

# Universal Multiple Description Scalar Quantization: Analysis and Design

Chao Tian, *Student Member, IEEE*, and Sheila S. Hemami, *Senior Member, IEEE*

**Abstract**—This paper introduces a new high-rate analysis of the multiple description scalar quantizer (MDSQ) with balanced descriptions. The analysis provides insight into the structure of the MDSQ, suggesting the nonoptimality of uniform central quantizer cell lengths, as well as a method to approximate optimal cell lengths. For both level-constrained and entropy-constrained MDSQ, new upper bounds on the granular distortion for sources with smooth probability density functions (pdfs) are derived under the mean-squared error measure, which are 0.4 dB lower than previous results. Based on the insights, a universal multiple description scalar quantizer (UMDSQ) is proposed which, at high rate, can achieve nearly the same performance as the fully optimized entropy-constrained MDSQ (ECMDSQ), without requiring extensive training. The proposed UMDSQ has only two control parameters, and a continuum of tradeoff points between the central and side distortions can be achieved as the two parameters are varied.

**Index Terms**—Asymptotic analysis, multiple description, scalar quantization.

## I. INTRODUCTION

THE multiple description (MD) problem was first introduced at the 1979 IEEE Information Workshop by Gersho, Wittenhausen, Wolf, Wyner, Ziv, and Ozarow. The problem can be defined as follows: consider a stochastic process  $X_1, X_2, X_3, \dots$  where the  $X_i$ 's are independent and identically distributed (i.i.d.) according to some known distribution  $p(x)$ . Two descriptions must be generated at rates  $R_1$  and  $R_2$ , respectively. Three single-letter distortion measures  $d_1, d_2, d_0$  are given. The problem is to find the possible reconstruction distortions simultaneously, by using only description 1, only description 2, and both descriptions, which are usually denoted as  $D_1, D_2$ , and  $D_0$ , respectively. The problem is difficult in the sense that if the two individual descriptions each achieve good reconstruction performance, then they must be very alike, and the joint reconstruction using both does not improve significantly over the only-one-description reconstruction.

The MD problem is an information-theoretic representation of a communication system with channel failures. Consider the case when a transmission uses two distinct channels. At the transmitter side, the encoder does not know which channel(s)

Manuscript received February 4, 2003; revised April 29, 2004. This work was supported in part by the Multidisciplinary University Research Initiative (MURI) under the Office of Naval Research Contract N00014-00-1-0564. The material in this paper was presented in part at the Data Compression Conference, Snowbird, UT, March 2003, and at the 37th Annual Conference on Information Sciences and Systems, Baltimore MD, March 2003.

The authors are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853 USA (email: ctian@ece.cornell.edu; hemami@ece.cornell.edu).

Communicated by R. Zamir, Associate Editor for source Coding.

Digital Object Identifier 10.1109/TIT.2004.833344

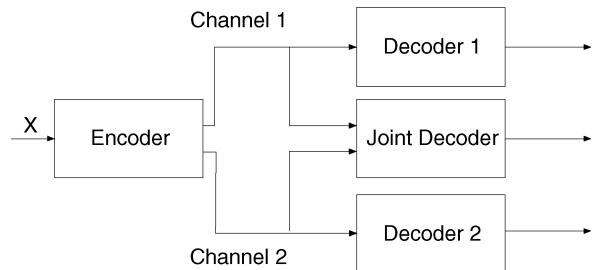


Fig. 1. Two-channels and three-receivers model.

will fail, so it must encode the source with only the knowledge that the channel(s) may fail with certain probability; at the receiver, the decoder knows which description(s) it is receiving, and thus decodes accordingly. In this setting, the distortion resulting from using both descriptions is called the *central distortion*, and the distortion resulting from using one description is called the *side distortion*. These terms come from the simplified “two channels and three receivers” model [1], as depicted in Fig. 1. The joint decoder is the *central decoder*; decoders 1 and 2 are the *side decoders*. When the two descriptions are *balanced*, the two side descriptions have the same rate and generate the same distortion. This paper assumes balanced descriptions.

An achievable rate region for the MD problem is defined as a set of rates  $(R_1, R_2)$  that is sufficient to achieve the fixed distortions of  $D_1, D_2$ , and  $D_0$ . El Gamal and Cover [2] gave such an achievable rate region for a memoryless source with a single-letter distortion measure, and it was conjectured that this region is tight (i.e., it is not only sufficient, but also necessary) for general sources and distortion measures. Ozarow [1] showed that it is tight for a memoryless Gaussian source and the mean-squared error measure. But Zhang and Berger [3] proved that this region is not tight generally for other sources and distortion measures. To date, a Gaussian source with the mean-squared error measure is the only case for which the region has been completely characterized for this problem. However, a set of outer and inner bounds for the MD problem under the mean-squared error measure was given by Zamir [4], and the outer bound was shown to be asymptotically tight under high resolution conditions.

As a practical system to achieve the MD property, MD scalar quantizer (MDSQ) [5] employs two steps, namely, a quantization step and an index assignment step. MDSQ can be optimized under two different constraints, which lead to level-constrained (also referred to as fixed-rate) MDSQ and entropy-constrained MDSQ (ECMDSQ). The asymptotic analysis of MDSQs [6] reveals that they are **exponentially** optimal at high rate in the rate-

distortion sense. An important assumption in the analysis presented in [6] is the uniformity of central quantizer cell lengths over a certain local area, which will be shown to be nonoptimal in this paper.

This paper introduces a new and more straightforward asymptotic analysis of the MDSQ for a special class of index assignments, which is also exponentially optimal in the rate-distortion sense. This analysis provides insight into the structure of the MDSQ, suggesting the nonoptimality of using uniform central quantizer cell lengths, as well as methods to approximate optimal cell lengths. New upper bounds on the granular distortion of level-constrained and entropy-constrained MDSQ are thus derived for sources with smooth probability density functions (pdfs), under the mean-squared error measure. Based on the insights, a class of MDSQs is proposed which is universal in nature and can achieve almost the same performance as the fully optimized ECMDSQ [7] at high rate, without requiring extensive training.

This paper is organized as follows. In Section II, previous design and analysis of MDSQ are briefly reviewed, and some unsolved problems are discussed. Section III introduces the structure of the proposed two-stage quantizer. Section IV presents the asymptotic analysis for this structure, and applies this analysis to the problem of optimizing two-stage MDSQ. Section V discusses the granular distortions for level-constrained and entropy-constrained MDSQ in the context of our analysis method. Section VI introduces the universal MD scalar quantizer (UMDSQ) and compares its performance with ECMDSQ with optimized codebooks. Section VII concludes the paper.

## II. MULTIPLE DESCRIPTION SCALAR QUANTIZER (MDSQ)

The design and analysis of MDSQs were given in a series of papers [5]–[7] by Vaishampayan *et al.*. In the following, related results are briefly reviewed and then the remaining unsolved problems are posed.

### A. Overview of MDSQ

The basic idea of MDSQ is to create two coarse side quantizers, each of which produces acceptable side distortion when used alone; the two coarse side quantizers are combined to produce a finer central quantizer, which provides lower distortion than the side quantizers. MDSQ uses two steps to achieve the MD property, namely, a central quantization step and an index assignment step. The system is depicted in Fig. 2.

In the first step, the encoder quantizes the sample  $x$  using the central quantizer, and generates a central quantizer index  $l$ ,  $l \in I$ , where  $I$  is the one-dimensional index space. Then the index is input to the index assignment block, which performs a mapping  $a : l \rightarrow (p, q)$ ,  $l \in I, (p, q) \in I^2$ , where  $I^2$  is the two-dimensional index space, and  $(p, q)$  are the first and second side quantizers' indices. In the side quantizers, a cell is not necessarily a continuous interval, but can be the union of (central quantizer) intervals. However, as noted in [5], a central quantizer cell should always be a continuous interval under the mean-squared error measure.

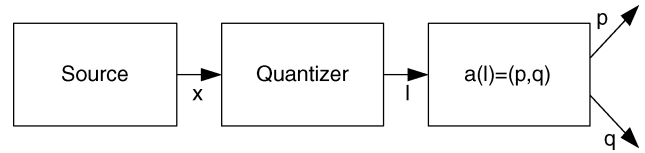


Fig. 2. Encoder diagram.

The mapping  $a$  can be represented as a matrix, called the index assignment matrix. Some of the elements in the matrix are assigned a number corresponding to the central quantizer index, and then the column and row indices of the elements are the side quantizer indices. Some of the elements in the matrix are not used; this is to achieve the desired central and side distortion tradeoffs. Generally, by using more diagonals in the index assignment matrix, better central distortion performance can be achieved, while sacrificing some side distortion performance. More details about the index assignment problem can be found in [5].

MDSQs can be classified by the constraints under which they are optimized. For level-constrained MDSQ, the optimization constraint is on the number of columns (and rows) in the index assignment matrix, or equivalently, the number of levels in the side quantizers. For ECMDSQ, the optimization constraint is on the entropy of the column (and row) indices in the index assignment matrix, or equivalently, the entropy of the side quantizer indices.

### B. Existing Performance Bounds for MDSQ

To bound the asymptotic performance of MDSQ, we will use the Shannon lower bound for MD [4], which is tight at high resolution. In the case of balanced description, the lower bound is given as

$$D_1 \geq P_x 2^{-2R}$$

$$D_0 \geq \frac{P_x 2^{-4R}}{1 - (\sqrt{\pi} - \sqrt{\Delta})^2}$$

where  $\pi = (1 - D_1/P_x)^2$ ,  $\Delta = D_1^2/P_x^2 - 2^{-4R}$ ,  $R$  is the rate for one description,  $P_x = (2\pi e)^{-1} 2^{2h(p)}$ , and  $h(p)$  is the differential entropy of the source,  $D_0$  and  $D_1$  are the central distortion and side distortion, respectively. Note since this lower bound is tight for high resolution, it can be used as the distortion-rate bound for the asymptotic analysis.

