achieved for some $y \in \mathcal{B} \cap [a_1, a_N]$. Notice that the latter set is non-empty when $\mathcal{B} = \mathbb{R}$ or $\mathcal{B}$ satisfies condition $C2$.

Let us prove now the second claim. Let $\mathbf{t}^k \in \mathcal{T}_k$ represent an optimal solution to problem (29), where k is the smallest integer with this property. Fix some $i, 1 \leq i < k$. Since the optimal solution is a nearest neighbour quantizer and since all elements of cell $\mathcal{C}_i$ are smaller than all elements of cell $\mathcal{C}_{i+1}$, it follows that $\hat{y}(\mathcal{C}_i) \leq \hat{y}(\mathcal{C}_{i+1})$. If $\hat{y}(\mathcal{C}_i)$ and $\hat{y}(\mathcal{C}_{i+1})$ were equal then cells $\mathcal{C}_i$ and $\mathcal{C}_{i+1}$ could be merged into a single cell without increasing the EED, thus contradicting the choice of k. Therefore, it follows that $\hat{y}(\mathcal{C}_i) < \hat{y}(\mathcal{C}_{i+1})$. Further, the fact that a_{t_i+1} cannot be closer to $\hat{y}(\mathcal{C}_i)$ than to $\hat{y}(\mathcal{C}_{i+1})$ implies that $\hat{y}(\mathcal{C}_i) < a_{t_i+1}$. The remaining inequality can be proved analogously. With this observation the proof is complete. ■

Let us define now

$$w'(u,v) \triangleq \min_{y \in \mathcal{B} \cap [a_u, a_{v+1}]} \Big(c_1 \sum_{a \in \mathcal{C}(u,v]} (a-y)^2 f(a) + c_2 (y-\nu)^2 \Big), \tag{31}$$

for $0 \leq u < v \leq N$. Note that the set $\mathcal{B} \cap [a_u, a_{v+1}]$ is non-empty in virtue of condition $C2$ (which is also satisfied when $\mathcal{B} = \mathbb{R}$). According to the above discussion, in order to solve (29) it is sufficient to solve

$$\min_{1 \leq k \leq K} \min_{\mathbf{t}^k \in \mathcal{T}_k} \sum_{i=1}^{k} w'(t'_{i-1}, t'_i). \tag{32}$$

Notice that the inner minimization in (32) is equivalent to the minimum weight k-edge path problem in the WDAG $G' = (V, E, w')$. Now it can be shown using similar arguments as in the proof of Lemma 4 in [20], that the edge weights of this graph satisfy the Monge property. This result is stated in the following lemma, whose proof is deferred to Appendix C.

Lemma 5. *The weights of edges of G' fulfill the Monge property [31], i.e., for all $0 \leq u_1 \leq u_2 < v_1 \leq v_2 \leq N$, the following relation holds*

$$w'(u_1, v_1) + w'(u_2, v_2) \leq w'(u_1, v_2) + w'(u_2, v_1).$$

Proposition 3. *Problem (29) can be solved in $O(KN)$ running time.*

[5]Another difference between the two problems is that in the noiseless case the optimal quantizer does not contain empty cells. Therefore, the corresponding graph problem is the minimum weight K-edge path problem. However, the dynamic programming solution to this problem still needs to compute the minimum weight k-edge path ending in every node v, for every $k, 1 \leq k \leq K$, as in the noisy channel case.

Proof: Let us denote by $\bar{y}(u,v)$ the value of y achieving the minimum in (31) for every $0 \leq u < v \leq N$. Then the following holds

$$\bar{y}(u,v) = \begin{cases} \hat{y}(\mathcal{C}(u,v]), & \text{if } a_u < \hat{y}(\mathcal{C}(u,v]) < a_{v+1} \\ \min\left(\mathcal{B} \cap [a_u, a_{v+1}]\right), & \text{if } \hat{y}(\mathcal{C}(u,v]) \leq a_u \\ \max\left(\mathcal{B} \cap [a_u, a_{v+1}]\right), & \text{if } \hat{y}(\mathcal{C}(u,v]) \geq a_{v+1} \end{cases}.$$

The aforementioned observations imply that, after having computed and stored the values $P(0,n]$, $m_1(0,n]$ and $m_2(0,n]$, for $1 \leq n \leq N$, the computation of each weight $w'(u,v)$ can still be carried out in $O(1)$ time.

