

# Learning-Theoretic Methods in Vector Quantization

Lecture Notes for the  
Advanced School on the Principles of Nonparametric Learning  
Udine, Italy, July 9-13, 2001.

To appear in:  
*Principles of Nonparametric Learning*  
L. Györfi, editor, CISM Lecture Notes, Wien, New York: Springer 2001.

**Tamás Linder**  
Department of Mathematics & Statistics  
and  
Department of Electrical & Computer Engineering  
Queen's University  
Kingston, Ontario, Canada, K7L 3N6  
email: [linder@mast.queensu.ca](mailto:linder@mast.queensu.ca)

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>The fixed-rate quantization problem</b>	<b>2</b>
<b>3</b>	<b>Consistency of empirical design</b>	<b>9</b>
<b>4</b>	<b>Finite sample upper bounds</b>	<b>15</b>
<b>5</b>	<b>Minimax lower bounds</b>	<b>21</b>
<b>6</b>	<b>Fundamentals of variable-rate quantization</b>	<b>32</b>
<b>7</b>	<b>The Lagrangian formulation</b>	<b>36</b>
<b>8</b>	<b>Consistency of Lagrangian empirical design</b>	<b>40</b>
<b>9</b>	<b>Finite sample bounds in Lagrangian design</b>	<b>45</b>

# 1 Introduction

The principal goal of data compression (also known as source coding) is to replace data by a compact representation in such a manner that from this representation the original data can be reconstructed either perfectly, or with high enough accuracy. Generally, the representation is given in the form of a sequence of binary digits (bits) that can be used for efficient digital transmission or storage.

Certain types of data, such as general purpose data files on a computer, require perfect reconstruction. In this case the compression procedure is called *lossless*, and the goal is to find a representation that allows perfect reconstruction using the fewest possible bits. Other types of data, such as speech, audio, images, and video signals, do not require (or do not even admit) perfect reconstruction. In this case, the goal is to find an efficient digital representation from which the original data can be reconstructed with a prescribed level of accuracy, as measured by a fidelity criterion between the original and reconstructed data. Such a compression procedure is called *lossy*. In these notes, we will focus on lossy data compression.

In our model, the data to be compressed is a sequence of  $d$ -dimensional random vectors  $Z_1, Z_2, \dots$ , such that all the  $Z_i$  have the same distribution. Such a data sequence is obtained, for example, by forming non-overlapping blocks of length  $d$  from a real-valued stationary process. To compress the data, each  $Z_i$  is mapped into a binary string  $b_i$ . Thus  $Z_1, Z_2, \dots$  is represented by the sequence of binary strings  $b_1, b_2, \dots$ . Typically,  $Z_i$  can take a continuum of values, while  $b_i$  is always discrete, and so this representation is lossy (not invertible). The data is reconstructed by mapping each  $b_i$  into another  $d$ -dimensional vector  $\hat{Z}_i$ , called the reproduction of  $Z_i$ .

The compactness of the representation is measured by its *rate*, defined as the average (expected) number of binary digits needed to obtain  $\hat{Z}_i$ , i.e., the average length of  $b_i$ . Note that this number is the same for all  $i$  since the  $Z_i$  have the same distribution. If  $b_i$  has fixed length, the compression procedure is called *fixed-rate vector quantization*. If the length of  $b_i$  is not fixed (i.e., it depends on the value of  $Z_i$ ), we talk about *variable-rate* vector quantization. The composition of mappings  $Z_i \rightarrow b_i \rightarrow \hat{Z}_i$  is called a fixed-rate vector quantizer in the first case, and a variable-rate vector quantizer in the second case.

Since  $Z_i \neq \hat{Z}_i$  in general, we need a way to measure how well  $\hat{Z}_i$  approximates  $Z_i$ . For this reason, we are given a nonnegative function  $d(\cdot, \cdot)$  of two vector variables, called a *distortion measure*, and we use the quantity  $d(Z_i, \hat{Z}_i)$  to measure the reconstruction error in representing  $Z_i$  by  $\hat{Z}_i$ . In

these notes, we will use the popular mean squared error given by the squared Euclidean distance between  $Z_i$  and  $\hat{Z}_i$ . The *distortion* of the scheme is characterized by a single number, the average (expected) value of  $d(Z_i, \hat{Z}_i)$ . Again, this quantity is the same for all  $i$  since the  $Z_i$  have the same distribution.

The fact that the rate and distortion do not depend on the particular index  $i$  allows us to focus on the problem of quantizing (compressing) a generic random vector  $X$ , called the *source*, which has the common distribution of the  $Z_i$ . The goal is to make the distortion in quantizing  $X$  as small as possible, while keeping the rate at a given threshold. Quantizers which are optimal in the sense of achieving minimum distortion under a rate constraint depend on the distribution of the source. If the source distribution is known, then the problem of optimal quantization under a rate constraint can be posed as a (typically rather hard) optimization problem.

On the other hand, if the source distribution is unknown (as is often the case in practice), then an approximation to an optimal quantizer must be constructed (learned) on the basis of a finite number of training samples drawn from the source distribution. Questions of a different flavor arise in this situation. For example, a fundamental problem is whether an optimal quantizer can be learned as the number of training samples increases without bound. We would also like to know how many training samples are needed and what methods to use to construct a quantizer whose performance is close to the optimum. Our main goal in these notes is to demonstrate how tools and techniques from nonparametric statistics and statistical learning theory can be used to tackle these and related problems concerning learning vector quantizers from empirical data.

## Notes

The model of data compression we consider is not the most general possible. For more general models and the information-theoretic framework for lossy data compression, see, e.g., Gray (1990). Fundamentals of vector quantization are given in Gersho and Gray (1992). A very thorough review of the history of quantization and an extensive survey of its literature can be found in Gray and Neuhoff (1998).

## 2 The fixed-rate quantization problem

A *fixed-rate  $N$ -point vector quantizer* ( $N \geq 1$  is an integer) is a Borel measurable mapping  $q : \mathbb{R}^d \rightarrow \mathcal{C}$ , where the *codebook*  $\mathcal{C} = \{y_1, \dots, y_N\}$  is an ordered collection of  $N$  distinct points in  $\mathbb{R}^d$ , called the *codevectors*.

