On the Training Distortion of Vector Quantizers

Tamás Linder, Member, IEEE

Abstract—The in-training-set performance of a vector quantizer as a function of its training set size is investigated. For squared error distortion and independent training data, worst case type upper bounds are derived on the minimum training distortion achieved by an empirically optimal quantizer. These bounds show that the training distortion can underestimate the minimum distortion of a truly optimal quantizer by as much as a constant times \( r_\gamma^{-3/2} \), where \( r_\gamma \) is the size of the training data. Earlier results provide lower bounds of the same order.

Index Terms—Empirical design, training distortion, vector quantization, worst case bounds.

I. INTRODUCTION

Vector quantizer design is usually based on a collection of example vectors, called the training set or training data. In general, the objective of a design algorithm (such as the popular generalized Lloyd algorithm [1]) is to find an empirically optimal quantizer, that is, a quantizer of a given codebook size whose distortion in quantizing the training data is minimum. The underlying principle of empirical design is that good performance inside the training set will imply good performance on other data produced by the source if the training set size is sufficiently large to represent well the source statistics. But training vectors may be costly to obtain and the computational cost of design may become prohibitive for large training sets. Therefore, it is of interest to quantify how the performance of the designed vector quantizer improves as the size of the training set increases.

Assume that the quantizer dimension and the codebook size are fixed. For any quantizer \( Q_n \) trained on \( n \) vectors, let \( D_n(Q_n) \) denote the training distortion of \( Q_n \) (its average distortion inside the training set) and let \( D(Q_n) \) denote the test distortion of \( Q_n \) (its distortion in coding independent test data). Note that both \( D_n(Q_n) \) and \( D(Q_n) \) are functions of the training set and therefore are random quantities. The quantity \( D(Q_n) \) is the “true” distortion of the designed quantizer; it is the performance figure one wants to be as close as possible to \( D(Q^*) \), the distortion of a truly optimal quantizer \( Q^* \). A design procedure is called consistent if the test distortion \( D(Q_n) \) of the resulting quantizer \( Q_n \) converges (in some sense) to its lower bound \( D(Q^*) \) as \( n \to \infty \).

Of particular interest are the empirically optimal quantizers \( Q_n^* \) minimizing the training distortion: \( D_n(Q_n^*) \) = \( \min Q_n D_n(Q_n) \). The consistency of empirically optimal quantizers was first investigated by Pollard [2], [3] for the case of mean-squared quantizer distortion. His results show, among other things, that for a stationary and ergodic training sequence, the test distortion \( D(Q_n^*) \) of an empirically optimal quantizer converges to \( D(Q^*) \) with probability one as \( n \to \infty \).

Pollard’s results imply that the performance of empirically optimal quantizers will approach the optimum performance as the training set size increases without bound. On the other hand, to determine the training set size sufficient for achieving a preassigned level of performance, one needs to study the dependence of \( D(Q_n^*) \) on \( n \). Assume that the training set consists of \( n \) independent sample vectors drawn from the source distribution and let \( E[D(Q_n^*)] \) denote the expected value (taken over the training sequence) of the mean-squared test distortion \( D(Q_n^*) \). In [4] it was shown that for all source distributions supported by a given bounded region, the test distortion of the empirically optimal quantizer satisfies

\[
E[D(Q_n^*)] - D(Q^*) \leq c n^{-1/2}
\]

for some positive constant \( c \). This upper bound was shown to have the right order in a minimax sense in [5], where it was demonstrated that for any quantizer design method, there exist “bad” source distributions for which the test distortion of the resulting quantizer \( Q_n \) is lower-bounded as \( E[D(Q_n^*)] - D(Q^*) \geq c_1 n^{-1/2} \) for another positive constant \( c_1 \). The sample behavior of \( D(Q_n^*) - D(Q^*) \) for a class of smooth source densities was studied by Chou [6], and upper bounds on \( E[D(Q_n^*)] - D(Q^*) \) for dependent (mixing) training data were developed by Zeevi [7]. The dependence of the test distortion on the training set size was also empirically investigated by Cosman et al. [8] and Cohn et al. [9] in the context of image coding.

In this correspondence, the focus of attention is the less studied training distortion \( D_n(Q_n^*) \). Since the value of \( D_n(Q_n^*) \) is obtained as a by-product of the design procedure without requiring additional test data, it can be considered an inexpensive estimate of \( D(Q_n^*) \) or \( D(Q^*) \). For an empirically optimal quantizer minimizing \( D_n(Q_n) \), one always has

\[
E[D_n(Q_n^*)] \leq D(Q^*) \leq E[D(Q_n^*)].
\]

The exact relationship between the training, test, and optimal distortions is only known in the special case of quantizers with codebook size \( k = 1 \). In this case, it is easy to see that the single unique codepoint of the empirically optimal quantizer is the arithmetic average of the \( n \) training vectors, and therefore,

\[
E[D_n(Q_n^*)] = D(Q^*) \left(1 - \frac{1}{n}\right)
\]

and

\[
E[D(Q_n^*)] = D(Q^*) \left(1 + \frac{1}{n}\right)
\]

for all source distributions with finite second moment.