Let us denote now by $\bar{W}'_k(v)$ the weight of the minimum weight k-edge path in G' from the source to vertex v, for each $k, 1 \leq k \leq K$, and each vertex $v \in V, v \geq k$. Then the following recurrence relation holds

$$\bar{W}'_k(v) = \min_{k-1 \leq u < v} \left(\bar{W}'_{k-1}(u) + w'(u,v)\right). \tag{33}$$

Based on Lemma 5 problem (33) can be solved for fixed k and all v in $O(N)$ time using the so-called SMAWK algorithm developed in [32]. This amounts to $O(KN)$ operations over all values of k. Since problem (32) is equivalent to computing the minimum of $\bar{W}'_k(v)$ over all $k, 1 \leq k \leq K$, and $v \in V, v \geq k$, the claim follows. ∎

Before ending this section we would like to point out that condition $C1$ was imposed in the case of a discrete reconstruction alphabet $\mathcal{B}$ in order to allow for the rounding operation in (28) to be performed in $O(1)$ time, leading further to the computation of each edge weight in $O(1)$ time as well. On the other hand, if $\mathcal{B}$ is an arbitrary finite set satisfying only condition $C2$, then the rounding operation in (28) can be performed in $O(\log|\mathcal{B}|)$ time. The computation of each edge weight for the graph G' takes then $O(\log|\mathcal{B}|)$ time leading to an overall time complexity of $O(KN\log|\mathcal{B}|)$ for solving problem (29).

V. Experimental Results

In this section we present experimental results to illustrate the suboptimality of the iterative algorithm for certain distributions, as well as its global optimality when the sufficient conditions for the uniqueness of a local optimum are satisfied.

First we present an example when the iterative algorithm fails to output the globally optimal solution. For this we consider a pdf defined on the interval $[-5,5]$ such that $f(x) = 2/7$ for all $x \in [-5,-3]$, $f(x) = 0$ for all $x \in [-3,2]$ and $f(x) = 1/7$ for all $x \in [2,5]$. The mean of the pdf is thus $\nu = -0.7857$ and the variance is 14.2874. The channel is a binary symmetric channel (BSC) with bit error rate (BER) of 0.01, and $K = 4$. We applied the iterative algorithm with the following initial partitions:

a) $(-5, -4.99, -4.98, -4.97, 5)$;
b) $(-5, -4.5, -4, -3.5, 5)$
c) $(-5, -2.5, 0, 2.5, 5)$;
d) $(-5, 2.75, 3.5, 4.25, 5)$.

The output partition and the EED $\bar{d}$ in each case are

a) $(-5, -2.374, -0.786, 1.324, 5)$; $\bar{d} = 0.9924$;
b) $(-5, -3.854, -2.07, 1.324, 5)$, $\bar{d} = 0.9924$;
c) $(-5, -2.374, 0.855, 3.244, 5)$; $\bar{d} = 0.9470$;
d) $(-5, -2.374, 0.855, 3.244, 5)$; $\bar{d} = 0.9470$.

On the other hand, the output of the globally optimal algorithm proposed in Section IV is the partition $(-5, -3.854, -3, 3.244, 5)$ and $\bar{d} = 0.8743$. The discretization of the pdf was obtained by applying a uniform quantizer with step size of 0.0001. As it can be seen, the iterative algorithm does not find the globally optimal solution in either of the four cases.

Additionally, we have considered a Gaussian distribution with 0 mean and variance 1, truncated to the interval $[-5,5]$ and ran both design algorithms for $K = 16$ and BER=$0.01, 0.05, 0.1, 0.2$. For the globally optimal algorithm we used a prequantization with a step size of 0.001. Since the Gaussian distribution satisfies the sufficient conditions for the uniqueness of a locally optimal solution according to Corollary 1, we expect that the two algorithms generate very similar results. Indeed, for all four BER's the EED values output by the two algorithms are very close with an absolute difference smaller than 10^{-7}. Specifically, the EED is $0.1, 0.322, 0.517, 0.773$ in the four cases, respectively. An interesting observation, however, is that, while the quantizers output by the iterative algorithm have 16 nonempty cells in all four cases, the algorithm for the discrete source outputs a quantizer with $16, 14, 10$ and 8 nonempty cells, for the four BER values, respectively. The reason is that as the BER increases the probability of the cells that are close to the mean in the optimal quantizer for the continuous source decreases and at some point becomes lower than the probability of a single sample of the discretized source.