In our model, the source is an  $\mathbb{R}^d$ -valued random vector  $X$ . In order to compress the source, the quantizer  $q$  represents the “input”  $X$  by the “output”  $\widehat{X} = q(X)$ . Since  $q(X)$  can only take  $N$  distinct values, it is possible to uniquely describe  $q(X)$  using only a finite number of bits, while such a description is in general not possible for  $X$ . In fact, for values of  $N$  such that  $\log_2 N$  is an integer, exactly  $\log_2 N$  bits are necessary and sufficient to uniquely specify the values of  $q(X)$  in a description using binary strings of a fixed length. In other words, for every realization of  $X$ , one needs to transmit or store  $\log_2 N$  bits so that the value of  $q(X)$  can be reconstructed. For this reason, the the *rate* of  $q$  is defined by

$$R(q) \triangleq \log_2 N.$$

In general, it is convenient (and customary) to define the rate by  $\log_2 N$  even if this number is not an integer.

To measure the reconstruction error in representing  $X$  by  $q(X)$ , we use the quantity  $d(X, q(X))$ , where  $d(x, y) \geq 0$ ,  $x, y \in \mathbb{R}^d$ , is a measurable function called a *distortion measure*. The *distortion* of  $q$  in quantizing  $X$  is the expected reconstruction error

$$D(\mu, q) \triangleq \mathbb{E} d(X, q(X)) = \int_{\mathbb{R}^d} d(x, q(x)) \mu(dx)$$

where  $\mu$  denotes the distribution of  $X$ . For simplicity, we assume mean squared distortion, i.e.,  $d(x, y) = \|x - y\|^2$ , where  $\|\cdot\|$  denotes the Euclidean norm in  $\mathbb{R}^d$ , so that

$$D(\mu, q) \triangleq \mathbb{E} \|X - q(X)\|^2.$$

Throughout we assume that  $\mathbb{E} \|X\|^2 < \infty$ , which implies  $D(\mu, q) < \infty$ .

Two quantizers have the same rate if they have the same number of codevectors. The primary goal of quantization is to find quantizers that have minimum distortion subject to a constraint on the rate, or equivalently, on the number of codevectors. Consequently, we define the optimal performance over  $N$ -point quantizers by

$$D_N^*(\mu) \triangleq \inf_{q \in \mathcal{Q}_N} D(\mu, q)$$

where  $\mathcal{Q}_N$  denotes the set of all  $N$ -point quantizers. A quantizer  $q^* \in \mathcal{Q}_N$  is called *optimal* if  $D(\mu, q^*) = D_N^*(\mu)$ .

It is useful to note that any  $N$ -point quantizer  $q$  is completely characterized by its codebook  $\{y_i\}_{i=1}^N$  and the *cells*  $S_i = \{x : q(x) = y_i\}$ ,  $i = 1, \dots, N$ , via the rule

$$q(x) = y_i \quad \text{if and only if} \quad x \in S_i.$$

Note that  $\{S_1, \dots, S_N\}$  is a partition of  $\mathbb{R}^d$ . In what follows we will often define quantizers by specifying their codebooks and partition cells.

The next lemma shows how to choose the partition cells optimally for a fixed codebook, and how to choose the codebook optimally for fixed partition cells.

**Lemma 1**

(NEAREST NEIGHBOR CONDITION) *Let  $q$  be an arbitrary  $N$ -point quantizer with codebook  $\mathcal{C} = \{y_i\}_{i=1}^N$ , and let  $q'$  be an  $N$ -point quantizer which has the same codebook and is defined by*

$$q'(x) = \arg \min_{y_i \in \mathcal{C}} \|x - y_i\|^2 \quad (1)$$

*where ties are broken arbitrarily. Then*

$$D(\mu, q') \leq D(\mu, q).$$

(CENTROID CONDITION) *Let  $q$  be an arbitrary quantizer with partition cells  $\{S_i\}_{i=1}^N$ . If  $q'$  is defined to have the same partition cells and codevectors given by*

$$y'_i = \arg \min_{y \in \mathbb{R}^d} \mathbb{E}[\|X - y\|^2 | X \in S_i] = \mathbb{E}[X | X \in S_i], \quad i = 1, \dots, N$$

*then*

$$D(\mu, q') \leq D(\mu, q).$$

PROOF. To prove the nearest neighbor condition, note that (1) is equivalent to

$$\|x - q'(x)\|^2 = \min_{1 \leq i \leq N} \|x - y_i\|^2$$

i.e,  $q(x)$  is the nearest neighbor of  $x$  among  $\{y_i\}_{i=1}^N$ . Thus for any  $q$  with codebook  $\mathcal{C}$  and arbitrary partition cells  $\{S_i\}_{i=1}^N$ ,

$$\begin{aligned} \mathbb{E}\|X - q(X)\|^2 &= \sum_{j=1}^N \int_{S_j} \|x - y_j\|^2 \mu(dx) \\ &\geq \sum_{j=1}^N \int_{S_j} \min_{1 \leq i \leq N} \|x - y_i\|^2 \mu(dx) \\ &= \int_{\mathbb{R}^d} \min_{1 \leq i \leq N} \|x - y_i\|^2 \mu(dx) \\ &= \mathbb{E}\|X - q'(X)\|^2. \end{aligned}$$

To prove the centroid condition, note that for any measurable  $S \subset \mathbb{R}^d$  such that  $\mu(S) > 0$ , if  $y' = \mathbb{E}[X|X \in S]$  and  $y \in \mathbb{R}^d$  is arbitrary, then

$$\mathbb{E}[\|X - y\|^2|X \in S] = \mathbb{E}[\|X - y'\|^2|X \in S] + \|y - y'\|^2$$

since  $\mathbb{E}[(y - y') \cdot (X - y')|X \in S] = 0$ , where  $a \cdot b$  denotes the usual inner product of  $a, b \in \mathbb{R}^d$ . (This optimizing  $y'$  is often called the *centroid* of  $S'$ .) Hence  $y'$  is the unique point satisfying

$$\mathbb{E}[\|X - y'\|^2|X \in S] = \inf_{y \in \mathbb{R}^d} \mathbb{E}[\|X - y\|^2|X \in S].$$