The problem becomes nontrivial when quantizers with more than one codepoint are considered, and in general little is known about the size of the difference \( D(Q^*) - E[D_n(Q_n^*)] \). In this respect, Abaya and Wise [10] proved that under general conditions the expected training distortion is a consistent estimate of the optimal distortion in the sense that \( D(Q^*) - E[D_n(Q_n^*)] \to 0 \) as \( n \to \infty \). The size of the bias of \( D_n(Q_n^*) \) in estimating \( D(Q^*) \) was first investigated in a recent work by Kim and Bell [11] who showed that for squared error distortion

\[
E[D_n(Q_n^*)] \leq D(Q^*) \left(1 - \frac{1}{n}\right)
\]


Manuscript received July 14, 1999; revised February 1, 2000. This work was supported in part by Queen’s University, Kingston, Ont., Canada, and by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

The author is with the Department of Mathematics and Statistics, Queen’s University, Kingston, Ont., Canada K7L 3N6 (e-mail: linder@mast.queensu.ca).

Communicated by P. A. Chou, Associate Editor for Source Coding.
for any source distribution with a finite second moment. No matching lower bounds or sharper upper bounds seem to be available in the literature.

In this correspondence, we apply techniques developed in [5] for proving minimax bounds in quantizer design to show that in the worst case, the difference $D(Q^*) - E[D_n(Q_n^*)]$ is proportional to $n^{-1/2}$. After introducing the necessary definitions in Section II, three results concerning the mean-squared training distortion of an empirically optimal quantizer are given in Section III. Theorem 1 proves the existence of “badly behaved” distributions on a bounded support set for which

$$E[D_n(Q_n^*)] \leq D(Q^*) - \frac{c}{\sqrt{n}}$$

(2)

for a constant $c > 0$ which depends on the quantizer dimension, the codebook size (which is assumed to be at least 3), and the diameter of the support set. Theorem 2 reformulates this bound in terms of the training ratio $\beta = n/k$ (where $k \geq 3$ is the codebook size) by showing that there exist source distributions for which

$$E[D_n(Q_n^*)] \leq D(Q^*) - \frac{c_0}{\sqrt{\beta}}$$

(3)

where $c_0 > 0$ is a universal constant. Theorem 3 presents an improved, explicit form of an earlier result in [4] to show that the lower bound

$$E[D_n(Q_n^*)] \geq D(Q^*) - \frac{\hat{c}}{\sqrt{n}}$$

holds for a $\hat{c} > 0$, uniformly for all sources supported on a given bounded set. This shows that bound (2) is tight in the sense that only the constants may be improved. The proofs of these results are given in Section IV.

The bounds (2) and (3) immediately demonstrate that for larger values of $n$ any bound in the form of (1) will be very loose for some source distributions. On the other hand, note that (1) holds for all source distributions while (2) and (3) are worst case bounds. Thus our results do not exclude the possibility that the $n^{-1}$ term in (1) has the right order for a restricted class of “smooth” source distributions. Potential candidates are the source densities satisfying Pollard’s central limit theorem [12] for empirical quantizer design.

II. PRELIMINARIES AND PROBLEM FORMULATION

A vector quantizer $Q$ of dimension $d$ and codebook size $k$ is a (measurable) mapping of the $d$-dimensional Euclidean space $\mathbb{R}^d$ into a finite set of points $\{y_1, \ldots, y_k\}$. The points $y_i \in \mathbb{R}^d$, $i = 1, \ldots, k$ are called the codepoints or codevectors and the collection $\{y_1, \ldots, y_k\}$ is called the codebook.

For any $x \in \mathbb{R}^d$, let $||x||$ denote its Euclidean norm. Given a $d$-dimensional random vector $X$ with probability distribution $\mu_X$ and finite second moment $E[||X||^2] < \infty$, the mean-squared distortion of a vector quantizer $Q$ is

$$D(Q) = E[||X - Q(X)||^2] = \int_{\mathbb{R}^d} ||x - Q(x)||^2 \mu_X(dx).$$

A vector quantizer $Q$ with codebook $\{y_1, \ldots, y_k\}$ is called a nearest neighbor quantizer if for all $x \in \mathbb{R}^d$

$$||x - Q(x)||^2 = \min_{1 \leq i \leq k} ||x - y_i||^2.$$

For any source distribution, a nearest neighbor quantizer has minimum distortion among all other quantizers with the same codebook. This fact allows us to consider only nearest neighbor quantizers in this correspondence without loss of generality.

For any $k \geq 1$, let $\hat{Q}_k$ be the family of all $d$-dimensional nearest neighbor vector quantizers with $k$ codevectors. A quantizer $Q^* \in \hat{Q}_k$ is called an optimal $k$-point quantizer for $\mu_X$ if it has minimum distortion

$$D(Q^*) = \min_{Q \in \hat{Q}_k} E[||X - Q(X)||^2].$$

(An optimal $Q^*$ always exists if $E[||X||^2] < \infty$, see, e.g., [3].)

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) $d$-dimensional random vectors drawn according to $\mu_X$. The collection $\{X_i\}_{i=1}^n$ is called the training data or training set. The average squared distortion of a vector quantizer $Q$ on the training set is

$$D_n(Q) = \frac{1}{n} \sum_{i=1}^n ||X_i - Q(X_i)||^2.$$

Let $Q_n^*$ denote an empirically optimal quantizer in $\hat{Q}_k$, that is, $Q_n^*$ is a $k$-point quantizer which has minimum average squared distortion

$$D_n(Q_n^*) = \min_{Q \in \hat{Q}_k} \frac{1}{n} \sum_{i=1}^n ||X_i - Q(X_i)||^2$$

over the training set. The random quantity $D_n(Q_n^*)$ is called the training distortion of the empirically optimal quantizer. Note that the dependence of $Q_n^*$ and $D_n(Q_n^*)$ on the training data $\{X_i\}_{i=1}^n$ is suppressed in the notation.