VI. Conclusion

Existing algorithms for joint source-channel quantizer design iteratively optimize the encoder, respectively the decoder, while keeping the other component fixed. They can guarantee only a locally optimal solution in general and sufficient conditions for the global optimality of the solution are not known. In this work we address the design of noisy channel K-level scalar quantizer under the assumption of random index assignment. We first find sufficient conditions for the uniqueness of a local optimum, which thus becomes a global optimum. Furthermore, we present a globally optimal dynamic programming algorithm for general discrete distributions. A monotonicity property is additionally proved which allows for the acceleration of the solution algorithm to $O(KN)$ running time, where N is the size of the source alphabet.

Appendix A

In this appendix we present the proof of Lemma 1. The notations and observations made at the beginning of Section III will be used here.

Proof of Lemma 1: Let $V \leq a \leq \nu \leq b \leq W$, with $a < b$. Consider the function $h : [a,b] \to \mathbb{R}$ defined as follows $h(x) \triangleq cost(a,b) - cost(a,x) - cost(x,b)$. Then it can be easily verified that h is continuous on $[a,b]$ and $h(a) = h(b) = 0$. Additionally, h is differentiable on (a,b). We would like to

evaluate the derivative $h'(x)$ for $x \in (a,b)$. For this denote first for $V \leq \alpha < \beta \leq W$ and $y \in \mathbb{R}$,

$$F(\alpha,\beta,y) \triangleq c_1 \int_\alpha^\beta (t-y)^2 f(t)dt + c_2(y-\nu)^2.$$

Clearly, the function F is differentiable in α for $\alpha > V$, in β for $\beta < W$ and in y for $y \in \mathbb{R}$, and the following hold

$$F_1(\alpha,\beta,y) \triangleq \frac{\partial F}{\partial \alpha}(\alpha,\beta,y) = -c_1(\alpha-y)^2 f(\alpha),$$
$$\text{for } V < \alpha < \beta \leq W, y \in \mathbb{R},$$
$$F_2(\alpha,\beta,y) \triangleq \frac{\partial F}{\partial \beta}(\alpha,\beta,y) = c_1(\beta-y)^2 f(\beta),$$
$$\text{for } V \leq \alpha < \beta < W, y \in \mathbb{R}.$$

Additionally, let $F_3(\alpha,\beta,y) \triangleq \frac{\partial F}{\partial y}(\alpha,\beta,y)$. Then one has for all $V \leq a < b \leq W$,

$$F_3(\alpha,\beta,g(\alpha,\beta)) = 0.$$

Now let us compute $h'(x)$

$$\begin{aligned} h'(x) &= -\frac{\partial cost(a,x)}{\partial x} - \frac{\partial cost(x,b)}{\partial x} \\ &= -\frac{\partial F(a,x,g(a,x))}{\partial x} - \frac{\partial F(x,b,g(x,b))}{\partial x} \\ &= -F_2(a,x,g(a,x)) - F_3(a,x,g(a,x))g_2(a,x) - \\ &\quad F_1(x,b,g(x,b)) - F_3(x,b,g(x,b))g_1(x,b) \\ &= c_1 f(x)((g(x,b)-x)^2 - (g(a,x)-x)^2) \\ &= c_1 f(x)\left(g(x,b)-g(a,x)\right)\left(g(x,b)+g(a,x)-2x\right). \end{aligned}$$

Next we need to differentiate between the following three cases: 1) $\mu(a,b) > \nu$, 2) $\mu(a,b) < \nu$ and 3) $\mu(a,b) = \nu$.

Case 1. $\mu(a,b) > \nu$. Then $g(a,b) > \nu$ and $a > V$. Notice that

$$lim_{x \searrow a} h'(x) = c_1 f(a)(g(a,b)-\nu)(g(a,b)+\nu-2a).$$

The fact that $g(a,b) > \nu \geq a$ implies that $g(a,b)+\nu-2a > 0$. Using further the fact that $f(a) > 0$ and $g(a,b) > \nu$, we obtain that $lim_{x \searrow a} h'(x) > 0$. Since $h'(x)$ is continuous on (a,b) it further follows that there is some $0 < \epsilon < b-a$ such that $h'(x) > 0$ for all $x \in (a, a+\epsilon)$. Aided by the fact that h is continuous on $[a, a+\epsilon]$ we further obtain that $h(x) > h(a) = 0$ for all $x \in (a, a+\epsilon)$.