Thus if  $q$  has partition cells  $\{S_i\}_{i=1}^N$  and arbitrary codebook  $\{y_i\}_{i=1}^N$ , then

$$\begin{aligned} \mathbb{E}\|X - q(X)\|^2 &= \sum_{i=1}^N \mathbb{E}[\|X - y_i\|^2|X \in S_i]\mu(S_i) \\ &\geq \sum_{i=1}^N \mathbb{E}[\|X - y'_i\|^2|X \in S_i]\mu(S_i) \\ &= \mathbb{E}\|X - q'(X)\|^2. \end{aligned}$$

□

A quantizer  $q$  with codebook  $\mathcal{C} = \{y_i\}_{i=1}^N$  is called a *nearest neighbor* quantizer if for all  $x \in \mathbb{R}^d$ ,

$$\|x - q(x)\|^2 = \min_{y_i \in \mathcal{C}} \|x - y_i\|^2.$$

The nearest neighbor condition of Lemma 1 implies that it suffices to consider nearest neighbor quantizers when searching for an optimal quantizer. Equivalently,

$$D_N^*(\mu) = \inf_{\mathcal{C}:|\mathcal{C}|=N} \mathbb{E} \min_{y_i \in \mathcal{C}} \|X - y_i\|^2. \quad (2)$$

Note that (2) clearly implies

$$D_N^*(\mu) \leq D_{N+1}^*(\mu) \quad (3)$$

for any  $N \geq 1$ .

Although the partition cells of a nearest neighbor quantizer are not uniquely determined by the codebook  $\mathcal{C} = \{y_i\}_{i=1}^N$ , one can make the definition unique via a fixed tie-breaking rule. For example, a tie-breaking rule that favors smaller indices gives

$$S_1 = \{x : \|x - y_1\| \leq \|x - y_j\|, j = 1, \dots, N\}$$

and for  $i = 2, \dots, N$ ,

$$S_i = \{x : \|x - y_i\| \leq \|x - y_j\|, j = 1, \dots, N\} \setminus \bigcup_{k=1}^{i-1} S_k.$$

Any  $\{S_i\}_{i=1}^N$  obtained as the partition associated with a nearest neighbor quantizer with codebook  $\mathcal{C}$  is called a *Voronoi* (or nearest neighbor) *partition* of  $\mathbb{R}^d$  with respect to  $\mathcal{C}$ .

In view of (2), it is not hard to show that an optimal  $N$ -point quantizer always exists.

**Theorem 1** *There exists a nearest neighbor quantizer  $q^* \in \mathcal{Q}_N$  such that  $D(\mu, q^*) = D_N^*(\mu)$ .*

PROOF. For a positive integer  $m$  and  $(y_1, \dots, y_m) \in (\mathbb{R}^d)^m$  define

$$g_m(y_1, \dots, y_m) \triangleq \int_{\mathbb{R}^d} \min_{1 \leq i \leq m} \|x - y_i\|^2 \mu(dx).$$

Note that  $g_m$  is the distortion of a nearest neighbor quantizer with codevectors  $y_1, \dots, y_m$  (this quantizer may have less than  $m$  codevectors since the  $y_i$  are not necessarily distinct). Hence by (3),

$$\inf_{(y_1, \dots, y_m) \in (\mathbb{R}^d)^m} g_m(y_1, \dots, y_m) = D_m^*(\mu). \quad (4)$$

We can assume that  $N \geq 2$  since for  $N = 1$  the claim of the theorem follows from the centroid condition of Lemma 1. Also, to exclude the trivial case when  $\mu$  is concentrated on a single point, we assume that the support of  $\mu$  contains at least two distinct points. Then it is easy to show that

$$D_2^*(\mu) < D_1^*(\mu)$$

(to see this, note that by the proof of the centroid condition, if  $q$  is an arbitrary two-point nearest-neighbor quantizer with cells  $S_1, S_2$  and codepoints  $y_1, y_2$  such that  $\mu(S_i) > 0$  and  $y_i = \mathbb{E}[X|X \in S_i]$  for  $i = 1, 2$ , then  $D(\mu, q) < \mathbb{E}\|X - \mathbb{E}X\|^2 = D_1^*(\mu)$ ). Hence there is a unique integer  $2 \leq k \leq N$  such that

$$D_N^*(\mu) = \dots = D_k^*(\mu) < D_{k-1}^*(\mu). \quad (5)$$

(In fact, one can prove that if the support of  $\mu$  contains at least  $N$  points, then  $D_N^*(\mu) < D_{N-1}^*(\mu)$ , and so  $k = N$ ).

Let  $B_r \triangleq \{x : \|x\| \leq r\}$  denote the closed ball of radius  $r > 0$  centered at the origin. Fix  $\epsilon > 0$  such that

$$\epsilon < \frac{1}{2}(D_{k-1}^*(\mu) - D_k^*(\mu)) \quad (6)$$

and pick  $0 < r < R$  such that

$$(R-r)^2\mu(B_r) > D_k^*(\mu) + \epsilon, \quad 4 \int_{B_{2R}^c} \|x\|^2 \mu(dx) < \epsilon. \quad (7)$$