Our goal is to compare the expected training distortion

$$E[D_n(Q_n^*)] = E\left\{ \min_{Q \in \hat{Q}_k} \frac{1}{n} \sum_{i=1}^n ||X_i - Q(X_i)||^2 \right\}$$

of an empirically optimal quantizer $Q_n^*$ with the distortion $D(Q^*)$ of an optimal quantizer $Q^*$. Since

$$D_n(Q^*) \geq D_n(Q_n^*)$$

and

$$D(Q^*) = E[D_n(Q^*)]$$

we always have

$$D(Q^*) \geq E[D_n(Q_n^*)].$$

Moreover, it is easy to see that strict inequality holds whenever $D(Q^*) > 0$. Let $\mathcal{P}$ denote the collection of all source distributions which are supported by a given bounded set. Our results concern the maximum deviation over $\mathcal{P}$ of the expected training distortion from the optimal distortion, that is, the quantity

$$\sup_{\mu_X \in \mathcal{P}} \left( D(Q^*) - E[D_n(Q_n^*)] \right).$$

In order to be consistent with earlier work [13], [5] on worst case bounds in vector quantization, we will formulate our results in terms of classes $\mathcal{P}(B)$ containing all source distributions which satisfy the peak power constraint $P\{1/2||X||^2 \leq B\} = 1$. In other words, for any $B > 0$, the class $\mathcal{P}(B)$ consists of all source distributions whose support is contained in the ball $\{x : ||x|| \leq \sqrt{B}\}$.

III. RESULTS

Our first result shows that for training data of size $n$, the difference $D(Q^*) - E[D_n(Q_n^*)]$ of the minimum distortion of an optimal quantizer and the expected training distortion of the empirically optimal quantizer can be as large as constant times $n^{-1/2}$. 
Theorem 1: For any quantizer dimension \( d \geq 1 \) and codebook size \( k \geq 3 \) there exists a distribution \( \mu_X \in \mathcal{P}(B) \) such that for all training set size \( n \geq \frac{2}{3} k \)
\[
E[D_n(Q^*_n)] \leq D(Q^*) - \frac{c(B, d, k)}{\sqrt{n}} \tag{4}
\]
where
\[
c(B, d, k) = \frac{Bd \sqrt{k}^{1-\frac{d}{2}}}{2^d 3^d}.
\]

In the next result the relative difference of the training and optimal distortions is considered, in which case a very simple bound can be obtained in terms of the training ratio \( \beta = \frac{n}{k} \).

Theorem 2: For any quantizer dimension \( d \geq 1 \) and codebook size \( k \geq 3 \) there exists a distribution \( \mu_X \in \mathcal{P}(B) \) such that for all training set size \( n \geq \frac{2}{3} k \)
\[
E[D_n(Q^*_n)] \leq D(Q^*) \left(1 - \frac{c_0}{\sqrt{n}}\right)
\]
where \( c_0 = \frac{1}{2} \sqrt{\frac{2}{3}} \approx 0.27 \).

Theorems 1 and 2 are proved by using a construction of “bad” distributions introduced in [5]. This method uses discrete distributions supported by a finite number of points, although a modified construction using distributions with smooth densities is possible at the expense of complicating an already somewhat involved argument. An important point is that in [5] the choice of these “bad” distributions depends on the training set size \( n \). In our case, due apparently to the fact that we deal with the training distortion instead of the test distortion, we are able to construct one “bad” distribution which works for all large enough \( n \). Therefore, Theorem 1 guarantees the existence of at least one fixed source distribution in \( \mathcal{P}(B) \) such that
\[
\lim_{n \to \infty} \inf \sqrt{n} (D(Q^*) - E[D_n(Q^*_n)]) > 0.
\]

Next we examine in what sense (if any) the bound of Theorem 1 is tight. The constant \( c(B, d, k) \) is rather small and can probably be improved. But the more fundamental question is whether \( n^{-1/2} \) can be replaced with something larger. To answer this question in the negative, we note that for all \( \mu_X \in \mathcal{P}(B) \)
\[
D(Q^*) - E[D_n(Q^*_n)] = E \left\{ \sup_{Q \in \mathcal{Q}_k} [D(Q) - D_n(Q)] \right\}
\]
where \( \mathcal{Q}_k \) denotes the family of all \( k \)-point nearest neighbor quantizers with codepoints inside the sphere \( \{ x : \| x \| \leq \sqrt{d}B \} \) (see the proof of Theorem 3). Any uniform upper bound on the expectation on the right-hand side will result in a uniform lower bound on \( E[D_n(Q^*_n)] \).

The existence of such an upper bound of order \( n^{-1/2} \) has been pointed out in [4] (see [4, the discussion following Corollary 1]) although in an asymptotic form and without explicit constants. Nevertheless, such a bound implies that the bound of Theorem 1 is essentially tight.

The following theorem presents a new form of this lower bound which is tighter than those given by existing results and has a more attractive, nonasymptotic form.

Theorem 3: For any quantizer dimension \( d \geq 1 \), codebook size \( k \geq 1 \), training set size \( n \geq 1 \), we have
\[
E[D_n(Q^*_n)] \geq D(Q^*) - \frac{\hat{c}(B, d, k)}{\sqrt{n}}
\]
for all \( \mu_X \in \mathcal{P}(B) \), where
\[
\hat{c}(B, d, k) = 96 B d^{3/2} \sqrt{k}.
\]

The result is based on a nonasymptotic upper bound on
\[
E \left\{ \sup_{Q \in \mathcal{Q}_k} [D(Q) - D_n(Q)] \right\}.
\]

At the core of the proof is a simple and elegant version of the classic “metric entropy” bound [14], [15] of empirical process theory, recently proved by Cesa-Bianchi and Lugosi [16], which allows us to provide an explicit form of the constant \( \hat{c}(B, d, k) \).