Through some algebra, it can be shown that when  $R \rightarrow \infty$ , if the side distortion is in the form of

$$D_1 = b 2^{-2(1-\eta)R} \quad (1)$$

for some  $b > 0$  and  $\eta \in (0, 1)$ , then the central distortion is given by

$$D_0 \geq \frac{P_x^2}{4b} 2^{-2(1+\eta)R} \quad (2)$$

which for the Gaussian source with unit variance yields

$$D_0 \geq 2^{-2(1+\eta)R}/4b \quad (3)$$

as given in [6].

The product of the central and side distortion bounds is nearly constant at a fixed rate, except for the special cases when  $\eta = 0$  or  $\eta = 1$  (see [6] for a more detailed discussion). This product is therefore used as the information-theoretic bound for the quantizers. It suggests there is a performance tradeoff between the central and side quantizers, which is simply abbreviated as the *tradeoff* throughout this paper.

In [6], the analysis of the MDSQs introduced a *compander* function, which is widely used in high-resolution quantizer analysis, on the central quantizer. The central and side mean-squared errors were given, respectively, by

$$D_{Q0} \geq \frac{1}{12N^2} \int_{-x_0}^{x_0} \frac{p(x)}{g^2(x)} dx \quad (4)$$

$$D_{Q1} \geq \frac{2\alpha_k(k+1)^2}{(2k+1)N^2} \int_{-x_0}^{x_0} \frac{p(x)}{g^2(x)} dx \quad (5)$$

where

$$\alpha_k = \sum_{u=1}^k u^2. \quad (6)$$

In the preceding equations,  $N$  is the total number of central quantizer levels;  $2k+1$  is the number of diagonals used in the index assignment matrix;  $[-x_0, x_0]$  is the support of the source samples;  $p(x)$  is the probability distribution of the source; and  $g(x)$  is the central quantization point density function [2]. For ECMDSQ, the rate of the quantizer was approximated by

$$H_Q = \log_2\left(\frac{N}{2k+1}\right) + \int_{-x_0}^{x_0} p(x) \log_2 g(x) dx + h(p) \quad (7)$$

where  $h(p)$  is the differential entropy of the source pdf and  $H_Q$  is the quantizer rate. It should be pointed out that the compander function introduced on the central quantizer implies that the central quantizer is uniform over a certain local area, but this assumption will later be shown to be not true in general. As such, (4) and (5) are nonoptimal in general.

Equations (4), (5), and (7) were used to show that the optimal ECMDSQ should have a uniform central quantizer, and then they were combined to derive the granular distortion of ECMDSQ in [6]. For sources with smooth pdfs, the gap between the product of the central and side distortions of ECMDSQ and that of the distortion-rate bound was shown to be 3.07 dB. (This difference between the quantizer performance and the distortion-rate bound at high rate will be referred to as the *granular gap* in this paper.) Applying similar methods to level-constrained MDSQ yielded a gap of 8.69 dB for the Gaussian source under the mean-squared error measure.<sup>1</sup> This result can be interpreted as implying that this granular performance of ECMDSQ is optimal, **if** the central and side quantizers are considered to be simply two scalar quantizers. With this assumption, the central and side quantizers should each be 1.53 dB away from the distortion-rate bound, and their product would thus be 3.07 dB

<sup>1</sup>When the decibel (dB) is used as the unit to measure the difference between the power  $P(A)$  and  $P(B)$  of two signals  $A$  and  $B$ , it is often defined as  $10 \log_{10}(\frac{P(A)}{P(B)})$ . Here, since we are comparing the product of powers, it would be more natural to use  $5 \log_{10}(\cdot)$  as the definition of the decibel. In this work, we still choose to use the former definition adopted in [5]–[7] for comparison purposes.

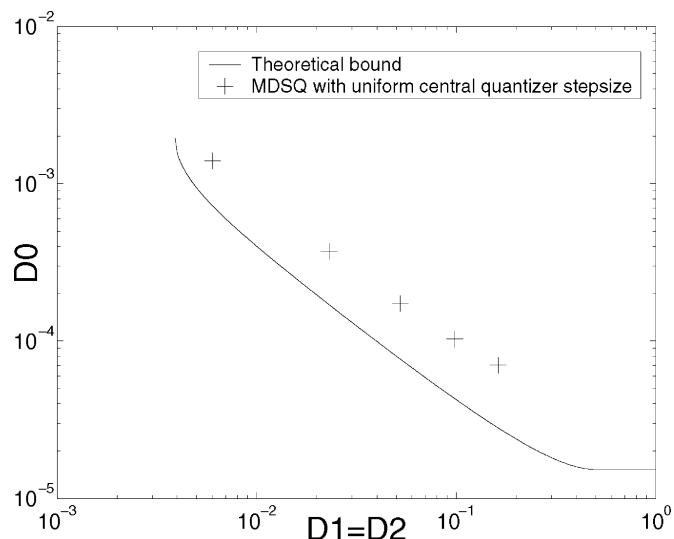


Fig. 3. ECMDSQ with a uniform central quantizer: discrete tradeoff points.

worse than that of the distortion-rate bound. However, we will show that the above explanation is not justified, by showing that the granular gap between the optimal ECMDSQ and the distortion-rate bound is actually smaller than 3.07 dB.

### C. The Remaining Unsolved Problems

The remaining problems can be categorized into two classes: inconsistencies between the asymptotic analysis and the performance in practice, and practical difficulty in the usage of MDSQ.

1) *Inconsistencies Between the Asymptotic Analysis and the Performance in Practice:* The results mentioned in the last subsection are based on the assumption that the central cells are nearly uniform over a certain local area, which allows an approximation of the reconstruction point density function as a continuous function and allows the use of an integral. However, it will be shown that this assumption is not valid in most cases; i.e., it is only valid when the constraint on the side distortion is so loose that it becomes trivial to satisfy. This assumption leads to the following discrepancies.

In the design of MDSQ [5], [7], it was shown that for a given  $k$ , MDSQ can achieve different tradeoff points, which are optimal in a rate-distortion sense, by adjusting the thresholds and reconstruction values. Subsequently, by varying the value of  $k$ , tradeoffs can be achieved continuously over a wide range. However, with a given  $k$ , using the expressions for the central and side distortions in the previous section ((4) and (5)), the central and side distortions can be minimized simultaneously by choosing the proper quantization point density function  $g(x)$  and the total number of central quantizer levels  $N$ . This implies there is only one rate-distortion optimal tradeoff point for a given  $k$  (which is achieved by using a uniform central quantizer for ECMDSQ, as asserted by (4), (5), and (7) that the optimal ECMDSQ should use). By varying the value of  $k$ , optimal tradeoffs can only be achieved discretely (see Fig. 3, and this “discrete” effect was also observed by Goyal and Kovačević [8]), which contradicts the aforementioned achievability of a continuum of tradeoffs.

2) *Practical Difficulty in the Usage of MDSQ*: Several authors have used MDSQ in their design of MD encoders for images [9], [10]. Practical usage of MDSQ poses a dilemma, requiring a choice between simple implementation with minimal opportunity for tradeoffs, or complex implementation allowing more tradeoffs.

- If the central quantizer uses only uniform step size, then the tradeoff points achievable by varying the number of diagonals used in the index assignment matrix is so limited, that, in the tradeoff range of typical interest, it provides no other choice than the index assignment which maximally favors the side distortion [9]. Achieving intermediate operating points between these discrete points is desirable.
- If more tradeoff points at various rates than are achievable with uniform step size are desired, codebooks must be stored at the encoder/decoder. In designing these optimized codebooks, extensive training is required and the operating points on the convex hull must be identified while eliminating the operating points which are only locally optimal for a particular selection of diagonal number and index assignment.

### III. THE STRUCTURE OF TWO-STAGE MDSQ

To resolve the inconsistencies between the asymptotic analysis and the performance in practice, we propose a new structure for the index assignment of MDSQ. Two classes of index assignments were given in [5]: linear and nested. Let  $k$  be the number of diagonals used in the index assignment matrix above the main diagonal. Then the linear index assignment in [5] can be taken as filling the matrix in a zigzag manner along the diagonal, and changing directions every  $2k$  rounds. Fig. 4 provides an example of a portion of the index assignment matrix for  $k = 2$ . The matrix is assumed to be infinite when rate is high, and thus the border problem can be neglected. For a thorough explanation of the linear index assignment, see [5].

This section provides a new structure for the MDSQ index assignment, by first introducing the base index assignment problem, and then by demonstrating that MDSQ can be performed by a two-stage quantization procedure.

#### A. The Base Index Assignment

When the number of diagonals is increased, the side distortion increases in exchange for a decrease in the central distortion. As the requirement on the side distortion tightens, the number of diagonals decreases, and finally degenerates into two identical descriptions, where only the main diagonal is used in the index assignment matrix. In this case, the *base index assignment* is the main-diagonal-only index assignment (Fig. 5(a)), and at this point the requirement on the side distortion tightens to its extreme. This seems to be a natural choice, but in fact this is not optimal in a rate-distortion sense for MDSQ.

Consider the *staggered index assignment* illustrated in Fig. 5(b) (see [11], [9]) for ECMDSQ, and compare the main-diagonal-only and staggered index assignments at high rate. In particular, consider the case when the central quantization step size of the staggered index assignment is half that of the step

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

Fig. 4. Linear index assignment at  $k = 2$ .

...							
	1						
		2					
			3				
				4			
					5		
						6	
							7
							8
							9
							10
							11
							...
							...