Choose  $(y_1, \dots, y_k)$  satisfying  $g_k(y_1, \dots, y_k) < D_k^*(\mu) + \epsilon$ , and suppose without loss of generality that the codevectors are indexed so that  $\|y_1\| \leq \dots \leq \|y_k\|$ . Then  $\|y_1\| \leq R$ , since otherwise, by the triangle inequality,  $\min_{1 \leq i \leq k} \|x - y_i\|^2 \geq (R-r)^2$  for all  $x \in B_r$ , and so

$$D_k^*(\mu) + \epsilon > \int_{B_r} \min_{1 \leq i \leq k} \|x - y_i\|^2 \mu(dx) \geq (R-r)^2\mu(B_r)$$

contradicting (7). We will show that  $\|y_j\| \leq 5R$  for all  $j$ . Assume to the contrary that  $\|y_k\| > 5R$ . Then by the triangle inequality, for all  $x \in \mathbb{R}^d$ ,

$$\|x - y_1\| \leq \|x - y_k\| I_{\{x \in B_{2R}\}} + 2\|x\| I_{\{x \in B_{2R}^c\}} \quad (8)$$

where  $I_A$  denotes the indicator of the set  $A$ . Then letting  $\{S_i\}_{i=1}^k$  be a Voronoi partition with respect to  $\{y_i\}_{i=1}^k$ , we obtain

$$\begin{aligned} g_{k-1}(y_1, \dots, y_{k-1}) &= \sum_{j=1}^k \int_{S_j} \min_{1 \leq i \leq k-1} \|x - y_i\|^2 \mu(dx) \\ &\leq \sum_{j=1}^{k-1} \int_{S_j} \|x - y_j\|^2 \mu(dx) + \int_{S_k} \|x - y_1\|^2 \mu(dx) \\ &\leq \sum_{j=1}^k \int_{S_j} \|x - y_j\|^2 \mu(dx) + 4 \int_{B_{2R}^c} \|x\|^2 \mu(dx) \\ &\leq g_k(y_1, \dots, y_k) + \epsilon \leq D_k^*(\mu) + 2\epsilon < D_{k-1}^*(\mu) \end{aligned}$$

where the second inequality follows from (8), the third from (7), and the last one from (6). This contradicts (4), so we obtain that  $g_k(y_1, \dots, y_k) < D_k^*(\mu) + \epsilon$  implies  $(y_1, \dots, y_k) \in (B_{5R})^k$ . Therefore

$$D_k^*(\mu) = \inf_{(y_1, \dots, y_k) \in (B_{5R})^k} g_k(y_1, \dots, y_k).$$

Since  $(B_{5R})^k \subset (\mathbb{R}^d)^k$  is compact and  $g_k$  is continuous (as can be seen by an application of the dominated convergence theorem), there exists  $(y_1^*, \dots, y_k^*)$  in  $(B_{5R})^k$  with  $g_k(y_1^*, \dots, y_k^*) = D_k^*(\mu)$ . Thus there is a nearest neighbor quantizer with at most  $k$  codevectors achieving  $D_k^*(\mu)$ . Since  $k \leq N$  and  $D_k^*(\mu) = D_N^*(\mu)$ , this implies that there must exist an  $N$ -point nearest neighbor quantizer  $q^*$  that achieves  $D_N^*(\mu)$ .  $\square$

**Remark** Lemma 1 gives rise to an iterative algorithm for designing  $N$ -point quantizers. Start with an arbitrary  $N$ -point quantizer  $q_0$  with codebook  $\mathcal{C}_0$  and partition  $\mathcal{S}_0$ . In the  $m$ th iteration ( $m = 1, 2, \dots$ ), first let  $\mathcal{S}_m = \{S_i^{(m)}\}_{i=1}^N$  be the Voronoi partition with respect to  $\mathcal{C}_{m-1}$ , and then set  $\mathcal{C}_m = \{y_i^{(m)}\}_{i=1}^N$ , where  $y_i^{(m)} = \mathbb{E}[X|X \in S_i^{(m)}]$  for  $i = 1, \dots, N$ . If  $q_m$  denotes the quantizer defined by  $\mathcal{C}_m$  and  $\mathcal{S}_m$ , and we set  $D_m = D(\mu, q_m)$ , then Lemma 1 implies

$$D_m \leq D_{m-1}$$

and so  $\lim_{m \rightarrow \infty} (D_{m-1} - D_m) = 0$ . The algorithm stops (after a finite number of iterations) when the drop in distortion falls below a given threshold. The distortion  $D_m$  is not guaranteed to converge to the minimum distortion  $D_N^*(\mu)$ , but quantizers obtained by this method or its variants yield sufficiently low distortion for practical applications.

Unless the dimension  $d$  is very small, computing  $D_m$  and the conditional expectations  $\mathbb{E}[X|X \in S_i^{(m)}]$ ,  $i = 1, \dots, N$ , is hard for a general source distribution (given, e.g., by its probability density function). In practice, the algorithm is usually run with  $\mu$  replaced by the empirical distribution of a finite number of samples drawn according to  $\mu$ . As explained at the end of the next section, the implementation becomes straightforward in this case.

## Notes

The optimality conditions of Lemma 1 were first derived by Lloyd (1957) (for the scalar  $d = 1$  case) and by Steinhaus (1956) (who considered a problem equivalent to three-dimensional fixed-rate quantization). Theorem 1 is due to Pollard (1982a); the existence of optimal quantizers for more general distortion measures was shown, for example, by Pollard (1981), Abaya and Wise (1982), and Sabin (1984). Except for trivial cases, optimal quantizers and the minimum distortion  $D_N^*(\mu)$  are very hard to determine analytically, but approximations that become tight as  $N \rightarrow \infty$  can be derived for a large class of source distributions. We refer to Gray and Neuhoff (1998) and Graf and Luschgy (2000) for such asymptotic (high-rate) results. The design algorithm sketched above is basically also due to Lloyd (1957) and Steinhaus (1956). An extension to the vector case and to more general distortion mea-

tures was given by Linde, Buzo, and Gray (1980), and the algorithm is often referred to as “the LBG algorithm.” For more details and other methods of vector quantizer design, see Gersho and Gray (1992).