In summary, Theorems 1 and 3 show that for independent training data of size \( n \), the maximum difference \( D(Q^*) - E[D_n(Q^*_n)] \) of the distortion of an optimal quantizer and the expected training distortion of the empirically optimal quantizer is of order \( n^{-1/2} \). More formally, these results imply that for all \( k \geq 3 \) and large enough \( n \)
\[
\frac{c}{\sqrt{n}} \leq \sup_{\mu_X \in \mathcal{P}(B)} (D(Q^*) - E[D_n(Q^*_n)]) \leq \frac{\hat{c}}{\sqrt{n}}
\]
for some constants \( c, \hat{c} > 0 \) depending on \( d, k, \) and \( B \).

IV. PROOFS

Proof of Theorem 1: We simplify the notation by assuming that \( B = 1 \) (that is, \( \mu_X \) has to satisfy \( P(||X||^2 \leq \sqrt{d}) = 1 \)). Since we consider mean-squared distortion, for arbitrary \( B > 0 \) the result follows by straightforward scaling.

To demonstrate the existence of a \( \mu_X \) satisfying the bound of the theorem, we will use a modified form of a construction introduced in [5, the proof of Theorem 1]. Just as in [5], the basic idea is to construct a source distribution such that with constant positive probability, the empirically optimal quantizer is sufficiently “far” from the optimal quantizer. However, new techniques are needed to derive the desired bound since we consider the training distortion (the empirically optimal quantizer is a function of the data on which its distortion is evaluated), while in [5] the test distortion was considered (the distortion is evaluated on independent data).

Assume that \( k \geq 3 \) is divisible by 3 (we will relax this assumption later) and let \( m = \frac{k}{3} \) (note that \( m \) is even). Let \( \Delta > 0 \) be a constant to be specified later and let \( z_1, \ldots, z_m \) be points in \( \mathbb{R}^d \) satisfying \( \| z_i - z_j \| \geq 3\Delta \) for all \( i \neq j \). Let \( w \) denote the \( d \)-vector \( w = (\Delta, 0, \ldots, 0) \). The proposed \( \mu_X \) is the uniform distribution concentrated on the \( 2m \) points \( \{ z_i, z_i + w, i = 1, \ldots, m \} \), that is,
\[
\mu_X (\{ z_i \}) = \mu_X (\{ z_i + w \}) = \frac{1}{2m}, \quad 1 \leq i \leq m.
\]

The parameters of \( \mu_X \) are \( \Delta \) and the points \( z_1, \ldots, z_m \). We assume that \( z_1, \ldots, z_m \) and \( \Delta \) are such that \( \mu_X \in \mathcal{P}(1) \), i.e.,
\[
\max_{1 \leq i \leq n} (\| z_i \|, \| z_i + w \|) \leq \sqrt{d}.
\]

Clearly, if \( \Delta \) is small enough this is always possible; the specific choice of \( \Delta \) will be given later. A key feature of \( \mu_X \) is that an optimal quantizer \( Q^* \) for \( \mu_X \) with \( k = \frac{3}{2} m \) has a very simple structure.

Lemma 1: Let \( \mu_X \) be defined by (6) and assume that \( \| z_i - z_j \| \geq 3\Delta \) for all \( i \neq j \). Let \( S \) be any subset of \( \{ 1, \ldots, m \} \) of cardinality \( |S| = m/2 \). Then the quantizer which has one codepoint at \( z_i + \frac{1}{\sqrt{d}} w \) for each \( i \in S \) and has codepoints at both \( z_i \) and \( z_i + w \) for each \( i \in \{ 1, \ldots, m \} \setminus S \) is an optimal \( k \)-point quantizer for \( \mu_X \).

The assertion of the lemma is intuitively clear; the proof is given in [5, the Appendix]. Note that the optimal quantizer is not unique, and in fact there are \( \binom{m}{m/2} \) optimal quantizers for \( \mu_X \).

Let the training data \( X_1, X_2, \ldots, X_n \) be drawn independently from \( \mu_X \) and let \( N \) be the number of training data points falling in the set \( \{ z_i, z_i + w \} \), i.e.,
\[
N_i = |\{ j : X_j = z_i \text{ or } X_j = z_i + w, j = 1, \ldots, n \}|
\]
Let $Q^*$ have one codepoint at $z_i + \frac{1}{2} w$ for each $i \leq m/2$ and two codepoints at $z_i$ and $z_i + w$ for each $m/2 + 1 \leq i \leq m$. Then $Q^*$ is an optimal $k$-point quantizer by Lemma 1, and its distortion is given in terms of the $N_i$ by

$$D(Q^*) = E \left\{ \frac{1}{n} \sum_{j=1}^{n} ||X_j - Q^*(X_j)||^2 \right\}$$

$$= E \left\{ \frac{\Delta^2}{4} \frac{1}{n} \sum_{i=1}^{m/2} N_i \right\}$$

(7)

where the second equality holds because $||Q^*(X_j) - X_j||^2 = \Delta^2/4$ if $X_j$ takes value in $\bigcup_{i=1}^{m/2} [z_i, z_i + w]$ and $||Q^*(X_j) - X_j||^2 = 0$ otherwise.