(a)

(b)

Fig. 5. Base index assignment: (a) main-diagonal-only and (b) staggered.

size used with the main-diagonal-only index assignment. In this case, the quantizer rates of the two index assignments are equal, and the side distortions are also nearly equal (neglecting the border area of the matrix at high rates), but the central distortion using the staggered index assignment is reduced by a factor of four when compared with that using the main-diagonal-only assignment. By using the staggered index assignment and a uniform central quantizer, the side quantizers are in fact optimal, because the optimal entropy-constrained quantizer is uniform at high rate [12]. Thus, no further improvement can be made to reduce the side distortion, even when the number of diagonals is decreased to one. The staggered index assignment is indeed a better choice for the base index assignment than the main-diagonal-only matrix.

To generalize this staggered index assignment, the index assignment should always include these two diagonals (*the twin diagonals*); if more diagonals are needed, they can be added symmetrically above and below the twin diagonals to maintain balanced descriptions. By doing this, transition of tradeoffs is achieved by symmetrically increasing or decreasing the number of diagonals, and the degradation to the main-diagonal-only index assignment is avoided. It will become clear in later sections that this change of base index assignment not only facilitates an implementation of MDSQ to achieve good performance, but also provides a two-stage structure which leads to a more elegant analysis. Despite these advantages, the staggered index assignment by itself is neither sufficient nor necessary to provide the 0.4-dB performance again asymptotically, which will be explained in more details in Section V.

The index assignment is thus modified as follows. Let  $k$  be the number of diagonals used in the index assignment above the twin diagonals on each column. The index assignment proceeds in one direction (northeast or southwest) for  $2k + 1$  rounds and

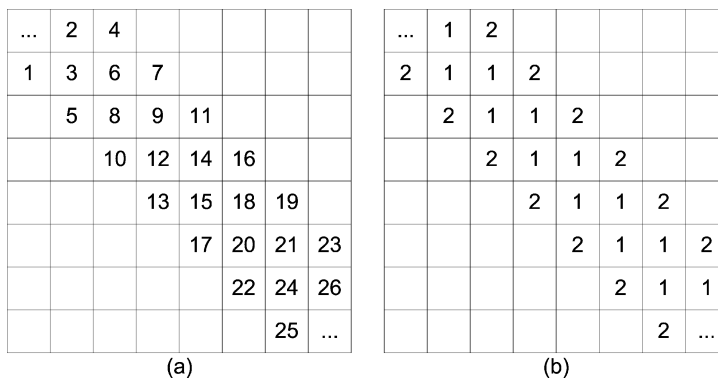


Fig. 6. (a) General index assignment using staggered base index assignment for  $k = 1$ . (b) Labeling the cells according to their distances from the twin diagonals.

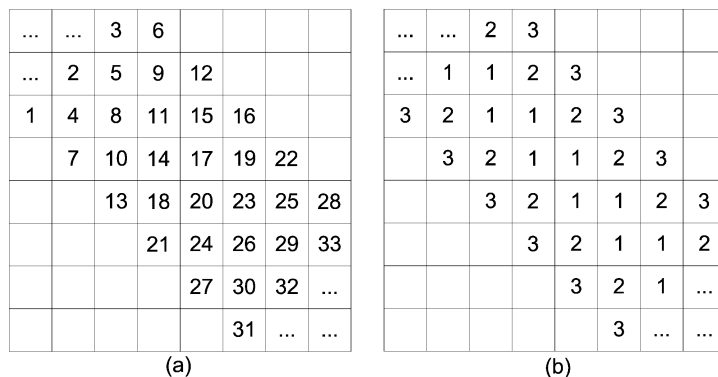


Fig. 7. (a) General index assignment using staggered base index assignment for  $k = 2$ . (b) Labeling the cells according to their distances from the twin diagonals.

TABLE I  
INDICES AND LABELS BY DIAGONAL POSITION FOR  $k = 1$  IN FIG. 6. MACROCELLS ARE DELINEATED WITH SOLID LINES, WHILE INNER CELLS ARE DELINEATED WITH DOTTED LINES

Central Quantizer Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...
Relabel of cells	2	1	1	2	2	1	2	1	1	2	2	1	2	1	1	2	...

TABLE II  
INDICES AND LABELS BY DIAGONAL POSITION FOR  $k = 2$  IN FIG. 7

Central Quantizer Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	...
Relabel of cells	3	1	2	2	1	3	3	1	2	2	1	3	3	1	2	3	1	2	...

then changes directions. As examples, Figs. 6(a) and 7(a) illustrate the index assignments for  $k = 1$  and  $k = 2$ , respectively.

**B. Two-Stage MDSQ**

It is observed in [13] that when the constraint on the side distortion tightens, the central quantizer cells farther away from the main diagonal (which are also the farther cells from their reconstruction values in the side quantizers), become smaller. This shrinking decreases the side distortion since these farther cells generate larger side distortion. By shrinking their lengths, the cells closer to the main diagonal have to expand to compensate for it, which increases the central distortion. Shrinking the farther cells is therefore a technique to decrease the “actual” number of diagonals used in the index assignment matrix.

A classification of the central quantizer cells will be made based on the distance of the cells from the twin diagonals. In the above example with  $k = 1$ , if the cells on the twin diagonals are labeled 1 and the cells one step from the twin diagonals are labeled 2, Fig. 6(b) results, and the corresponding indices

are shown in Table I. (For the example, with  $k = 2$ , they are given in Fig. 7(b) and Table II.) Now consider every  $(2, 1)$  or  $(1, 2)$  to be a larger cell, the boundaries of which segment the entire source range into continuous intervals. Such a larger cell is defined as a *macro-cell*, which is the union of central quantizer cells on a single antidiagonal; an *inner-cell* is then defined as one of the central quantizer cells within each macro-cell. Within each macro-cell, the inner-cells are indexed  $1, 2, 3, \dots$ , in an increasing order of their distances to the twin diagonals in the index assignment matrix.

The MDSQ can therefore be modeled in two steps: the first step is to identify which macro-cell  $i$  contains the sample, and the second step is to identify which inner-cell  $j$  contains the sample. An index assignment mapping

$$a' : (i, j) \rightarrow (p, q), (i, j) \in I^2, (p, q) \in I^2$$

is still required, but because of the periodicity in the index assignment matrix, this mapping is trivial after performing the

above two steps. Note this mapping  $a'$  is different from the mapping  $a$  in Section II-A.

#### IV. ASYMPTOTIC ANALYSIS OF TWO-STAGE MDSQ

The above macro-cell/inner-cell structure is used to develop a straightforward asymptotic analysis of the MDSQ. This method is used to optimize the MDSQ, and to suggest strategies for the design of UMDSQ.

The basic idea in this analysis is that using a continuous function to approximate the distribution of the inner-cells is a poor approximation when the number of diagonals is low; this encompasses many cases of practical interest. Here, such an approximation is not made; however, were this to be done, the asymptotic results (in  $k$ ) in the next section would still be obtained, but the generality would be lost when the number of diagonals were low.

The well-accepted conjecture made by Gersho [14] states that at high rate the cells of the optimal vector quantizer for a vector uniformly distributed on a convex set are all congruent to a certain (optimal) polytope. In MDSQ, a similar situation occurs, only instead of the shape of each vector quantizer cell, we have the inner-cell structure of each macro-cell. In order to make the derivation traceable, we make a similar conjecture that at high rate for a source uniformly distributed in  $[0, 1]$ , all the macro-cells of the optimal MDSQ have identical inner-cell structure. Note that even if this conjecture is not true, our derivation still suffices to provide an upper bound on the distortions, since then MDSQs with such a property are just a special class in the general MDSQs.

Let  $s_j \in [0, 1]$  be the *normalized length factor* for inner-cells of class  $j$ , which represents the ratio between the length of inner-cell  $j$  and that of its macro-cell, and thus  $\sum_{j=1}^{k+1} s_j = 1$ . Denote the vector  $(s_1, s_2, \dots, s_{k+1})$  as  $\mathbf{s}^{k+1}$ . Let  $N$  denote the total number of macro-cells in an MDSQ. For a sequence of MDSQs with strictly increasing number of macro-cells, quantizer  $Q_N$  has a total of  $N$  macro-cells, and the  $k_N + 1$  inner-cells in each macro-cell has inner-cell structure following  $\mathbf{s}_N^{k_N+1}$ . We will denote the parameters of  $Q_N$  as  $(N, k_N, \mathbf{s}_N^{k_N+1})$ , or without ambiguity simply as  $(k, \mathbf{s})_N$ . The following theorem can be proved even without assuming the aforementioned conjecture.

*Theorem 1:* For any sequence of parameters  $(k, \mathbf{s})_N$  such that  $\lim_{N \rightarrow \infty} \frac{k}{N} = 0$ , there exists a sequence of MDSQs with such parameters on a source uniformly distributed on  $[0, 1]$ , and a sequence of vectors  $\mathbf{w}^{k+1} = (w_1, w_2, \dots, w_{k+1})$ , where  $j - 1 < w_j < j$ , such that

$$D_{Q0} = \frac{1}{12N^2} \sum_{j=1}^{k+1} s_j^3 \quad (8)$$

$$\lim_{N \rightarrow \infty} \frac{D_{Q1}}{\frac{1}{N^2} \sum_{j=1}^{k+1} w_j^2 s_j} = 1. \quad (9)$$

Theorem 1 essentially states that for high-resolution MSDQs, when the number of diagonals is relatively small compared with the number of macro-cells, the performances in (8) and (9) are achievable. In fact,  $w_j^2$  has its physical meaning as the expected squared distance between the points falling in inner-cells of

class  $j$  and their side quantizer reconstruction points, normalized by the squared length of a macro-cell. More details on  $s_j$  and  $w_j$  will be given as we proceed. Note that different tradeoff points can still be achieved by using different values of  $s_j$ 's for a given  $k$ , in contrast with the expressions given in [6] (see Section II-B (4) and (5), which give the central and side distortions for ECMDSQs with a uniform central quantizer at high rates), where only one optimal tradeoff point is achievable for a given  $k$ .