Case 2. $\mu(a,b) < \nu$. In this case $g(a,b) < \nu \leq b < W$. Then

$$lim_{x \nearrow b} h'(x) = c_1 f(b)(\nu - g(a,b))(\nu + g(a,b) - 2b) < 0.$$

Then there is some $0 < \epsilon < b-a$ such that $h'(x) < 0$ for all $x \in (b-\epsilon, b)$. Further, it follows that $h(x) > h(b) = 0$ for all $x \in (b-\epsilon, b)$.

Case 3. $\mu(a,b) = \nu$. Since $a < \mu(a,b) < b$ it follows that $a < \nu < b$. Note that $g(a,x) - x$ is a continuous function of x for $x \in (a,b)$. Since $lim_{x \searrow a}(g(a,x)-x) = \nu - a > 0$ it follows that there is some $0 < \epsilon < \nu - a$ such that $g(a,x) > x$ for all $x \in [a, a+\epsilon]$. For $x \in (a, a+\epsilon]$ we have $\mu(x,b) > \mu(a,b) = \nu$ and $\mu(a,x) < x < \nu$ because μ is increasing in both arguments. Therefore, it follows that $g(x,b) > \nu > g(a,x) > x$. Corroborating with $f(x) > 0$ we further obtain that $h'(x) > 0$ for all $x \in (a, a+\epsilon]$. Using the fact that h is continuous on $[a, a+\epsilon]$ it follows that $h(x) > h(a) = 0$ for all $x \in (a, a+\epsilon]$. Thus, the proof is complete. ■

APPENDIX B

In this appendix we present the proof of Lemma 3. For this we need another auxiliary result.

Lemma 6. *Assume that relation (18) is true for all $V < a < b < W$. Let $V < a_0 < b_0 < W$, $V < a_1 < b_1 < W$ such that $a_0 \leq a_1$, $a_0 < g(a_0,b_0) < b_0$, $a_1 < g(a_1,b_1) < b_1$. Then the following hold.*

i) *If $a_1 - a_0 \leq b_1 - b_0$ then $b_0 - g(a_0,b_0) \leq b_1 - g(a_1,b_1)$. Moreover, if $b_1 > b_0$ then $b_0 - g(a_0,b_0) < b_1 - g(a_1,b_1)$.*

ii) *If $a_1 - a_0 > b_1 - b_0$ and $a_1 > a_0$ then $g(a_0,b_0) - a_0 > g(a_1,b_1) - a_1$.*

Proof: Define the functions $a(t) = a_0 + (a_1 - a_0)t$ and $b(t) = b_0 + (b_1 - b_0)t$ for all $t \in [0,1]$. Clearly, $a(t) < b(t)$ and $a(t), b(t)$ and $g(a(t),b(t))$ are differentiable and $a'(t) \geq 0$ for all $t \in [0,1]$.

Let us prove now claim i). Assume that $a_1 - a_0 \leq b_1 - b_0$. Then one has $0 \leq a'(t) \leq b'(t)$ for all $t \in [0,1]$. Using (10) and the fact that $g(a_0,b_0) > a_0$ and $g(a_1,b_1) > a_1$, it follows that $g_1(a_0,b_0) > 0$ and $g_1(a_1,b_1) > 0$. Now define $t_0 \in [0,1]$ in the following way. If $g_1(a(t),b(t)) > 0$ for all $t \in [0,1]$ then let $t_0 = 0$. Otherwise, let $t_0 = \sup\{t \in (0,1) | g_1(a(t),b(t)) \leq 0\}$. In the latter case, the continuity of g_1 and the fact that $g_1(a_1,b_1) > 0$ imply that $t_0 < 1$. Thus, we have $g_1(a(t),b(t)) > 0$ for all $t \in (t_0,1]$, while $g_1(a(t_0),b(t_0)) = 0$ when $t_0 > 0$. The inequality $a'(t) \leq b'(t)$ implies that $a'(t)g_1(a(t),b(t)) \leq b'(t)g_1(a(t),b(t))$ for all $t \in [t_0,1]$. Using further (18) and the fact that $b'(t) \geq 0$ it follows that