We now define a training-set-dependent quantizer $Q_n$ to approximate the empirically optimal $k$-point quantizer $Q_n^*$. Let $\sigma(1), \cdots, \sigma(m)$ be the permutation of $1, \cdots, m$ obtained by switching the positions of the indices $i$ and $m/2 + i$ (i.e., letting $\sigma(i) = m/2 + i$ and $\sigma(m/2 + i) = i$) for each $i \leq m/2$ such that $N_i > N_{m/2+i}$. Furthermore, let $Q_n$ be the $k$-point quantizer whose codepoints are $z_{\sigma(i)} + \frac{1}{2} w$ for $i \leq m/2$, and $z_{\sigma(i)} + w$ for $m/2 + 1 \leq i \leq m$. Then we have

$$E[D_n(Q_n)] = E \left\{ \frac{1}{n} \sum_{j=1}^{n} ||X_j - Q_n(X_j)||^2 \right\}$$

$$= E \left\{ \frac{\Delta^2}{4} \frac{1}{n} \sum_{i=1}^{m/2} N_{\sigma(i)} \right\}$$

(8)

since $||Q_n(X_j) - X_j||^2 = \Delta^2/4$ if

$$X_j \in \bigcup_{i=1}^{m/2} [z_{\sigma(i)}, z_{\sigma(i)} + w]$$

and $||Q_n(X_j) - X_j||^2 = 0$ otherwise. Since the empirically optimal quantizer $Q_n^*$ minimizes the training distortion over all $k$-point quantizers, we have

$$E[D_n(Q_n)] \geq E[D_n(Q_n^*)].$$

Therefore, using (7) and (8), we can lower-bound the difference $D(Q^*) - E[D_n(Q_n^*)]$ as

$$D(Q^*) - E[D_n(Q_n^*)] \geq \frac{\Delta^2}{4} \frac{1}{n} E \left\{ \sum_{i=1}^{m/2} (N_i - N_{\sigma(i)}) \right\}.$$ 

(9)

In the rest of the proof we will demonstrate that the expectation on the right-hand side is of order $n^{-1/2}$. First note that for all $i \leq m/2$ we have $N_{\sigma(i)} = N_i$ if $N_i \leq N_{m/2+i}$, and $N_{\sigma(i)} = N_{m/2+i}$ otherwise. Therefore, $N_i - N_{\sigma(i)} = (N_i - N_{m/2+i})^+$, where $x^+ = \max(x, 0)$. Thus

$$E \left\{ \sum_{i=1}^{m/2} (N_i - N_{\sigma(i)}) \right\} = E \left\{ \sum_{i=1}^{m/2} (N_i - N_{m/2+i})^+ \right\}$$

$$= \frac{m}{2} E[(N_1 - N_{m/2+i})^+]$$

(10)

since the pairs $(N_i, N_{m/2+i})$ have the same distribution. For each $j \in \{1, \cdots, m\}$ define the random variable $Y_j$ as follows: $Y_j = 1$ if the training vector $X_j$ contributes to $N_1$, $Y_j = -1$ if the training vector $X_j$ contributes to $N_{m/2+1}$, and $Y_j = 0$ otherwise. Then

$$P\{Y_j = 1\} = P\{Y_j = -1\} = \frac{1}{m}$$

and

$$P\{Y_j = 0\} = 1 - \frac{2}{m}.$$ 

Define

$$S_n = \sum_{j=1}^{n} Y_j.$$

Then $S_n = N_1 - N_{m/2+1}$ and since $S_n$ is distributed symmetrically about zero

$$E[(N_1 - N_{m/2+1})^+] = \frac{1}{2} E|S_n|.$$ 

(11)

To lower-bound the last expectation we will use the following useful inequality: for any random variable $Z$ with finite fourth moment

$$E[Z] \geq \left( \frac{E[Z^2]}{(E[Z])^2} \right)^{3/2}$$

(12)

(see [17, p. 194] or [18, Lemma A.4]). Since the $Y_j$ are independent and identically distributed, and have zero mean, we have

$$E[S_n^2] = \frac{2n}{m} E[Y_1^2] = \frac{2n}{m}.$$ 

On the other hand, expanding

$$S_n = \left( \sum_{j=1}^{n} Y_j \right)^4$$

yields

$$E[S_n^4] = n E[Y_1^4] + 3n(n-1) \left( E[Y_1^2] \right)^2$$

$$\leq 4 \left( \frac{2n}{m} \right)^2$$

where the inequality holds if $n \geq m$. Hence (12) gives

$$E|S_n| \geq \frac{1}{\sqrt{2}} \sqrt{\frac{n}{m}}.$$ 

(13)

Combine this with (11), (10), and (9) to obtain

$$D(Q^*) - E[D_n(Q_n^*)] \geq \frac{\Delta^2}{4m} \sqrt{\frac{m}{n}}.$$ 

To maximize this lower bound, we need to make $\Delta$ as large as possible under the constraint $\mu_N \in \mathcal{P}(1)$. A simple packing argument shows (see [5, Step 14, proof of Theorem 1]) that the choice

$$\Delta = \sqrt{\frac{7}{4m}}$$

is possible while maintaining the separation condition

$$||z_i - z_j|| \geq 3 \Delta, \quad i \neq j$$

and also satisfying

$$\max_{1 \leq i \leq m} (||z_i||, ||z_i + w||) \leq \sqrt{d}.$$
Substituting $\Delta = \frac{\sqrt{d}}{4(m+1)n}$ and $m = \frac{3}{4} k$ in (13) we can conclude that

$$D(Q^*) - E[D_n(Q_n^*)] \geq \frac{d}{2n} \sqrt{k - \frac{3}{n}}$$

(14)

which proves the statement of the theorem for all $k \geq 3$ divisible by 3 and $n \geq \frac{3}{2} k$.