### 3 Consistency of empirical design

In most situations, the distribution  $\mu$  of the source  $X$  to be quantized is unknown, and the only available information about  $\mu$  is in the form of *training data*, a finite sequence of vectors drawn according to  $\mu$ . More formally, the training data  $X_1^n \triangleq X_1, \dots, X_n$  consists of  $n$  independent and identically distributed (i.i.d.) copies of  $X$ . It is assumed that  $X_1^n$  and  $X$  are also independent. The training data is used to construct an  $N$ -point quantizer

$$q_n(\cdot) = q_n(\cdot, X_1, \dots, X_n).$$

Such a  $q_n$  is called an *empirically designed* quantizer. The goal is to “learn” the optimal quantizer from the data, i.e., to produce empirically designed quantizers with performance approaching (as  $n$  gets large) the performance of a quantizer optimal for  $X$ . We assume, as before, that  $\mathbb{E}\|X\|^2 < \infty$ .

We call the quantity

$$D(\mu, q_n) = \mathbb{E}[\|X - q_n(X)\|^2 | X_1^n]$$

the *test distortion* of  $q_n$ . Thus  $D(\mu, q_n)$  measures the distortion resulting when  $q_n$  is applied to  $X$ ; it is the “true” distortion of the empirically designed quantizer. Note that  $D(\mu, q_n)$  is a random variable since  $q_n$  depends on  $X_1^n$ .

Also of interest is the *training distortion* (or empirical distortion) of  $q_n$ , defined as the average distortion of  $q_n$  on the training data:

$$\frac{1}{n} \sum_{k=1}^n \|X_k - q_n(X_k)\|^2.$$

The empirical distribution  $\mu_n$  of the training data is defined by

$$\mu_n(A) = \frac{1}{n} \sum_{k=1}^n I_{\{X_k \in A\}}$$

for every Borel measurable  $A \subset \mathbb{R}^d$ , i.e.,  $\mu_n$  places weight  $1/n$  at each point  $X_k$ ,  $k = 1, \dots, n$ . Thus the training distortion becomes

$$\frac{1}{n} \sum_{k=1}^n \|X_k - q_n(X_k)\|^2 = D(\mu_n, q_n).$$

Intuitively, if  $q_n$  performs well on the training set, it should also have good performance on the source, assuming the training set is sufficiently large. We define an *empirically optimal* quantizer as an  $N$ -point quantizer  $q_n^*$  that minimizes the training distortion:

$$q_n^* \triangleq \arg \min_{q \in \mathcal{Q}_N} \frac{1}{n} \sum_{k=1}^n \|X_k - q(X_k)\|^2.$$

In other words,  $q_n^* \in \mathcal{Q}_N$  satisfies

$$D(\mu_n, q_n^*) = \inf_{q \in \mathcal{Q}_N} D(\mu_n, q) = D_N^*(\mu_n).$$

Note that by Theorem 1,  $q_n^*$  always exists. In fact, since  $\mu_n$  is supported on at most  $n$  points, the existence of  $q_n^*$  is easy to show without resorting to Theorem 1. Note also that the definition of  $q_n^*$  outside the support of  $\mu_n$  does not affect the training distortion, so  $q_n^*$  is not uniquely defined even if its codebook happens to be unique. We will resolve this problem by always requiring (as we may by Lemma 1) that  $q_n^*$  be a nearest neighbor quantizer.

Our goal is to show that the design based on empirical distortion minimization is *consistent* in the following sense: As the size of the training data grows, the sequence of test distortions converges almost surely (i.e., for almost every realization of the training sequence) to the minimum distortion achieved by an optimal quantizer. If this is the case, then for  $n$  large enough,  $q_n^*$  can effectively replace an optimal quantizer  $q^*$  in quantizing  $X$ .

**Theorem 2** (CONSISTENCY OF EMPIRICAL DESIGN) *For any  $N \geq 1$  the sequence of empirically optimal  $N$ -point nearest neighbor quantizers  $q_n^*$ ,  $n = 1, 2, \dots$ , satisfies*

$$\lim_{n \rightarrow \infty} D(\mu, q_n^*) = D_N^*(\mu) \quad a.s.$$

To prove the theorem we need some intermediate results. The basic idea is that for large  $n$  the empirical distribution  $\mu_n$  is a good estimate of  $\mu$ , so the optimal quantizer for  $\mu_n$  should provide a good approximation to the optimal quantizer for  $\mu$ . In order to formalize this idea, we need a measure of closeness for probability distributions that is appropriate in quantization arguments.

Let  $\mu$  and  $\nu$  be probability distributions on  $\mathbb{R}^d$  with finite second moment. The  $L_2$  Wasserstein distance between  $\mu$  and  $\nu$  is defined by

$$\rho(\mu, \nu) \triangleq \inf_{X \sim \mu, Y \sim \nu} (\mathbb{E} \|X - Y\|^2)^{1/2}$$

where the infimum is taken over all joint distributions of two random vectors  $X$  and  $Y$  such that  $X$  has distribution  $\mu$ , and  $Y$  has distribution  $\nu$  (denoted by  $X \sim \mu$  and  $Y \sim \nu$ , respectively). It is easy to see that  $\rho(\mu, \nu)$  is finite; in fact,  $\rho(\mu, \nu) \leq (\mathbb{E}\|X\|^2)^{1/2} + (\mathbb{E}\|Y\|^2)^{1/2}$  by the triangle inequality for the  $L_2$  norm.

**Lemma 2** *The infimum defining  $\rho(\mu, \nu)$  is a minimum. Moreover,  $\rho(\mu, \nu)$  is a metric on the space of probability distributions on  $\mathbb{R}^d$  with finite second moment.*

SKETCH OF PROOF. Let  $\mathcal{P}(\mu, \nu)$  denote the family of probability distributions on  $\mathbb{R}^d \times \mathbb{R}^d$  with marginals  $\mu$  and  $\nu$ . Then

$$\rho(\mu, \nu) = \inf_{P \in \mathcal{P}(\mu, \nu)} \left( \int \|x - y\|^2 P(dx, dy) \right)^{1/2}.$$