$$\begin{aligned} &b'(t) - a'(t)g_1(a(t),b(t)) - b'(t)g_2(a(t),b(t)) \geq \\ &\quad b'(t)\left(1 - g_1(a(t),b(t)) - g_2(a(t),b(t)\right) \geq 0 \end{aligned} \tag{34}$$

for all $t \in [t_0,1]$. Then

$$\begin{aligned} &(b_1 - g(a_1,b_1)) - (b(t_0) - g(a(t_0),b(t_0))) = \\ &\quad \int_{t_0}^1 \left(b(t) - g(a(t),b(t))\right)' dt = \\ &\quad \int_{t_0}^1 b'(t) - a'(t)g_1(a(t),b(t)) - \\ &\quad\quad b'(t)g_2(a(t),b(t))dt \geq 0. \end{aligned} \tag{35}$$

When $t_0 = 0$ the inequality $b_0 - g(a_0,b_0) \leq b_1 - g(a_1,b_1)$ follows immediately. When $t_0 > 0$, the fact that $g_1(a(t_0),b(t_0)) = 0$ implies that $g(a(t_0),b(t_0)) = a_0$ in view of (10) . Then

$$b(t_0) - g(a(t_0),b(t_0)) = b(t_0) - a(t_0) =$$
$$b_0 - a_0 + t_0(b_1 - b_0 - a_1 + a_0) \geq b_0 - a_0 > b_0 - g(a_0,b_0).$$

Using further (35) one obtains that $b_0 - g(a_0,b_0) \leq b_1 - g(a_1,b_1)$. Assume now that $b_1 > b_0$. Then $b'(t) > 0$ for all $t \in [0,1]$ implying that the last inequality in (34) and, consequently, the inequality in (35) are strict. It further follows that $b_0 - g(a_0,b_0) < b_1 - g(a_1,b_1)$.

Let us prove now claim ii). Assume that the inequalities $a_1 - a_0 > b_1 - b_0$ and $a_1 > a_0$ hold. These imply that $a'(t) > b'(t)$ and $a'(t) > 0$ for all $t \in [0,1]$. Additionally, notice that, in view of (11), one has $g_2(a_0,b_0) > 0$ and $g_2(a_1,b_1) > 0$. Define now t_1 as follows. If $g_2(a(t),b(t)) > 0$

for all $t \in [0,1]$ then let $t_1 = 1$. Otherwise, let $t_1 = \inf\{t \in (0,1) | g_2(a(t), b(t)) \leq 0\}$. The fact that $g_2(a_0, b_0) > 0$ and the continuity of g_2 imply that $t_1 > 0$. It follows that $g_2(a(t), b(t)) > 0$ for all $t \in [0, t_1)$, while $g_2(a(t_1), b(t_1)) = 0$ when $t_1 < 1$. Based on the fact that $a'(t) > b'(t)$ we further obtain that $a'(t)g_2(a(t), b(t)) > b'(t)g_2(a(t), b(t))$ for all $t \in [0, t_1)$. Using further (18) and the fact that $a'(t) > 0$ it follows that

$$a'(t)g_1(a(t), b(t)) + b'(t)g_2(a(t), b(t)) - a'(t) < a'(t)\left(g_1(a(t), b(t)) + g_2(a(t), b(t)) - 1\right) < 0$$

for all $t \in [0, t_1)$. Thus, we further obtain

$$\begin{gathered}(g(a(t_1), b(t_1)) - a(t_1)) - (g(a_0, b_0) - a_0) = \\ \int_0^{t_1} \left(g(a(t), b(t) - a(t)\right)' dt = \\ \int_0^{t_1} a'(t)g_1(a(t), b(t)) + \\ b'(t)g_2(a(t), b(t)) - a'(t)dt < 0. \end{gathered} \tag{36}$$

The claim follows immediately when $t_1 = 1$. If $t_1 < 1$, then the equality $g_2(a(t_1), b(t_1)) = 0$ implies that $g(a(t_1), b(t_1)) = b(t_1)$ in view of (11). Then

$$\begin{gathered}g(a(t_1), b(t_1)) - a(t_1) = b(t_1) - a(t_1) = \\ b_1 - a_1 + (1 - t_1)(a_1 - a_0 - b_1 + b_0) > \\ b_1 - a_1 > g(a_1, b_1) - a_1.\end{gathered}$$