The proof for the case when $k$ is not a multiple of 3 involves a slightly modified construction. In this case, we let $m$ be the unique even positive integer satisfying $k = 3m/2 + p$, where $p$ is either 1 or 2. In the definition of the modified $\mu_X$ the points $z_i, z_i + w$ are assigned probability $\frac{1}{m+1}$, and we augment the support of $\mu_X$ by one additional point with probability $\frac{1}{(m+1)^2}$ (when $p = 1$), or a pair of points, each having probability $\frac{1}{2(m+1)}$ (when $p = 2$). Since we now have $m + 1$ pairs, we set

$$\Delta = \frac{\sqrt{d}}{4(m+1)n}.$$

The details of the derivation are omitted since these are almost identical to the case when $k$ is divisible by 3. Instead of (13), in this case we obtain the slightly weaker bound

$$D(Q^*) - E[D_n(Q_n^*)] \geq \frac{\Delta^2}{2} \sqrt{\frac{m}{m+1}} \sqrt{\frac{1}{n}}$$

$$\geq \frac{\Delta^2}{2} \sqrt{\frac{2}{3}} \sqrt{\frac{m}{m+1}}$$

$$\geq \frac{d}{2} \sqrt{\frac{k - \frac{3}{n}}{n}}$$

(15)

where the second inequality holds because $m \geq 2$ and the third holds because $m \geq \frac{3}{4} (k - 2)$ and $k \geq 4$.

**Proof of Theorem 2:** The construction in the proof of Theorem 1 is used again. Assume first that $k$ is divisible by 3. Then by (7) we have

$$D(Q^*) = \frac{\Delta^2}{8},$$

since $E(N_0) = \frac{n}{m}$. Hence (13) can be rewritten as

$$E[D_n(Q_n^*)] \leq D(Q^*) - \frac{\Delta^2}{2} \sqrt{\frac{m}{m+1}} \sqrt{\frac{1}{n}}$$

$$= D(Q^*) \left(1 - \frac{1}{2} \sqrt{\frac{m+1}{m}} \right)$$

$$= D(Q^*) \left(1 - \frac{1}{2} \sqrt{\frac{k}{m}} \right).$$

If $k \geq 3$ is not divisible by 3, then $\mu_X$ is modified as in the last part of the proof of Theorem 1. In this case, the distortion of $Q^*$ is

$$D(Q^*) = \frac{\Delta^2}{8} \frac{m}{m+1}$$

where $m$ is the unique even positive integer such that $k = 3m/2 + p$, where $p$ is either 1 or 2. Then (15) implies

$$E[D_n(Q_n^*)] \leq D(Q^*) - \frac{\Delta^2}{2} \sqrt{\frac{m}{m+1}} \sqrt{\frac{1}{n}}$$

$$= D(Q^*) \left(1 - \frac{1}{2} \sqrt{\frac{m+1}{m}} \right)$$

$$\leq D(Q^*) \left(1 - \frac{1}{2} \sqrt{\frac{k}{m}} \right),$$

where the second inequality holds since $m \geq \frac{3}{4} (k - 2)$ and $k \geq 4$.

**Proof of Theorem 3:** As in the proof of Theorem 1, we assume that $B = 1$ and obtain the result for general $B$ by scaling. Since $X, X_1, \ldots, X_n$ is an i.i.d. sequence and $Q^*$ is a $k$-point quantizer with minimum distortion, we can write

$$D(Q^*) - E[D_n(Q_n^*)] \leq \frac{d}{2n} \sqrt{k - \frac{3}{n}}$$

$$\leq \frac{E \left\{ \sup_{Q \in \mathcal{Q}_k} \right\}}{E \left\{ \sup_{Q \in \mathcal{Q}_k} \right\}}$$

(16)

where $\mathcal{Q}_k$ denotes the family of all $k$-point nearest neighbor quantizers with codepoints inside the sphere $S(\sqrt{d}) = \{ x : ||x|| \leq \sqrt{d} \}$. The second inequality holds since $P(\|X\| \leq \sqrt{d}) = 1$ for all $i$ and therefore the codepoints of $Q^*$ are inside $S(\sqrt{d})$ with probability one.

For any $Q \in \mathcal{Q}_k$ let the random variable $T^{(Q)}_n$ be defined by

$$T^{(Q)}_n = \frac{1}{2} \sum_{i=1}^n (E[||X_i - Q(X_i)||^2] - ||X_i - Q(X_i)||^2)$$

$$= \frac{n}{2} (D(Q) - D_n(Q))$$

so that by (16)

$$D(Q^*) - E[D_n(Q_n^*)] \leq \frac{2}{n} \sup_{Q \in \mathcal{Q}_k} \left\{ T^{(Q)}_n \right\}.$$  

(17)

We will use a standard but effective technique of empirical process theory to upper-bound the expectation on the right-hand side.

First we recall some definitions. Let $(S, \rho)$ be a totally bounded metric space. For any $F$ $\subseteq$ $S$ and $\epsilon > 0$ the $\epsilon$-covering number $N_S(F, \epsilon)$ of $F$ is defined as the minimum number of closed balls with radius $\epsilon$ whose union covers $F$.

A family $\{T_i : s \in S\}$ of zero-mean random variables indexed by the metric space $(S, \rho)$ is called subgaussian in the metric $\rho$ if for any $\lambda > 0$ and $s, s' \in S$ we have

$$E(e^{\lambda \rho(s, s')}) \leq e^{\lambda^2 \rho(s, s')/2}.$$

The family $\{T_i : s \in S\}$ is called sample continuous if for any sequence $s_1, s_2, \ldots \in S$ such that $s_j \to s \in S$ we have $T_{s_j} \to T_s$ with probability one.