The purpose of presenting this theorem is to rigorously provide certain justifications for the approximations that will be made during the derivation (at least for a simple source), and the proof of this theorem is delayed to the Appendix. In the sequel, we start the derivation with more general sources using a heuristic approach by introducing a point density function on the macro-cells.

##### A. Central Distortion Integral

Using the concept of "asymptotic fractional density of quanta" by Lloyd [15], define the *macro-cell density function* as

$$g_N(x) = \frac{1}{NL(C_i)}, \quad x \in C_i, i = 1, 2, \dots, N \quad (10)$$

where  $L(C_i)$  denotes the length of the macro-cell  $C_i$  and  $N$  is the total number of macro-cells.

When the rate is high,  $N$  is very large, and  $g_N(x)$  can be approximated closely by a continuous density function  $g(x)$ . Then  $g(x)\Delta L(x)$  may be taken as the fraction of macro-cells located in an incremental length element  $\Delta L(x)$ , which contains  $x$ . The length of the macro-cell  $C_i$  into which the sample  $x$  falls is given approximately by

$$L(C_i) \approx \frac{1}{Ng(x)}. \quad (11)$$

Assume that the overload region is properly chosen such that the overload distortion can be ignored, and thus only distortions generated on the finite support of  $[-x_0, x_0]$  need to be considered (see [12] for a complete explanation on this topic). Consider the central distortion

$$\begin{aligned} D_{Q0} &= \int_{-x_0}^{x_0} (x - q(x))^2 p(x) dx \\ &= \sum_{i=1}^N \sum_{j=1}^{k+1} \int_{L_{ij}}^{U_{ij}} (x - y_{ij})^2 p(x) dx \end{aligned} \quad (12)$$

where  $q(x)$  is the quantized value of  $x$  by the central quantizer,  $y_{ij}$  is the  $j$ th reconstruction value (for the  $j$ th inner-cell) in the macro-cell  $i$ ,  $L_{ij}$  is the lower threshold of the inner-cell  $j$  in macro-cell  $i$ ,  $U_{ij}$  is the upper threshold of the inner-cell  $j$  in macro-cell  $i$ , and there are  $k + 1$  inner-cells in a macro-cell.

Since at high rate

$$p(x) \approx p(y_{ij}) \approx p(y_i) \quad (13)$$

where  $y_i$  is taken as the midpoint of macro-cell  $i$ , then

$$D_{Q0} \approx \sum_{i=1}^N p(y_i) \sum_{j=1}^{k+1} \int_{L_{ij}}^{U_{ij}} (x - y_{ij})^2 dx. \quad (14)$$

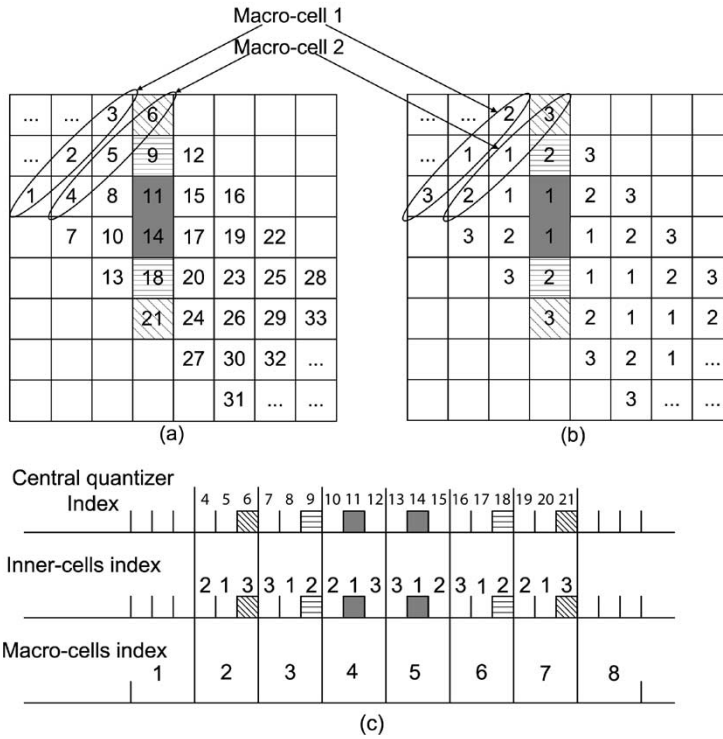


Fig. 8. (a) The index assignment matrix for  $k = 2$ . (b) Labeling with the inner-cell indices. (c) The quantizer cells on the real line. The union of the shaded cells in the quantizer are the side quantizer cell in shading in the index assignment matrix, while the different shadings of the shaded areas stand for different classes of inner-cells.

Let  $s_{ij}$  denote the length of the  $j$ th inner-cell of macro-cell  $i$ . Assuming midpoint reconstruction, the distortion in an inner-cell is approximated as

$$\int_{L_{ij}}^{U_{ij}} (x - y_{ij})^2 dx \approx \int_{-s_{ij}/2}^{s_{ij}/2} x^2 dx \approx s_{ij}^3/12 \quad (15)$$

and then

$$D_{Q0} \approx \sum_{i=1}^N p(y_i) \sum_{j=1}^{k+1} \frac{s_{ij}^3}{12}. \quad (16)$$

In the high-resolution case, every macro-cell covers only a very small fraction of  $[-x_0, x_0]$ , thus, we can now invoke the Gershho-type conjecture for MDSQ. Then take

$$s_{ij} = s_j L(C_i) = s_j \frac{1}{Ng(y_i)} \quad (17)$$

where  $s_j$  is the normalized length factor of a inner-cell. Recall that it is the fraction that the inner-cell  $j$  occupies in the macro-cell to which it belongs, and they must satisfy

$$\sum_{j=1}^{k+1} s_j = 1. \quad (18)$$

From (16), the central distortion is given as

$$D_{Q0} \approx \sum_{i=1}^N \frac{1}{12} p(y_i) \frac{1}{(Ng(y_i))^3} \sum_{j=1}^{k+1} s_j^3 \quad (19)$$

which can be approximated as

$$\begin{aligned} D_{Q0} &\approx \sum_{i=1}^N \frac{1}{12} p(y_i) \frac{1}{(Ng(y_i))^2} \frac{1}{Ng(y_i)} \sum_{j=1}^{k+1} s_j^3 \\ &= \sum_{i=1}^N \frac{1}{12} p(y_i) \frac{1}{(Ng(y_i))^2} L(C_i) \sum_{j=1}^{k+1} s_j^3 \end{aligned}$$

$$= \frac{1}{12} \sum_{j=1}^{k+1} s_j^3 \int_{-x_0}^{x_0} \frac{p(y)}{(Ng(y))^2} dy \quad (20)$$

with the fact that

$$L(C_i) \approx \frac{1}{Ng(x)} \approx \delta y. \quad (21)$$

### B. Side Distortion Integral

To obtain an expression for the side distortion, the side quantizer cell structure is exploited. It can be expressed in terms of the macro-cell length and probability based on the classification of the inner-cells, and is given by

$$\begin{aligned} D_{Q1} &= \int_{-x_0}^{x_0} (x - q^1(x))^2 p(x) dx \\ &= \sum_{i=1}^N \sum_{j=1}^{k+1} \int_{L_{ij}}^{U_{ij}} (x - y_{ij}^1)^2 p(x) dx \quad (22) \end{aligned}$$

where  $q^1(x)$  is the quantized value of  $x$  by the side quantizer 1; and  $y_{ij}^1$  is the reconstruction value for side quantizer 1 for inner-cell  $j$  of the macro-cell  $i$ .

An example can illustrate the structure of the cells specified by an index assignment matrix, and the example with  $k = 2$  from the previous section is redrawn in Fig. 8 for convenience. Notice that each column has  $2(k + 1)$  inner-cells, and these inner-cells belong to  $2(k + 1)$  macro-cells, which approximately makes the center of the union of these inner-cells (in the same column) the center of two innermost of these  $2(k + 1)$  macro-cells. Observe that in Fig. 8, the side quantizer cell in shading is consisted of central quantizer cells (6, 9, 11, 14, 18, 21), and these cells are contained in macro-cells (2, 3, 4, 5, 6, 7), respectively. The reconstruction value for this column is approximated

at the boundary between macro-cell 4 and 5. This approximation is less accurate for small  $k$ , but improves as  $k$  increases. Nevertheless, this is a conservative approximation, which suffices to provide an upper bound on the granular distortion.

The square root of the expected squared distance of points falling in inner-cells of class  $j$  to their side quantizer reconstruction points can be represented by

$$w_j \times \frac{1}{Ng(y)}, \quad j = 1, 2, \dots, k+1 \quad (23)$$

and  $w_j$  denotes the same value normalized by the macro-cell length, and notice that  $j-1 < w_j < j$ . An implicit assumption is made that the macro-cell lengths are approximately uniform over a local area with  $2(k+1)$  macro-cells, which requires that  $\frac{2(k+1)}{N} \rightarrow 0$  as  $R \rightarrow \infty$ . In other words, the index assignment matrix is a "thin"-banded matrix.<sup>2</sup>

Then it can be shown that the side distortion is approximated as

$$\begin{aligned} D_{Q1} &\approx \sum_{i=1}^N p(y_i) \sum_{j=1}^{k+1} \int_{L_{ij}}^{U_{ij}} (x - y_{ij}^1)^2 dx \\ &\approx \sum_{i=1}^N p(y_i) \sum_{j=1}^{k+1} (w_j \frac{1}{Ng(y_i)})^2 L(C_i) s_j \\ &\approx \sum_{j=1}^{k+1} w_j^2 s_j \int_{-x_0}^{x_0} \frac{p(y)}{(Ng(y))^2} dy. \end{aligned} \quad (24)$$

Equations (20) and (24) share a common factor

$$\int_{-x_0}^{x_0} \frac{p(y)}{(Ng(y))^2} dy$$

which depends only on the macro-cell density function, the source pdf, and the total number of macro-cells. Thus, to simultaneously minimize  $D_{Q0}$  and  $D_{Q1}$  while allowing a tradeoff between these two, the common factor should be minimized first, and then the normalized length factors  $s_j$  and the number of diagonals  $k$  should be properly selected to achieve different tradeoffs.