Using further (36) the claim follows completing the proof. ■

Proof of Lemma 3: $T1$) Let $V < b_0 < b_1 < W$. Define $b(t) = b_0 + (b_1 - b_0)t$, for $t \in [0,1]$. Then one obtains

$$\begin{gathered}(b_1 - g(V, b_1)) - (b_0 - g(V, b_0)) = \\ \int_0^1 (b(t) - g(V, b(t))' dt = \\ \int_0^1 b'(t)(1 - g_2(V, b(t)))dt > 0\end{gathered}$$

since $b'(t) > 0$ and $1 - g_2(V, b(t)) > 0$ for all $t \in [0,1]$ according to the second inequality in (17). Thus, the claim is proved.

$T2$) This claim follows immediately from Lemma 6 point i) by letting $a_0 = a_1 = a$.

$T3$) Let $V < a_0 < a_1 < W$. Define $a(t) = a_0 + (a_1 - a_0)t$ for $t \in [0,1]$. Then one has

$$\begin{gathered}(g(a_1, W) - a_1) - (g(a_0, W) - a_0) = \\ \int_0^1 (g(a(t), W) - a(t))' dt = \\ \int_0^1 a'(t)(g_1(a(t), W) - 1))dt < 0\end{gathered}$$

due to $a'(t) > 0$ and $g_1(a(t), W) < 1$, for $t \in [0,1]$, proving the claim.

$T4$) Let $V < a_0 < b_0 < W$, $V < a_1 < b_1 < W$, such that $a_0 < a_1$, $a_0 < g(a_0, b_0) < b_0$ and $a_1 < g(a_1, b_1) < b_1$ and $g(a_0, b_0) - a_0 < g(a_1, b_1) - a_1$. Then Lemma 6 point ii) implies that $0 < a_1 - a_0 \leq b_1 - b_0$. Further point i) of Lemma 6 leads to the conclusion that $b_0 - g(a_0, b_0) < b_1 - g(a_1, b_1)$, concluding the proof. ■

APPENDIX C

In this appendix we present the proof of Lemma 5. Before proceeding to the proof consider the following notation. For every $0 \leq u < v \leq N$, and $y \in \mathbb{R}$ denote by

$$\Omega(u, v, y) \triangleq c_1 \sum_{a \in \mathcal{C}(u,v]} (a - y)^2 f(a) + c_2 (y - \nu)^2.$$

It follows that for $0 \leq u < v \leq N$, one has

$$w'(u, v) = \Omega(u, v, \bar{y}(u, v)).$$

Proof of Lemma 5: Notice first that the cases when $u_1 = u_2$ or $v_1 = v_2$ are trivial. Therefore, let us assume that $0 \leq u_1 < u_2 < v_1 < v_2 \leq N$. Further, denote $\eta_1 = \bar{y}(u_2, v_1)$ and $\eta_2 = \bar{y}(u_1, v_2)$. The definition of η_1 implies that

$$a_{u_2} \leq \eta_1 \leq a_{v_1+1}. \tag{37}$$

Additionally, one has

$$w'(u_1, v_2) + w'(u_2, v_1) = \Omega(u_1, v_2, \eta_2) + \Omega(u_2, v_1, \eta_1). \tag{38}$$

To proceed we need to consider two cases.

Case 1. $\eta_1 \leq \eta_2$. Relations (37) and the fact that $u_1 < u_2$ imply that $a_{u_1} < \eta_1 \leq a_{v_1+1}$. Using further the definition of $w'(u_1, v_1)$ and of $\Omega(u_1, v_1, \eta_1)$ one obtains that

$$w'(u_1, v_1) \leq \Omega(u_1, v_1, \eta_1). \tag{39}$$

Further, the fact that $\eta_1 \leq \eta_2$ together with (37) imply that $a_{u_2} \leq \eta_2$. Corroborating with the definition of η_2 it follows that $a_{u_2} \leq \eta_2 \leq a_{v_2+1}$. Therefore, one has

$$w'(u_2, v_2) \leq \Omega(u_2, v_2, \eta_2). \tag{40}$$

Relations (38-40) imply that, in order to establish the validity of Lemma 5, it is sufficient to prove the following inequality

$$\Omega(u_1, v_1, \eta_1) + \Omega(u_2, v_2, \eta_2) \leq \Omega(u_1, v_2, \eta_2) + \Omega(u_2, v_1, \eta_1).$$