The following result gives an upper bound on the expected supremum of the random variables $\{T_i : s \in S\}$ in terms of the covering number of the index space. It provides a version of a classical result in empirical process theory (see, e.g., [15]) with an explicit constant.

**Lemma 2 ([16, Proposition 3]):** If $\{T_i : s \in S\}$ is subgaussian and sample continuous in the metric $\rho$, then

$$E \left\{ \sup_{s \in S} T_i \right\} \leq 12 \int_0^{\text{diam}(S)/2} \sqrt{\ln N_S(S, \epsilon)}d\epsilon$$

where $\text{diam}(S) = \sup_{s, s' \in S} \rho(s, s')$ is the diameter of $S$.

To apply the above result we need to show that when $\mathcal{Q}_k$ is equipped with a suitable metric, the family of random variables $\{T^{(Q)}_n : Q \in \mathcal{Q}_k\}$ is subgaussian and sample continuous. For any $Q, Q' \in \mathcal{Q}_k$ define

$$\rho_n(Q, Q') = \sqrt{n} \sup_{||x||^2 \leq \epsilon} \left(||x - Q(x)||^2 - ||x - Q'(x)||^2\right).$$

Clearly, $\rho_n$ is a metric on $\mathcal{Q}_k$. Also, for any $Q, Q' \in \mathcal{Q}_k$ we have

$$T^{(Q)}_n - T^{(Q')}_n \leq \sqrt{n} \rho_n(Q, Q')$$

(18)

with probability one, which implies that $\{T^{(Q)}_n : Q \in \mathcal{Q}_k\}$ is sample continuous. To show that $\{T^{(Q)}_n : Q \in \mathcal{Q}_k\}$ is subgaussian in $\rho_n$, we recall Hoeffding’s inequality [19] which states that if $Y_1, \ldots, Y_n$
are independent zero-mean random variables such that $a \leq Y_i \leq b$, $i = 1, \ldots, n$ with probability one, then for all $\lambda > 0$
\[
E \left[ e^{\lambda \sum_{i=1}^{n} Y_i^2} \right] \leq e^{\lambda^2 (b-a)^2/8}.
\]
For $i = 1, \ldots, n$ let
\[
Y_i = \frac{1}{2} \left( D(Q) - ||X_i - Q(X_i)||^2 \right) - \frac{1}{2} \left( D(Q') - ||X_i - Q'(X_i)||^2 \right).
\]
Then
\[
T_n^{(Q)} - T_n^{(Q')} = \sum_{i=1}^{n} Y_i
\]
where the $Y_i$ are independent, have zero mean, and
\[
|Y_i| \leq \frac{1}{\sqrt{n}} \rho_n(Q, Q')
\]
for all $i$. Hence Hoeffding’s inequality implies
\[
E \left[ e^{\lambda (T_n^{(Q)} - T_n^{(Q')})} \right] \leq e^{\lambda^2 \rho_n(Q, Q')^2/2}
\]
proving that $\{T_n^{(Q)} : Q \in \mathcal{Q}_k\}$ is subgaussian in $\rho_n$. Therefore, Lemma 2 gives
\[
E \left\{ \sup_{Q \in \mathcal{Q}_k} T_n^{(Q)} \right\} \leq 12 \int_0^{\min \{\rho_n(Q, Q')\}^2/2} \sqrt{\ln N_n(Q_k, \epsilon)} d\epsilon. \quad (19)
\]
To evaluate the integral we need the following bound on the covering number of $\mathcal{Q}_k$.

Lemma 3 ([5, Corollary 1]): For any $0 < \epsilon \leq 4d$ and $k \geq 1$, the covering number of $\mathcal{Q}_k$ in the metric
\[
\rho(Q, Q') = \sup_{||x||^2 \leq d} ||x - Q(x)||^2 - ||x - Q'(x)||^2
\]
is bounded as
\[
N_k(Q_k, \epsilon) \leq \left( \frac{16d}{\epsilon} \right)^{kd}.
\]
Since $\rho_n(Q, Q') = \sqrt{n} \rho(Q, Q')$, the preceding lemma implies that
\[
N_{kn}(Q_k, \epsilon) \leq \left( \frac{16d \sqrt{n}}{\epsilon} \right)^{kd}
\]
for all $0 < \epsilon \leq \sqrt{n} 4d$. Moreover, since
\[
\sup_{||x||^2 \leq d} ||x - Q(x)||^2 \leq 4d
\]
for all $Q \in \mathcal{Q}_k$, we have $\text{diam}(\mathcal{Q}_k) \leq \sqrt{n} 4d$. Therefore, (17) and (18) imply
\[
D(Q') = E[D_n(Q_n^* - Q_n)] \leq 24 \int_0^{\sqrt{n} 2d} \sqrt{\ln \left( \frac{16d \sqrt{n}}{\epsilon} \right)^{kd}} d\epsilon = 24 \sqrt{kd} \int_0^{\sqrt{n} 2d} \ln \left( \frac{16d \sqrt{n}}{\epsilon} \right) d\epsilon.
\]
We can upper-bound the last integral as
\[
\int_0^{\sqrt{n} 2d} \ln \left( \frac{16d \sqrt{n}}{\epsilon} \right) d\epsilon = 16d \sqrt{n} \int_0^{1/8} \ln \left( \frac{1}{x} \right) dx
\]
\[
\leq 2d \sqrt{n} \ln 8 + 1
\]
\[
\leq 4d \sqrt{n}
\]
where we first used the change of variable $x = \epsilon/(16d \sqrt{n})$ and then applied Jensen’s inequality to the concave function $f(t) = \sqrt{t}$. Combining this bound with (20) proves Theorem 3.