### C. Rate of the Quantizer

1) *Rate of ECMDSQ*: For the entropy-constrained case, the rate is the entropy of the side quantizer indices. To compute this, first note that the sum of the lengths (and probabilities) of the inner-cells in every column (row) of the index assignment are approximately equal to the sum of the lengths (and probabilities) of two macro-cells, because each column/row has two sets of inner-cells, and each set consists of inner-cells with inner-cell indices to make a macro-cell. Thus,<sup>3</sup>

$$p_i \approx \frac{p(y_{2i})}{Ng(y_{2i})} + \frac{p(y_{2i+1})}{Ng(y_{2i+1})} \approx \frac{2p(y_{2i})}{Ng(y_{2i})} \quad (25)$$

<sup>2</sup>This is true if  $k+1 = 2^{\eta R}$ ,  $0 < \eta < 1$ , when  $R \rightarrow \infty$ . *Proof*: suppose it is false, then  $N$  has to be at most  $\Theta(k+1)$ , which implies the size of the matrix has to be at most  $\Theta(k+1)$ . Thus, the rate of the quantizer is at most  $\Theta(\eta R)$ , regardless of whether it is an entropy-constrained or a level-constrained quantizer, and this contradicts the fact that the rate of the quantizer is  $R$ .  $\square$

<sup>3</sup>Note the labeling here is not consistent with the example in Fig. 8, because of the border effect in the index assignment matrix. However, under the high-rate condition, the border effect can be neglected and this approximation holds.

where  $p_i$  is the probability that the sample falls in the column with side quantizer index  $i$ . The rate integral is given by

$$\begin{aligned} H_Q &= - \sum_i p_i \log_2(p_i) \\ &\approx - \sum_i \frac{2p(y_{2i})}{Ng(y_{2i})} \log_2 \frac{2p(y_{2i})}{Ng(y_{2i})} \\ &\approx - \sum_i \frac{p(y_{2i})}{Ng(y_{2i})} \log_2 \frac{2p(y_{2i})}{Ng(y_{2i})} \\ &\quad - \sum_i \frac{p(y_{2i+1})}{Ng(y_{2i+1})} \log_2 \frac{2p(y_{2i+1})}{Ng(y_{2i+1})} \\ &= - \sum_i \frac{p(y_i)}{Ng(y_i)} \log_2 \frac{2p(y_i)}{Ng(y_i)} \\ &\approx - \int_{-x_0}^{x_0} p(y) \log_2 p(y) dy \\ &\quad - 1 + \sum_i p(y_i) \log_2(Ng(y_i)) \delta y \\ &\approx h(p) - 1 + \int_{-x_0}^{x_0} p(y) \log_2(Ng(y)) dy \end{aligned} \quad (26)$$

where  $h(p)$  is the differential entropy of the random variable  $x$ .

In this case, applying Jensen's inequality to (26) gives

$$\int_{-x_0}^{x_0} \frac{p(y)}{(Ng(y))^2} dy \geq 2^{-2(H_Q - h(p) + 1)}. \quad (27)$$

The optimal  $g(y)$  which makes (27) hold with equality is a constant, corresponding to uniformly distributed macro-cells. This suggests that for ECMDSQ, in the first stage of an optimal two-stage implementation of quantization, a uniform step size can be used for the macro-cells regardless of the structure within the macro-cell. And also the common factor in the central and side distortion integrals ((20) and (24)) only depends on the rates, but not on the tradeoff.

Consider now the special case when the inner-cells have equal length; i.e.,  $s_j = 1/(k+1)$ . For a Gaussian source with unit variance, and when  $k$  is large, taking either  $w_j = j$  (or  $w_j = j-1$ ) leads to the distortion product

$$D_{Q0} D_{Q1} \approx \frac{\pi^2 e^2}{144} 2^{-4H_Q} \quad (28)$$

which agrees with the result in [6] for ECMDSQ with a uniform central quantizer.

2) *Rate of Level-Constrained MDSQ*: For the level-constrained case, the total number of side quantizer indices

$$\lfloor N/2 \rfloor + k + 1 = 2^R$$

is given. As previously mentioned,  $\frac{2(k+1)}{N} \rightarrow 0$  as  $R \rightarrow \infty$ , and thus,  $N \rightarrow 2^{R+1}$  as  $R \rightarrow \infty$ . Thus, the constraint of quantizer level is actually on  $N$  in this case. Then the optimal macro-cell point density function is given in [16]

$$g(x) = \frac{p^{1/3}(x)}{\int_{-x_0}^{x_0} p^{1/3}(x) dx}. \quad (29)$$

This suggests that in the level-constrained case, the common factor

$$\int_{-x_0}^{x_0} \frac{p(y)}{(Ng(y))^2} dy$$

is also only controlled by the rate by varying  $N$  at high resolution.

Thus, in both entropy-constrained and level-constrained cases, the two-stage quantization reduces the optimization problem to a decoupled two-step procedure: optimize the product of the macro-cell density function and  $N$  subject to the rate constraint; and optimize the normalized inner-cell length factors to achieve different tradeoffs.

#### D. The Optimal Normalized Inner-Cell Length

To achieve an optimal tradeoff, the normalized inner-cell lengths must vary according to the distortion and rate constraints. Redefine the problem as follows:

$$\text{minimize } D_{Q0}$$

subject to constraints on both the side distortion and the rate

$$D_{Q1} \leq D_s$$

$$H_Q \leq R$$

where  $D_s$  is the constraint on the side distortion. This formulation and the following analysis are for ECMDSQs only, but a similar approach also applies to level-constrained MDSQs. Also notice that this problem can be formulated in a Lagrangian approach [17], which is to minimize  $D_{Q0} + \lambda_1 D_{Q1} + \lambda_2 H_Q$  for this specific problem.

Using the equations in Sections IV-B–IV-D, proper selection of  $Ng(y)$  satisfies the second constraint. The problem, therefore, reduces to one involving the normalized length factors for the inner-cells

$$\text{minimize } \sum_{j=1}^{k+1} s_j^3$$

subject to the constraints

$$\sum_{j=1}^{k+1} w_j^2 s_j \leq C_s$$

$$\sum_{j=1}^{k+1} s_j = 1$$

$$s_j \geq 0, \quad j = 1, 2, \dots, k+1$$

where the constant  $C_s$  is defined as

$$C_s = \frac{D_s}{\int_{-x_0}^{x_0} \frac{p(y)}{(Ng(y))^2} dy} \quad (30)$$

and is completely determined by the constraint  $D_s$  and the function  $Ng(y)$ , which are known.

Application of the Kuhn–Tucker conditions to this problem yields

$$s_j^2 = \begin{cases} (\lambda - \lambda_1 w_j^2), & \lambda \geq \lambda_1 w_j^2 \\ 0, & \text{else} \end{cases} \quad (31)$$

and

$$\sum_{j=1}^{k+1} w_j^2 s_j \leq C_s \quad (32)$$

$$\sum_{j=1}^{k+1} s_j = 1 \quad (33)$$

$$\lambda \geq 0 \quad (34)$$

$$\lambda_1 \geq 0 \quad (35)$$

$$\lambda_1 \left( \sum_{j=1}^{k+1} w_j^2 s_j - C_s \right) = 0. \quad (36)$$

Consider these conditions under the following cases.

*High Side-Distortion Case:* If  $C_s$  is large, the constraint (32) can be satisfied with inequality and thus,  $\lambda_1 = 0$ , which gives  $s_j = \sqrt{\lambda} = 1/(k+1)$ . A uniform inner-cell length factor is therefore optimal in the high side-distortion case; i.e., when the side-distortion constraint is so loose as to be satisfied with inequality, uniform inner-cell length is optimal.

*Low Side-Distortion Case:* As the constraint on the side distortion tightens, (32) must be satisfied with equality. Thus, some  $s_j$  corresponding to large values of  $w_j$  go to zero, and without loss of generality, choose  $\lambda = \beta \lambda_1$ , with  $\beta$  chosen so  $w_m^2 < \beta < w_{m+1}^2$ ,  $1 \leq m \leq k$ , yielding

$$s_j^2 = \begin{cases} \lambda_1(\beta - w_j^2), & j \leq m \\ 0, & j > m. \end{cases} \quad (37)$$

(Recall that  $j-1 < w_j < j$ .)

This result suggests that the farthest cells from diagonal are eliminated first, and the remaining nonzero cells are sized accordingly. Uniform inner-cells will not provide optimal MDSQ performance, unless the constraint on the side distortion is so trivial to satisfy that it has no influence on the optimization problem. As the value of  $\beta$  changes continuously, the normalized inner-cell length factors vary accordingly, and a continuum of tradeoffs is subsequently achieved.

Similar behaviors were observed by Goyal *et al.* [18] in the case of multiple description vector quantizer (MDVQ) with a lattice codebook. Though their observation was reached by applying a generalized Lloyd–Max algorithm without explicit analysis such as presented here, the diminishing of the “farthest” cells when the side distortion constraint tightens and the nonuniformity of central quantizer cells are very similar. The topic of MDVQ is beyond the scope of this paper; however, the analysis presented here can be extended to MDVQ with minor modifications [19].

#### V. GRANULAR DISTORTION REVISITED

The granular distortion of MDSQ was calculated in [6] using a compander function on the central quantizer. In this section, the granular distortion problem is re-examined using our analysis. The granular distortions for ECMDSQ and level-constrained MDSQ are derived, respectively, and compared with the rate-distortion bounds.