Upon applying the definition of Ω and canceling the like terms, the above inequality becomes equivalent to

$$\begin{gathered}\sum_{a \in \mathcal{C}(u_1, v_1]} (a - \eta_1)^2 f(a) + \sum_{a \in \mathcal{C}(u_2, v_2]} (a - \eta_2)^2 f(a) \leq \\ \sum_{a \in \mathcal{C}(u_1, v_2]} (a - \eta_2)^2 f(a) + \sum_{a \in \mathcal{C}(u_2, v_1]} (a - \eta_1)^2 f(a).\end{gathered}$$

After expanding the summations and canceling the like terms, the above relation becomes

$$\sum_{a \in \mathcal{C}(u_1, u_2]} (a - \eta_1)^2 f(a) \leq \sum_{a \in \mathcal{C}(u_1, u_2]} (a - \eta_2)^2 f(a). \tag{41}$$

Notice that the fact that $a_{u_2} \leq \eta_1 \leq \eta_2$ and that $f(a) > 0$ for all a, imply that $(a - \eta_1)^2 f(a) \leq (a - \eta_2)^2 f(a)$ for all $a_{u_1} < a \leq a_{u_2}$, thus establishing the validity of (41). This observation completes the proof of Case 1.

Case 2. $\eta_1 > \eta_2$. The proof for this case is symmetrical. ■

REFERENCES

[1] A. Kurtenbach and P. Wintz, "Quantizing for noisy channels", *IEEE Trans. Comm. Techn.*, vol. COM-17, no. 2, pp. 291-302, Apr. 1969.

[2] H. Kumazawa, M. Kasahara, and T. Namekawa, "A construction of vector quantizers for noisy channels", *Electron. Eng. Japan*, vol. 67-B, no. 4, pp. 39-47, 1984.

[3] N. Farvardin and V. Vaishampayan, "Optimal quantizer design for noisy channels: an approach to combined source-channel coding", *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 827-838, Nov. 1987.