\section*{Acknowledgment}
The author wishes to thank G. Lugosi for helpful discussions. Also, thanks are due to an anonymous reviewer for pointing out how to simplify the proof of Theorem 1 in a way that improved the bound.

\section*{References}
Transform Coding with Backward Adaptive Updates

Vivek K Goyal, Member, IEEE, Jun Zhuang, and Martin Vetterli, Fellow, IEEE

Abstract—The Karhunen–Loève transform (KLT) is optimal for transform coding of a Gaussian source. This is established for all scale-invariant quantizers, generalizing previous results. A backward adaptive technique for combating the data dependence of the KLT is proposed and analyzed. When the adapted transform converges to a KLT, the scheme is universal among transform coders. A variety of convergence results are proven.

I. INTRODUCTION

The essence of transform coding is to apply a linear transform to a source vector and then apply scalar quantization, as opposed to applying scalar quantization directly to the source vector. Heuristically, transform coding works because the transform can eliminate correlation between components of the source vector, producing a vector of transform coefficients more amenable to scalar quantization and entropy coding. Transform codes are popular because they provide an attractive compromise between computational complexity and performance. In the parlance of vector quantization, the point-density and oblongitis losses of scalar quantization are eliminated or reduced, leaving predominantly only a space-filling loss [1].

With a Gaussian source model, the optimal transform is a Karhunen–Loève (KLT), an orthonormal transform that produces uncorrelated transform coefficients. The optimality of the KLT is well known for high rates 2] or when optimal fixed-rate quantizers are employed [3], but holds more generally (see Appendix I). However, the KLT is rarely used in practice for a variety of reasons. One prominent reason is that the KLT is signal-dependent; the transform used in the encoder and decoder must be adjusted to correspond to the covariance of the source in order to maintain optimality. A second reason is that since the KLT has no special structure, it requires more operations to compute than a harmonic transform such as a discrete cosine transform. For vectors of length of \(N\), the complexity difference is roughly \(N^2\) compared to \(N \log N\), which is not overwhelming for small values of \(N\).

This correspondence addresses only the first issue—the matching of transform to source. A backward adaptive method for transform adaptation is proposed and analyzed. In backward adaptation the encoder and decoder adapt in unison based on the coded data without the explicit transmission of coder parameters. Backward adaptation is also called adaptation without side information or on-line adaptation.

The use of backward adaptation for transform adaptation in transform coding seems to be unprecedented, though backward adaptive techniques have a long history. For example, adaptation of prediction filters in speech coders is often backward adaptive [4], [5] and ADPCM includes not only backward adaptation of filter taps but also of quantizer scaling [6]. Similar to the quantizer scaling in ADPCM is the backward adaptive context modeling and quantizer scaling of the EQ image coder [7]. It is also possible to adapt a quantizer more generally without side information [8].

The incompletely realized aim of our work is to show that backward adaptation can result in a transform code that is universal for Gaussian sources. “Universal” is used here to mean that the performance approaches that of an ideal transform code designed with a priori knowledge of the source distribution. The results along these lines are asymptotic in the data length, but the transform or block size is fixed. Empirical evidence and partial analyses are provided. Such a code would be an “on-line” alternative to the “universal codebook” approach to universal transform coding by Effros and Chou [9]. Forward adaptive techniques that are not necessarily universal are discussed, e.g., in [11].

The results of [9] were inspiring to this study because they indicated superior performance of weighted universal transform coding over weighted universal vector quantization for image compression with reasonable vector dimensions. It was also shown that there are sizable gains to be realized by varying the transform, a result that runs counter to the conventional wisdom in image compression.

In the remainder of the correspondence, the aforementioned ideas are made more precise. The sources and coding structures under consideration are described in Section II. Unable to satisfactorily analyze the original coding structure, we give several analyses based on simplifying assumptions. The main results are stated in Section III and proven in Appendix II. Section IV describes ways in which the encoding algorithms can be modified to reduce computational complexity or to track a varying source. Concluding comments appear in Section V.

II. PROPOSED BACKWARD ADAPTIVE CODING STRUCTURE

Let \(\{x_n\}_{n \in \mathbb{Z}}\) be a sequence of independent and identically distributed (i.i.d.), zero-mean Gaussian random vectors of dimension \(N\) with covariance matrix \(R_x = E[x x^T]\). If \(R_x\) is not diagonal, i.e., the components of \(x\) are correlated, one obtains better rate-distortion performance with transform coding than with direct scalar quantization and scalar entropy coding of the source vectors.

In transform coding, a square, invertible linear transform \(T\) is applied to each source vector to get a vector of transform coefficients \(y_n = T x_n\). The transform coefficients undergo scalar quantization

\[1\] See the taxonomy of universal coding methods by Zhang and Wei [10] for explanations of the quoted terms.

\[2\] Throughout the correspondence, \(R_x\) will be used to denote the (exact) covariance matrix \(E[vv^T]\) of a random vector \(v\). \(R_x^T\) denotes an estimate of \(R_x\), obtained from a finite-length observation. Aside from this convention, subscripts indicate the time index of a variable, except where two subscripts are given to indicate the row and column indices of a matrix. A superscript \(T\) indicates a transpose.