Note that in the proposed two-stage structure, macro-cells have the same number of inner-cells, and subsequently are assumed to have the same inner-cell structure. This is different from the previous structure of MDSQ by Vaishampayan [5], [7], which interleaves two types of macro-cell structures, when the main-diagonal-only matrix is used as the base index assignment matrix. These two types of macro-cells have different numbers of inner-cells, thus, they cannot be assumed to have the same inner-cell structure. However, with some modification, the asymptotic analysis method presented here can be used on the MDSQs designed with a main-diagonal-only matrix as the base index assignment matrix, which leads to similar expressions for  $D_0$ ,  $D_1$ , and  $H_Q$ , and the same granular gap results can also be derived

through a less straightforward route. The choice of the staggered matrix as the base index assignment matrix against the main-diagonal-only matrix is more from a design perspective, rather than to reach different asymptotic results. An even more accurate approximation for the side distortion can be derived by using refined classification of “types” as presented in [20], which is beyond the scope of this paper. However, the derivation presented here is reached by a conservative approximation, which suffices to provide an upper bound on the granular distortions of MDSQ.

#### A. ECMDSQ

To achieve different tradeoffs by using different  $k$  value, let  $k + 1 = 2^{\eta H_Q}$ , which is also used in [6] to derive the granular distortions. Using the expressions from the previous section, the optimal entropy-constrained quantizer can achieve

$$D_{Q0} = \frac{1}{48} \sum_{j=1}^{k+1} s_j^3 2^{-2H_Q + 2h(p)}$$

$$D_{Q1} = \frac{1}{4} \sum_{j=1}^{k+1} w_j^2 s_j 2^{-2H_Q + 2h(p)}.$$

Thus upper bounds are needed for the quantities

$$B_0 = \sum_{j=1}^{k+1} s_j^3 \quad \text{and} \quad B_1 = \sum_{j=1}^{k+1} w_j^2 s_j. \quad (38)$$

To find these upper bounds, we show that some sets of valid  $s_j$  can achieve satisfactory distortions.

*Definition 1:* A set of  $s_j$  is said to be admissible for  $k + 1$  if  $\sum_{i=1}^{k+1} s_i = 1$  and  $s_j \geq 0$  for  $j = 1, 2, \dots, k + 1$ .

For any admissible  $s_j$

$$\sum_{j=1}^{k+1} w_j^2 s_j \leq \sum_{j=1}^{k+1} j^2 s_j \quad (39)$$

which is trivial because  $w_j < j$ . Thus, if the quantity

$$B'_1 = \sum_{j=1}^{k+1} j^2 s_j \quad (40)$$

with a set of admissible  $s_j$  is upper-bounded, then this bound is an upper bound for  $B_1$ .

Consider the following admissible  $s_j$  for a certain  $k + 1$ , call it  $\hat{s}_j$

$$\hat{s}_j = \frac{(\beta - j^2)^{1/2}}{\sum_{j=1}^{k+1} (\beta - j^2)^{1/2}} \quad (41)$$

where  $\sqrt{\beta} = k + 1 + \epsilon$  with  $1 \geq \epsilon > 0$ .

Then

$$B_0 = \beta^{-1} \sum_{j=1}^{k+1} \left(1 - \frac{j^2}{\beta}\right)^{3/2} \beta^{-1/2}$$

$$\times \left(\sum_{j=1}^{k+1} \left(1 - \frac{j^2}{\beta}\right)^{1/2} \beta^{-1/2}\right)^{-3} \quad (42)$$

$$B'_1 = \beta \sum_{j=1}^{k+1} \frac{j^2}{\beta} \left(1 - \frac{j^2}{\beta}\right)^{1/2} \beta^{-1/2}$$

$$\times \left(\sum_{j=1}^{k+1} \left(1 - \frac{j^2}{\beta}\right)^{1/2} \beta^{-1/2}\right)^{-1}. \quad (43)$$

Let  $z = \frac{j}{\sqrt{\beta}}$ , and since  $1 \leq j \leq k + 1$ , then  $0 < z < 1$ . Recall that  $\sqrt{\beta} \approx k + 1 = 2^{\eta H_Q}$ , and as  $k \rightarrow \infty$ ,  $\frac{1}{\sqrt{\beta}} = \Delta z = dz$ . Thus, the above quantities can be approximated as integrals

$$B_0 \approx 2^{-2\eta H_Q} \int_0^1 (1 - z^2)^{3/2} dz \left(\int_0^1 (1 - z^2)^{1/2} dz\right)^{-3}$$

$$= \frac{12}{\pi^2} 2^{-2\eta H_Q} \quad (44)$$

$$B'_1 \approx 2^{2\eta H_Q} \int_0^1 z^2 (1 - z^2)^{1/2} dz \left(\int_0^1 (1 - z^2)^{1/2} dz\right)^{-1}$$

$$= \frac{1}{4} 2^{2\eta H_Q}. \quad (45)$$

Thus, the optimal central and side distortions are upper-bounded by

$$D_{Q0} \approx \frac{1}{4\pi^2} 2^{-2(1+\eta)H_Q + 2h(p)} \quad (46)$$

$$D_{Q1} \approx \frac{1}{16} 2^{-2(1-\eta)H_Q + 2h(p)}. \quad (47)$$

For ECMDSQ with the mean-squared error measurement, the sets of  $\hat{s}_j$  can achieve arbitrarily close to

$$D_{Q0} D_{Q1} \approx \frac{1}{64} \frac{1}{\pi^2} 2^{-4H_Q + 4h(p)} \quad (48)$$

while the rate distortion bound is ((1) and (2))

$$D_0 D_1 = \frac{1}{4} \frac{1}{(2\pi e)^2} 2^{-4H_Q + 4h(p)}. \quad (49)$$

Thus, an upper bound for the granular gap of the product of central and side distortions for ECMDSQ is  $e^2/4 = 2.67$  dB. The performance difference between ECMDSQ with the optimal inner-cell lengths and ECMDSQ with uniform inner-cell lengths is  $9/\pi^2 = 0.4$  dB, in terms of distortion product.

#### B. Level-Constrained MDSQ

For level-constrained MDSQ, a similar result is obtained by taking  $N = 2^{R+1}$ ,  $k + 1 = 2^{\eta R}$ , and using the optimal  $g(x)$  as given in (29); thus, an upper bound is given as

$$D_{Q0} \approx \frac{1}{4\pi^2} \left(\int_{-x_0}^{x_0} p^{1/3}(x) dx\right)^3 2^{-2(1+\eta)R}$$

$$D_{Q1} \approx \frac{1}{16} \left(\int_{-x_0}^{x_0} p^{1/3}(x) dx\right)^3 2^{-2(1-\eta)R}.$$

Compare these with the results when a uniform central quantizer is used [6], which are

$$D'_{Q0} \approx \frac{1}{48} \left(\int_{-x_0}^{x_0} p^{1/3}(x) dx\right)^3 2^{-2(1+\eta)R}$$

$$D'_{Q1} \approx \frac{1}{12} \left(\int_{-x_0}^{x_0} p^{1/3}(x) dx\right)^3 2^{-2(1-\eta)R}.$$

There is also a  $9/\pi^2 = 0.4$  dB difference in terms of distortion product. In the special case of a Gaussian source with unit variance, optimal inner-cell lengths can achieve

$$D_{Q0}D_{Q1} \approx \frac{27}{16}2^{-4R} \quad (50)$$

and comparing with the rate-distortion result of

$$D_0D_1 = \frac{1}{4}2^{-4R} \quad (51)$$

which demonstrates a 8.29-dB granular gap, while previously this gap was given as 8.69 dB [6].

## VI. DESIGN OF UNIVERSAL MSDQ (UMDSQ)

Since quantizers are in practice typically followed by entropy coding, in this section we apply our analysis results to the design of ECMDSQ and introduce the UMDSQ.

### A. A Greedy Approximation to the Optimal Inner-Cell Lengths

In order to achieve a continuum of tradeoff points while exploiting the simplicity of uniform inner-cell structure, the following technique is considered: for fixed  $k$ , keep the macro-cell length fixed (which approximately fixes the rate), shrink gradually the farthest inner-cell from diagonal, and make the other inner-cells uniform. This approximation method is *greedy*, because it starts with uniform inner-cells and their tradeoffs, then greedily achieves other tradeoffs. As the diagonal number is varied, repeating the above process yields a continuous tradeoff curve.

The performances of the greedy and optimal inner-cell lengths are compared using simulation methods. Fig. 9 illustrates the results when  $R = 4$  bps/channel. The difference between the greedy and optimal methods is almost undistinguishable, which is not surprising, since asymptotically this difference is only 0.4 dB. The optimal method gives smoother transitions between tradeoffs, while the greedy method produces some small bumps on the curve. The optimal method therefore only provides a minor improvement over the greedy method, and the latter is applied to the design of universal MDSQ in the next section.

### B. Implementation of UMDSQ

Based on the two-stage MDSQ structure and the greedy inner-cell structure, UMDSQ is proposed in this section as follows.

- 1) Quantize the sample using the macro-cell step size  $\Delta_1$ , to get the macro-cell index  $i$ . Suppose sample  $x \in [x_i, x_i + \Delta_1)$ , where  $x_i$  and  $x_i + \Delta_1$  are macro-cell thresholds, then compute the residual error as  $e_1 = x - x_i$ , which will give  $0 \leq e_1 < \Delta_1$ .
- 2) Quantize the residual error  $e_1$  with one of two classes of the inner-cell quantizers, which are uniform except for the farthest cell from the twin diagonals, to get the inner-cell index  $j$ .
- 3) Compute the index assignment based on the above two steps.

Details of the three steps are now described.