First we show that the infimum is in fact a minimum. Fix  $\epsilon > 0$  and pick a closed ball  $B \subset \mathbb{R}^d$  with large enough radius such that  $\mu(B) \geq 1 - \epsilon$  and  $\nu(B) \geq 1 - \epsilon$ . Then any  $P \in \mathcal{P}(\mu, \nu)$  has  $P(B \times B) \geq 1 - 2\epsilon$ . Since  $B \times B$  is compact, we obtain that the set of distributions  $\mathcal{P}(\mu, \nu)$  is tight (see, e.g., Ash (2000)). If  $P_k \in \mathcal{P}(\mu, \nu)$ ,  $k = 1, 2, \dots$ , is a sequence such that

$$\int \|x - y\|^2 P_k(dx, dy) < \rho(\mu, \nu)^2 + \frac{1}{k}$$

then by Prokhorov's theorem (see, e.g., Theorem 7.2.4 in Ash (2000)) there is a subsequence of  $\{P_k\}$ , say  $\{P'_k\}$ , such that, as  $k \rightarrow \infty$ ,  $P'_k$  converges weakly to some probability distribution  $P'$ . Clearly,  $P' \in \mathcal{P}(\mu, \nu)$ , and it is easy to show using a truncation argument that  $\int \|x - y\|^2 P'(dx, dy) = \rho(\mu, \nu)^2$ .

To show that  $\rho$  is a metric, note that  $\rho(\mu, \nu) = \rho(\nu, \mu) \geq 0$ , and if  $\rho(\mu, \nu) = 0$ , then by the preceding argument there exist  $X \sim \mu$  and  $Y \sim \nu$  such that  $\mathbb{E}\|X - Y\|^2 = 0$ , implying  $\mu = \nu$ . Thus it only remains to verify that  $\rho$  satisfies the triangle inequality.

Let  $\mu$ ,  $\nu$ , and  $\lambda$  be probability distributions on  $\mathbb{R}^d$  having finite second moment. Assume  $P \in \mathcal{P}(\mu, \nu)$  achieves  $\rho(\mu, \nu)$  and  $P' \in \mathcal{P}(\nu, \lambda)$  achieves  $\rho(\nu, \lambda)$ . Construct a jointly distributed triplet  $(X, Y, Z)$  by specifying that  $(X, Y) \sim P$ ,  $(Y, Z) \sim P'$ , and that  $X$  and  $Z$  are conditionally independent given  $Y$  (i.e.,  $X, Y, Z$  form a Markov chain in this order). Since  $X \sim \mu$  and  $Z \sim \lambda$ , by the triangle inequality for the  $L_2$  norm

$$\begin{aligned} \rho(\mu, \lambda) &\leq (\mathbb{E}\|X - Z\|^2)^{1/2} \\ &\leq (\mathbb{E}\|X - Y\|^2)^{1/2} + (\mathbb{E}\|Y - Z\|^2)^{1/2} \\ &= \rho(\mu, \nu) + \rho(\nu, \lambda). \end{aligned}$$

□

The next lemma justifies the choice of  $\rho$  by showing that if two distributions are close in  $\rho$  metric, then any nearest neighbor quantizer will quantize these distributions with similar distortion. This fact also implies the stability of optimal quantizer performance with respect to the  $\rho$  metric.

**Lemma 3** *If  $q$  is a nearest neighbor quantizer, then*

$$|D(\mu, q)^{1/2} - D(\nu, q)^{1/2}| \leq \rho(\mu, \nu).$$

Consequently,

$$|D_N^*(\mu)^{1/2} - D_N^*(\nu)^{1/2}| \leq \rho(\mu, \nu).$$

PROOF. To prove the first bound let  $\{y_1, \dots, y_N\}$  denote the codebook of  $q$ , and let  $X \sim \mu$  and  $Y \sim \nu$  achieve the minimum defining  $\rho(\mu, \nu)$ . Then

$$\begin{aligned} D(\mu, q)^{1/2} &= \left\{ \mathbb{E} \min_{1 \leq i \leq N} \|X - y_i\|^2 \right\}^{1/2} \\ &= \left\{ \mathbb{E} \left( \min_{1 \leq i \leq N} \|X - y_i\| \right)^2 \right\}^{1/2} \\ &\leq \left\{ \mathbb{E} \left( \min_{1 \leq i \leq N} (\|X - Y\| + \|Y - y_i\|) \right)^2 \right\}^{1/2} \\ &= \left\{ \mathbb{E} \left( \|X - Y\| + \min_{1 \leq i \leq N} \|Y - y_i\| \right)^2 \right\}^{1/2} \\ &\leq \left\{ \mathbb{E} \|X - Y\|^2 \right\}^{1/2} + \left\{ \mathbb{E} \min_{1 \leq i \leq N} \|Y - y_i\|^2 \right\}^{1/2} \\ &= \rho(\mu, \nu) + D(\nu, q)^{1/2}. \end{aligned}$$

The inequality  $D(\nu, q)^{1/2} - D(\mu, q)^{1/2} \leq \rho(\mu, \nu)$  is proved similarly.

To prove the second bound, assume  $q^*$  is an optimal  $N$ -point (nearest neighbor) quantizer for  $\nu$ . Then

$$\begin{aligned} D_N^*(\mu)^{1/2} - D_N^*(\nu)^{1/2} &= D_N^*(\mu)^{1/2} - D(\nu, q^*)^{1/2} \\ &\leq D(\mu, q^*)^{1/2} - D(\nu, q^*)^{1/2} \\ &\leq \rho(\mu, \nu) \end{aligned}$$

by the first bound of the lemma. The inequality  $D_N^*(\nu)^{1/2} - D_N^*(\mu)^{1/2} \leq \rho(\mu, \nu)$  is proved in a similar fashion by considering an  $N$ -point optimal quantizer for  $\mu$ . □

The following corollary relates the test distortion of the empirically optimal quantizer to the distortion of the optimal quantizer in terms of the  $\rho$  distance between the the empirical distribution and the true source distribution.