[4] N. Farvardin and V. Vaishampayan, "On the performance and complexity of channel-optimized vector quantizers", *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 155-160, Jan. 1991.
[5] D. Miller and K. Rose, "Combined source-channel vector quantization using deterministic annealing", *IEEE Trans. Commun.*, vol. 42, no. 2, pp. 347-356, Feb. 1994.
[6] S. Gadkari and K. Rose, "Robust vector quantizer design by noisy channel relaxation", *IEEE Trans. Commun.*, vol. 47, no. 8, pp. 1113-1116, Aug. 1999.
[7] H. Jafarkhani and N. Farvardin, "Design of channel-optimized vector quantizers in the presence of channel mismatch", *IEEE Trans. Commun.*, vol. 48, no. 1, pp. 118-124, Jan. 2000.
[8] I. Kozintsev and K. Ramchandran, "Multiresolution joint source-channel coding using embedded constellations for power-constrained time-varying channels", *Proc. ICASSP'1996*, Atlanta, Georgia, pp. 2343-2346, May 1996.
[9] J. Max, "Quantizing for minimum distortion", *IRE Trans. Inform. Theory*, vol. IT-6, no. 1, pp. 7-12, Jan. 1960.
[10] S. P. Lloyd, "Least squares quantization in PCM", *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 129-137, Mar. 1982.
[11] P. E. Fleischer, "Sufficient conditions for achieving minimum distortion in a quantizer", *IEEE Int. Conv. Rec.*, 1964, part 1, pp. 104-111.
[12] J. C. Kieffer, "Exponential rate of convergence for Lloyd's Method I", *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 205-210, Mar. 1982.
[13] A. V. Trushkin, "Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions", *IEEE Trans. Inform. Theory*, vol. 28, no. 2, pp. 187-198, Mar. 1982.
[14] J. C. Kieffer, "Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function", *IEEE Trans. Inform. Theory*, vol. IT-29, no. 1, pp. 42-47, Jan. 1983.
[15] A. V. Trushkin, "Monotony of Lloyd's Method II for log-concave density and convex error weighting function", *IEEE Trans. Inform. Theory*, vol. 30, no. 2, pp. 380-383, March 1984.
[16] S. Dumitrescu and X. Wu, "On properties of locally optimal multiple description scalar quantizers with convex cells," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5591–5606, Dec. 2009.
[17] J. D. Bruce, "Optimum quantization", Sc. D. thesis, M. I. T., May 14, 1964.
[18] D. K. Sharma, "Design of absolutely optimal quantizers for a wide class of distortion measures", *IEEE Trans. Inform. Theory*, vol. IT-24, no. 6, pp. 693-702, Nov. 1978.
[19] X. Wu, "Optimal Quantization by matrix searching", *Journal of Algorithms*, vol. 12, no. 4, pp. 663-673, Dec. 1991.
[20] X. Wu and K. Zhang, "Quantizer monotonicities and globally optimal scalar quantizer design", *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 1049-1053, May 1993.
[21] A. Aggarwal, B. Schieber, and T. Tokuyama, "Finding a minimum-weight k-link path in graphs with the concave the property and applications", *Discrete Comput. Geometry*, vol. 12, no. 1, pp. 263-280, Dec. 1994.
[22] B. Schieber, "Computing a minimum-weight k-link path in graphs with the concave the property", *Proc. ACM-SIAM Symp. Algorithms'95*, pp. 405-411.
[23] S. Dumitrescu and X. Wu, "Algorithms for optimal multi-resolution quantization", *J. Algorithms*, vol. 50, no. 1, pp. 1-22, Jan. 2004.
[24] D. Muresan and M. Effros, "Quantization as histogram segmentation: optimal scalar quantizer design in network systems", *IEEE Trans. Inform. Theory*, vol. 54, no. 1, pp. 344-366, Jan. 2008.
[25] S. Dumitrescu and X. Wu, "Optimal two-description scalar quantizer design", *Algorithmica*, vol. 41, no. 4, pp. 269-287, Feb. 2005.
[26] S. Dumitrescu and X. Wu, "Lagrangian optimization of two-description scalar quantizers", *IEEE Trans. Inform. Theory*, vol. 53, no. 11, pp. 3990-4012, Nov. 2007.
[27] A. Gyorgy, and T. Linder, G. Lugosi, "Tracking the best quantizer", *IEEE Trans. Inform. Theory*, vol. 54, no. 4, pp. 1604-1625, Apr. 2008.
[28] X. Yu, H. Wang, and E. -H. Yang, "Design and analysis of optimal noisy channel quantization with random index assignment" *IEEE Trans. Inform. Theory*, vol. 56, no. 11, pp. 5796–5801, Nov. 2010.
[29] F. Teng, E. -H. Yang, and X. Yu, Optimal multiresolution quantization for broadcast channels with random index assignment, *Proc. IEEE Intern. Symp. Inf. Theory*, Austin, TX, pp. 181-185, Jun. 2010.
[30] J. Ho and E. -H. Yang, "Designing optimal multiresolution quantizers with error detecting codes", *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3588-3599, July 2013.
[31] R. E. Burkard, B. Klinz, and R. Rudolf, "Perspectives of Monge properties in optimization", *Discrete Applied Mathematics* vol. 70, no. 2, pp. 95-161, Sept. 1996.
[32] A. Aggarval, M. Klave, S. Moran, P. Shor, and R. Wilber, "Geometric applications of a matrix-searching algorithm", *Algorithmica*, vol. 2, no. 1-4, pp.195-208, Nov. 1987.
[33] X. Wu, "On convergence of Lloyd Method I", *IEEE Trans. Inform. Theory*, vol. 38, no. 1, pp. 171-174, Jan. 1992.

PLACE PHOTO HERE

Sorina Dumitrescu (M'05-SM'13) received the B.Sc. and Ph.D. degrees in mathematics from the University of Bucharest, Romania, in 1990 and 1997, respectively. From 2000 to 2002 she was a Postdoctoral Fellow in the Department of Computer Science at the University of Western Ontario, London, Canada. Since 2002 she has been with the Department of Electrical and Computer Engineering at McMaster University, Hamilton, Canada, where she held Postdoctoral, Research Associate, and Assistant Professor positions, and where she is currently an Associate Professor. Her current research interests include multimedia coding and communications, network-aware data compression, multiple description codes, joint source-channel coding, signal quantization. Her earlier research interests were in formal languages and automata theory. Dr. Dumitrescu held an NSERC University Faculty Award during 2007-2012.