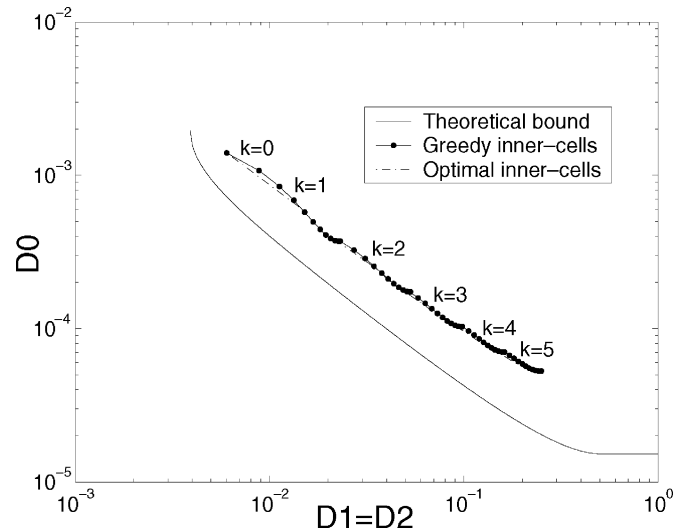


Fig. 9. Comparison of optimal versus greedy methods at  $R = 4$  bps/channel.

TABLE III  
TWO INNER-CELL QUANTIZERS

class	Normalized inner-cell length factor
1	$\underbrace{\frac{1}{k+1+\alpha}, \frac{1}{k+1+\alpha}, \dots, \frac{1}{k+1+\alpha}}_{k+1}, \frac{\alpha}{k+1+\alpha}$
2	$\frac{\alpha}{k+1+\alpha}, \underbrace{\frac{1}{k+1+\alpha}, \frac{1}{k+1+\alpha}, \dots, \frac{1}{k+1+\alpha}}_{k+1}$

The inner-cell quantization can be taken as quantizing a sample on a finite support. Define an index assignment to have *redundancy parameter*  $k + \alpha$ , such that the macro-cell is divided into  $k + 2$  inner-cells, and the cell length ratios are  $1 : 1 : \dots : 1 : \alpha$ , i.e., the  $k + 1$  inner-cells closer to the twin diagonals have same cell lengths, except that the inner-cell farthest from the twin diagonals has normalized inner-cell length factor  $\alpha/(k + 1 + \alpha)$ . Note that only  $0 \leq \alpha \leq 1$  is allowed, and  $k$  is a positive integer. If  $\alpha = 0$ , then the  $k + 1$  inner-cells will have uniform threshold, i.e., the macro-cell is divided uniformly into  $k + 1$  cells; and if  $\alpha = 1$ , then the  $k + 2$  inner-cells have uniform step size. Now there are only two parameters to control this whole system: the redundancy parameter  $k + \alpha$  and the macro-cell length  $\Delta_1$ .

Because of the periodically changing direction of index assignment matrix, in the second stage quantization, there are two classes of quantizers as in Table III. Class 2 quantizers are used in macro-cells in which the inner-cell with the smallest central quantizer index is farthest from the twin diagonals; otherwise, class 1 quantizers are used. These two quantizers have nearly identical structures (they differ only in that one is the reverse of the other) and almost uniform step sizes, and therefore they are easy to implement.

To achieve perfectly balanced descriptions, the fact that the sources in practical application usually have symmetric distribution about zero is used. An example index assignment matrix for  $k = 1$ , which takes this fact into account, is given in Fig. 10. Notice that there are negative indices in this matrix, and special care should be taken to retain the symmetry about zero. Details for this technique can be found in [5].

	-4	-3	-2	-1	0	1	2	3	4
-4	...	-14	-12						
-3	-13	-11	-9	-7					
-2		-10	-8	-5	-4				
-1			-6	-3	-2				
0				-1	0	2	4		
1					1	3	5	7	
2						6	8	9	12
3							10	11	14
4								13	...

Fig. 10. Index assignment using the symmetric property of the source. The numbers outside the matrix are the column/row indices.

Reconstructing using the cell centers rather than centroids dramatically deteriorates the performance of UMDSQ at low rates, especially the side distortion. To maintain universality, on-the-fly training can be employed for several reconstruction values around zero [21]. The encoder uses the uniform two-stage quantizer to encode and simultaneously, the reconstruction values of quantizers are trained using the data being encoded. The decoder uses the trained reconstruction values to reconstruct the data, instead of using center reconstruction. Training the side quantizer reconstruction values with indices  $(-3, -2, -1, 1, 2, 3)$  (see Fig. 10) and the central reconstruction values with indices  $(-2, -1, 1, 2)$  achieves performance with negligible differences from centroid reconstruction (0.3–0.6 dB, compared with the granular gap of 2.67 dB). Notice that the trained reconstruction values need to be transmitted, which results in a higher rate. However, these values need to be transmitted only once after a block of data are transmitted, at the expense of delay, and the overall rate increase over the whole block is usually minor.

The design of ECMDSQ presented in [7] requires extensive training on the quantizer thresholds and reconstruction values for a given index assignment matrix, with a generalized Lloyd–Max algorithm. Furthermore, running this algorithm to convergence only once might not be enough, because the operation points thus reached are only locally optimal, and they have to be identified whether they are indeed on the convex hull of the rate-distortion region. If they are not on the convex hull, which itself is difficult to determine, multiple runs with different index assignment matrices are required. For practical applications, the pdfs of the sources are usually unknown, which makes such training either impossible or undesirable. In contrast, UMDSQ does not require training, while it still achieves a similar performance to ECMDSQ with optimized codebooks at high rates. And it is for this reason that UMDSQ is universal to various sources with smooth pdfs. In a sense, UMDSQ is similar to the uniform quantizer followed by entropy coder in classical one-description system, which is widely used in practical applications and is (close to) optimal at reasonably high rate.

### C. Simulation Results

The performance of UMDSQ is evaluated and compared with that of ECMDSQ with optimized codebooks for a memory-

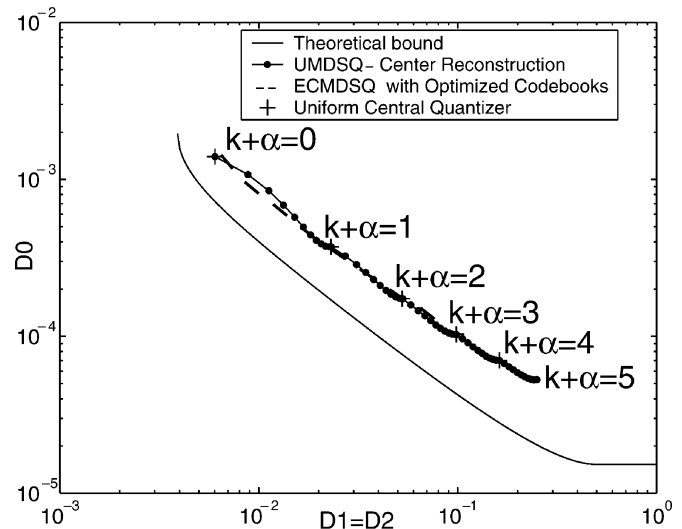


Fig. 11. Comparison of UMDSQ (with center reconstruction) versus ECMDSQ with optimized codebooks at  $R = 4$  bps/channel. The “Uniform Central Quantizer” marks the discrete tradeoff points that can be achieved by using uniform inner-cells, while UMDSQ connects them nicely.

less Gaussian source with unit variance. The granular region is taken large enough so that the rate and distortion contribution in the overload region can be ignored within machine precision. The parameter  $k + \alpha$  is increased in increments of 0.1 at both 4 bps/channel in Fig. 11 and 1 bps/channel in Fig. 12. At 4 bps/channel, as  $k + \alpha$  is varied from 0 to 5, UMDSQ with center reconstruction can achieve nearly the same performance as ECMDSQ with optimized codebooks. At 1 bps/channel, the high-rate assumption does not hold. UMDSQ with centroid reconstruction is shown for comparison purposes and achieves comparable performance to ECMDSQ with optimized codebooks, while the performance with center reconstruction deteriorates dramatically. On-the-fly training provides a compromise between them. Note that while ECMDSQ with an optimized codebook favors the central distortion, UMDSQ favors the side distortion. UMDSQ implemented using a two-stage quantizer followed by an entropy coder requires selection of only two parameters, and provides a continuum of tradeoff as the two parameters are varied.

## VII. CONCLUSION

We provide a new and straightforward asymptotic analysis of a class of MDSQs. This analysis is more general and more insightful than previous analyses. We show that the uniform central quantizer is not optimal in general, and by using optimal inner-cell lengths, the granular gap of ECMDSQ can be reduced to 2.67 dB, instead of 3.07 dB. Similarly, for level-constrained MDSQ, the same 0.4-dB improvement can be achieved. These results in effect provide new upper bounds on the granular distortions of MDSQs. The tightness of these bounds relies on a generalized Gersho-type conjecture for MDSQs that at high rates all the macro-cells are identical on a source with uniform pdf, and thus it is possible that these bounds can be improved. Based on the result given by the analysis of the two-stage quantization, UMDSQ is proposed, which can achieve a continuum of

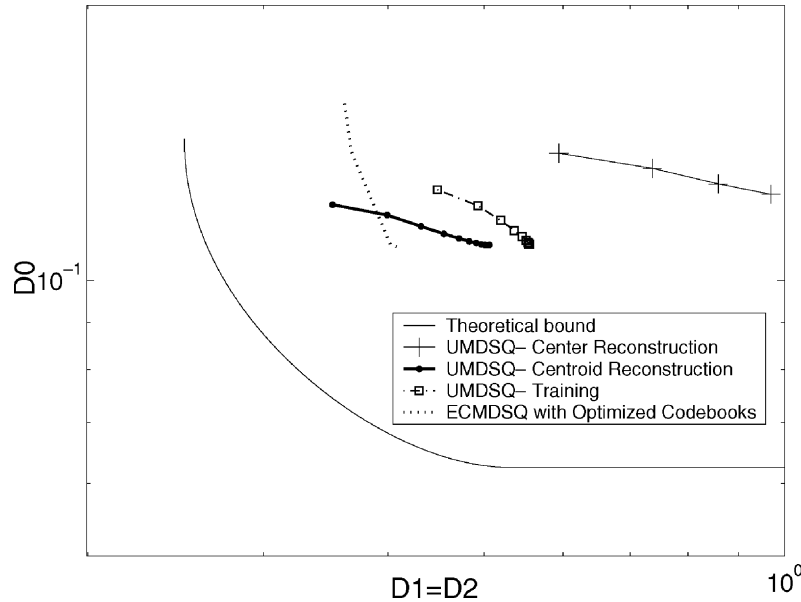


Fig. 12. Comparison of UMDSQ versus ECMDSQ with optimized codebooks at  $R = 1$  bps/channel.