**Corollary 1** *The test distortion of the empirically optimal  $N$ -point quantizer  $q_n^*$  is upper bounded as*

$$D(\mu, q_n^*)^{1/2} - D_N^*(\mu)^{1/2} \leq 2\rho(\mu, \mu_n).$$

PROOF. Let  $q^*$  be an optimal  $N$ -point (nearest neighbor) quantizer for  $\mu$ . Recall that we specified  $q_n^*$  to be a nearest neighbor quantizer. Then

$$\begin{aligned} & D(\mu, q_n^*)^{1/2} - D_N^*(\mu)^{1/2} \\ &= D(\mu, q_n^*)^{1/2} - D(\mu, q^*)^{1/2} \\ &= D(\mu, q_n^*)^{1/2} - D(\mu_n, q_n^*)^{1/2} + D(\mu_n, q_n^*)^{1/2} - D(\mu, q^*)^{1/2} \\ &\leq D(\mu, q_n^*)^{1/2} - D(\mu_n, q_n^*)^{1/2} + D(\mu_n, q^*)^{1/2} - D(\mu, q^*)^{1/2} \\ &\leq 2\rho(\mu, \mu_n) \end{aligned}$$

where the last inequality follows from Lemma 3.  $\square$

The consistency theorem is a consequence of Corollary 1 and the following lemma.

**Lemma 4** *Given  $\nu$  and a sequence  $\{\nu_n\}$ ,  $\lim_{n \rightarrow \infty} \rho(\nu, \nu_n) = 0$  if and only if  $\nu_n \rightarrow \nu$  weakly and  $\int \|x\|^2 \nu_n(dx) \rightarrow \int \|x\|^2 \nu(dx)$  as  $n \rightarrow \infty$ .*

PROOF. Suppose  $\lim_{n \rightarrow \infty} \rho(\nu, \nu_n) = 0$ . For each  $n$  let  $Y \sim \nu$ ,  $Y_n \sim \nu_n$  have joint distribution such that  $\rho(\nu, \nu_n) = (\mathbb{E}\|Y - Y_n\|^2)^{1/2}$ . Then  $\lim_{n \rightarrow \infty} \mathbb{E}\|Y - Y_n\|^2 = 0$ , implying  $Y_n \rightarrow Y$  in distribution (i.e.,  $\nu_n \rightarrow \nu$  weakly) and  $\mathbb{E}\|Y_n\|^2 \rightarrow \mathbb{E}\|Y\|^2$  (i.e.,  $\int \|x\|^2 \nu_n(dx) \rightarrow \int \|x\|^2 \nu(dx)$ ).

Conversely, suppose  $\nu_n$  converges weakly to  $\nu$ . Then by Skorohod's theorem (see, e.g., Theorem 11.7.2 in Dudley (1989)) there exist  $Y_n \sim \nu_n$  and  $Y \sim \nu$  jointly distributed such that  $Y_n \rightarrow Y$  a.s. By the triangle inequality,  $2\|Y_n\|^2 + 2\|Y\|^2 - \|Y_n - Y\|^2 \geq \|Y_n\|^2 + \|Y\|^2 - 2\|Y_n\|\|Y\| \geq 0$ ; hence Fatou's lemma implies

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \mathbb{E}\{2\|Y_n\|^2 + 2\|Y\|^2 - \|Y_n - Y\|^2\} \\ & \geq \mathbb{E}\{\liminf_{n \rightarrow \infty} (2\|Y_n\|^2 + 2\|Y\|^2 - \|Y_n - Y\|^2)\} \\ & = 4\mathbb{E}\|Y\|^2. \end{aligned}$$

Thus, if  $E\|Y_n\|^2 \rightarrow E\|Y\|^2$ , we obtain that  $\mathbb{E}\|Y_n - Y\|^2 \rightarrow 0$ , implying  $\rho(\nu, \nu_n) \rightarrow 0$ .  $\square$

PROOF OF THEOREM 2. Clearly, we have  $D(\mu, q_n^*) \geq D_N^*(\mu)$  for each  $n$ . Thus by Corollary 1 the theorem holds if the empirical distributions converge to the true source distribution in  $\rho$  metric almost surely, i.e., if

$$\lim_{n \rightarrow \infty} \rho(\mu, \mu_n) = 0 \quad \text{a.s.} \quad (9)$$

By Lemma 4 this happens if, almost surely,  $\mu_n \rightarrow \mu$  weakly and  $\int \|x\|^2 \mu_n(dx) \rightarrow \int \|x\|^2 \mu(dx)$ . A basic result (due to Varadayan) concerning empirical distributions is that

$$\mathbb{P}\{\mu_n \rightarrow \mu \text{ weakly}\} = 1$$

(see, e.g., Theorem 11.4.1 in Dudley (1989)). Also, by the strong law of large numbers

$$\begin{aligned} \lim_{n \rightarrow \infty} \int \|x\|^2 \mu_n(dx) &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|X_i\|^2 \\ &= \mathbb{E}\|X\|^2 = \int \|x\|^2 \mu(dx) \quad \text{a.s.} \end{aligned}$$

Thus

$$\mathbb{P}\left\{\mu_n \rightarrow \mu \text{ weakly, } \int \|x\|^2 \mu_n(dx) \rightarrow \int \|x\|^2 \mu(dx)\right\} = 1$$

completing the proof.  $\square$

**Remarks** (i) The consistency theorem says that when trained on a long enough training sequence, empirically optimal quantizers perform almost as well as optimal ones. In a similar manner, we can also show that the training distortion of the empirically optimal quantizer is a strongly consistent estimate of the optimal distortion. Indeed, applying the second bound of Lemma 3 with  $\nu = \mu_n$ ,

$$|D(\mu_n, q_n^*)^{1/2} - D_N^*(\mu)^{1/2}| = |D_N^*(\mu_n)^{1/2} - D_N^*(\mu)^{1/2}| \leq \rho(\mu, \mu_n)$$

and so (9) implies

$$\lim_{n \rightarrow \infty} D(\mu_n, q_n^*) = D_N^*(\mu) \quad \text{a.s.}$$

However, for finite  $n$  the training distortion is optimistically biased: If  $q^*$  is an optimal  $N$ -level quantizer for  $\mu$ , then  $D(\mu_n, q_n^*) \leq D(\mu_n, q^*)$ , so

$$\begin{aligned} \mathbb{E}D(\mu_n, q_n^*) &\leq \mathbb{E}D(\mu_n, q^*) \\ &= \mathbb{E}\left\{\frac{1}{n} \sum_{k=1}^n \|X_k - q^*(X_k)\|^2\right\} \\ &= D(\mu, q^*) = D_N^*(\mu). \end{aligned}$$

In fact, it is not hard to see that the above inequality is always strict unless  $\mu$  is a discrete distribution concentrating on  $N$  points or less.