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

Fig. 13. Top-left portion of the index assignment matrix.

tradeoffs between central and side distortions more gracefully, without using the extensive training in the design, as required by ECMDSQ with optimized codebooks.

#### APPENDIX

In this appendix we prove Theorem 1 by showing that a particular construction of two-stage MDSQ can achieve the claimed performances.

##### A. Proof

Consider an MDSQ partitioning the interval  $[0, 1]$  uniformly into a total of  $N$  macro-cells, and all the macro-cells have the identical inner-cell structure. It is trivial to show that when center reconstruction is used

$$D_{Q0} = \int_0^1 (x - q(x))^2 dx = \frac{1}{12N^2} \sum_{j=1}^{k+1} s_j^3 \quad (\text{A52})$$

where  $q(x)$  is the quantized value of  $x$  by the central quantizer. To derive the side distortions, observe the top-left portion of the matrix can be filled as shown in Fig. 13. Notice that the top-left corner is not filled with indices, which causes certain irregularity in the side distortion.

The side distortions generated by the  $2k+1$  macro-cells near each boundary are different from that of the inside macro-cells.

For the inside macro-cells, the side distortion can be calculated using the same approach as in Section IV-C, and we denote the contribution of distortion from these macro-cells as  $D_{Q1}^A$ . Correspondingly the contribution of distortion from the macro-cells near the boundaries is denoted as  $D_{Q1}^B$ . With each inside macro-cell contributes side distortion in the amount bounded in the interval

$$\left( \frac{\sum_{j=1}^{k+1} (j-1)^2 s_j}{N^3}, \frac{\sum_{j=1}^{k+1} j^2 s_j}{N^3} \right)$$

there exist  $w_j$ 's such that  $j-1 < w_j < j$  and

$$D_{Q1} = D_{Q1}^A + D_{Q1}^B > D_{Q1}^A = \frac{N-4k-2}{N^3} \sum_{j=1}^{k+1} w_j^2 s_j. \quad (\text{A53})$$

We denote the right-hand side of this inequality as  $D_L$ .

On the other hand, for the macro-cells near the boundaries, the distortion contribution can be bounded above, if we treat them in the same way as the inside macro-cells, which is obviously not optimal. And this gives

$$D_{Q1} = D_{Q1}^A + D_{Q1}^B < \frac{N-4k-2}{N^3} \sum_{j=1}^{k+1} w_j^2 s_j + \frac{4k+2}{N^3} \sum_{j=1}^{k+1} j^2 s_j \quad (\text{A54})$$

because  $w_j < j$ . We denote the right-hand side of this inequality as  $D_U$ .

Thus,

$$D_U > \frac{1}{N^2} \sum_{j=1}^{k+1} w_j^2 s_j > D_L \quad (\text{A55})$$

$$D_U > D_{Q1} > D_L. \quad (\text{A56})$$

Now it suffices to show that  $\lim_{N \rightarrow \infty} \frac{D_U}{D_L} = 1$ , which is equivalent to showing

$$\lim_{N \rightarrow \infty} \frac{\frac{4k+2}{N^3} \sum_{j=1}^{k+1} j^2 s_j}{\frac{N-4k-2}{N^3} \sum_{j=1}^{k+1} w_j^2 s_j} = 0. \quad (\text{A57})$$

Observe that  $12(w_j)^2 > j^2$  for any integer  $j > 1$ , because  $w_j > j - 1$ . Consider the two cases when  $s_1 > 0.5$  or  $s_1 \leq 0.5$ .

- $s_1 > 0.5$ :

In this case, each class 1 inner-cell occupies more than half of its macro-cell. Recall that  $w_j^2$  is the expected squared distance between points in class  $j$  inner-cell to their side quantizer reconstruction points, normalized by the squared length of a macro-cell. With the uniform distribution (see also Fig. 8), we have

$$w_1^2 \geq \int_0^{s_1} \frac{1}{s_1} x^2 dx = \frac{s_1^2}{3} > \frac{1}{12}. \quad (\text{A58})$$

Thus,  $12w_1^2 > 1$ , which implies  $12(w_j)^2 > j^2$  for any  $j \geq 1$ . Thus,

$$\begin{aligned} \frac{\frac{4k+2}{N^3} \sum_{j=1}^{k+1} j^2 s_j}{\frac{N-4k-2}{N^3} \sum_{j=1}^{k+1} w_j^2 s_j} &< \frac{(4k+2) \sum_{j=1}^{k+1} (12w_j^2) s_j}{(N-4k-2) \sum_{j=1}^{k+1} w_j^2 s_j} \\ &= 12 \frac{4k+2}{(N-4k-2)}. \end{aligned} \quad (\text{A59})$$

- $s_1 \leq 0.5$ :

In this case

$$\sum_{j=1}^{k+1} w_j^2 s_j \geq \sum_{j=2}^{k+1} s_j \geq 0.5$$

since  $w_j > j - 1$ . Thus,

$$\begin{aligned} \frac{\frac{4k+2}{N^3} \sum_{j=1}^{k+1} j^2 s_j}{\frac{N-4k-2}{N^3} \sum_{j=1}^{k+1} w_j^2 s_j} &< \frac{(4k+2)(s_1 + \sum_{j=1}^{k+1} (12w_j^2) s_j)}{(N-4k-2) \sum_{j=1}^{k+1} w_j^2 s_j} \\ &= \frac{4k+2}{(N-4k-2)} \left( \frac{s_1}{\sum_{j=1}^{k+1} w_j^2 s_j} + 12 \right) \\ &\leq 13 \frac{4k+2}{(N-4k-2)}. \end{aligned} \quad (\text{A60})$$

In both cases

$$\lim_{N \rightarrow \infty} \frac{\frac{4k+2}{N^3} \sum_{j=1}^{k+1} j^2 s_j}{\frac{N-4k-2}{N^3} \sum_{j=1}^{k+1} w_j^2 s_j} \leq \lim_{N \rightarrow \infty} 13 \frac{4k+2}{(N-4k-2)} = 0 \quad (\text{A61})$$

since  $\lim_{N \rightarrow \infty} \frac{k}{N} = 0$  as given in the condition. This completes the proof.  $\square$

## ACKNOWLEDGMENT

The authors are grateful to Sergio Servetto for the helpful discussions. They also thank Ram Zamir, Vivek Goyal, Vinay Vaishampayan, and the anonymous reviewers for many valuable and constructive suggestions and comments.

## REFERENCES

- [1] L. Ozarow, "On a source-coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, pp. 1909–1921, Dec. 1980.
- [2] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 851–857, Nov. 1982.
- [3] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 502–521, July 1987.
- [4] R. Zamir, "Gaussian codes and Shannon bounds for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. 45, pp. 2629–2636, Nov. 1999.
- [5] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 821–834, May 1993.
- [6] V. A. Vaishampayan and J. C. Batllo, "Asymptotic analysis of multiple description quantizers," *IEEE Trans. Inform. Theory*, vol. 44, pp. 278–284, Jan. 1998.
- [7] V. A. Vaishampayan and J. Domaszewicz, "Design of entropy-constrained multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 40, pp. 245–250, Jan. 1994.
- [8] V. K. Goyal and J. Kovačević, "Generalized multiple description coding with correlating transforms," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2199–2224, Sept. 2001.
- [9] S. D. Servetto, K. Ramchandran, V. A. Vaishampayan, and K. Nahrstedt, "Multiple description wavelet based image coding," *IEEE Trans. Image Processing*, vol. 9, pp. 813–826, May 2000.
- [10] M. Srinivansan and R. Chellappa, "Multiple description subband coding," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Oct. 1998, pp. 684–688.
- [11] S. Yang and V. A. Vaishampayan, "Low-delay communication for Rayleigh fading channels: An application of the multiple description quantizer," *IEEE Trans. Commun.*, vol. 43, pp. 2771–2783, Nov. 1995.
- [12] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2325–2383, Oct. 1998.
- [13] H. Jafarkhani and V. Tarokh, "Multiple description trellis-coded quantization," *IEEE Trans. Commun.*, vol. 47, pp. 799–803, June 1999.
- [14] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 373–380, July 1979.
- [15] S. P. Lloyd, "Least Squares Quantization in pcm," Bell Labs., unpublished memorandum, 1957.
- [16] P. F. Panter and W. Dite, "Quantization in pulse-count modulation with nonuniform spacing of levels," *Proc. IRE*, vol. 39, pp. 44–48, July 1951.
- [17] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 1, pp. 31–42, Jan. 1989.
- [18] V. K. Goyal, J. A. Kelner, and J. Kovačević, "Multiple description vector quantization with a coarse lattice," *IEEE Trans. Inform. Theory*, vol. 48, pp. 781–788, Mar. 2002.
- [19] C. Tian and S. S. Hemami, "Optimality and sub-optimality of multiple description vector quantization with a lattice codebook," *IEEE Trans. Inform. Theory*, to be published.
- [20] Y. Frank-Dayana and R. Zamir, "Dithered lattice-based quantizers for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. 48, pp. 192–204, Jan. 2002.
- [21] J. H. Kasner, M. W. Marcellin, and B. R. Hunt, "Universal trellis coded quantization," *IEEE Trans. Image Processing*, vol. 8, pp. 1677–1687, Jun. 1999.