(ii) Finding an empirically optimal quantizer  $q_n^*$  for a given realization  $x_1, \dots, x_n$  of the training sequence is a computationally hard problem for quantizer dimensions  $d \geq 2$ . However, the iterative algorithm sketched at the end of Section 2 can be used with the empirical distribution of the training sequence replacing  $\mu$  to find a suboptimal, but usually good enough approximation to  $q_n^*$ . The implementation of the algorithm is straightforward in this case: Let  $\mu_n$  denote the empirical distribution of the samples  $x_1, \dots, x_n$ , and let  $Y \sim \mu_n$ . Then the computation of the new codevectors in the  $m$ th iteration reduces to

$$y_i^{(m)} = \mathbb{E}[Y|Y \in S_i^{(m)}] = \frac{\frac{1}{n} \sum_{k=1}^n x_k I_{\{x_k \in S_i^{(m)}\}}}{\frac{1}{n} \sum_{k=1}^n I_{\{x_k \in S_i^{(m)}\}}}, \quad i = 1, \dots, N.$$

Calculating the distortion  $D_m$  has similar complexity. Note that it is easy to decide whether or not  $x_k \in S_i^{(m)}$  since  $\{S_j^{(m)}\}_{j=1}^N$  is a Voronoi partition with respect to  $\{y_j^{(m-1)}\}_{j=1}^N$ . Indeed, we have  $x_k \in S_i^{(m)}$  if  $y_i^{(m-1)}$  is the nearest neighbor of  $x_k$  in the codebook  $\{y_j^{(m-1)}\}_{j=1}^N$ .

### Notes

The proof of Theorem 2 is based on Pollard (1982a); see also Pollard (1981) and (1982b), Abaya and Wise (1984), and Graf and Luschgy (1994). See Graf and Luschgy (2000) for more recent stability and consistency results related to fixed-rate quantization. The  $L_2$  Wasserstein metric  $\rho$  is a special case of the  $\bar{\rho}$  (“ $\rho$ -bar”) distance between random processes introduced by Gray, Neuhoff, and Shields (1975). Lemma 3 is due to Gray and Davisson (1975) who were the first to use the  $L_2$  Wasserstein metric in quantization arguments. The problem of constructing an empirically optimal quantizer is often referred to as the  $k$ -means clustering problem in the statistical literature (see MacQueen (1967)). Convergence results for the empirical version of the iterative quantizer design algorithm (the LBG algorithm) are given in Sabin and Gray (1986).

## 4 Finite sample upper bounds

We know from the consistency theorem (Theorem 2) that as the training data size increases without bound, the performance of the empirically optimal quantizer will approximate the optimal performance with arbitrary

accuracy. But large amounts of training data may be “expensive” to acquire, and the computational cost of finding a quantizer that minimizes (at least approximately) the training distortion may become prohibitive for large training sets. It is therefore of interest to quantify how fast the test distortion of the empirically optimal quantizer converges to the optimal distortion with increasing training set size.

In this section we develop finite sample bounds on the expected test (and training) distortion in empirical design. Such bounds are of both theoretical and practical interest. For example, an upper bound on the test distortion, if sufficiently tight, can give a useful bound on the minimum number of training samples sufficient to guarantee a preassigned level of performance for the designed quantizer.

To simplify matters, we only consider source distribution with bounded support. For  $T > 0$ , let  $\mathcal{P}(T)$  denote the set of probability distributions on  $\mathbb{R}^d$  supported on  $B_T$ , the closed ball of radius  $T$  centered at the origin. Throughout this section we require that  $\mu \in \mathcal{P}(T)$ , that is,

$$\mathbb{P}\{\|X\| \leq T\} = 1.$$

Let  $\mathcal{Q}_N(T)$  denote the collection of all nearest neighbor quantizers with all their codevectors inside  $B_T$ . We start with a basic lemma.

**Lemma 5** *For any  $\mu \in \mathcal{P}(T)$ ,*

$$D(\mu, q_n^*) - D_N^*(\mu) \leq 2 \sup_{q \in \mathcal{Q}_N(T)} |D(\mu_n, q) - D(\mu, q)|.$$

PROOF. Let  $q^*$  be an optimal  $N$ -point nearest neighbor quantizer for  $\mu$ . We can repeat the argument in the proof of Corollary 1 without taking square roots to obtain

$$\begin{aligned} D(\mu, q_n^*) - D_N^*(\mu) \\ \leq D(\mu, q_n^*) - D(\mu_n, q_n^*) + D(\mu_n, q^*) - D(\mu, q^*). \end{aligned} \quad (10)$$

Now observe that since  $B_T$  is a convex set, if an arbitrary  $N$ -point quantizer  $q$  has a codevector  $y_i$  outside  $B_T$ , we can replace  $y_i$  by its projection  $y'_i$  on  $B_T$ . The new codevector  $y'_i$  is then closer to all  $x \in B_T$  than  $y_i$ , i.e.,  $\|x - y'_i\| < \|x - y_i\|$  if  $\|x\| \leq T$ ,  $\|y_i\| > T$ , and  $y'_i = Ty_i/\|y_i\|$ . Replacing all codevectors of  $q$  outside  $B_T$  by their projections, we obtain  $q'$  such that  $\|x - q'(x)\| \leq \|x - q(x)\|$  for all  $x \in B_T$ . Thus for any  $\nu \in \mathcal{P}(T)$ , we have  $D(\nu, q') \leq D(\nu, q)$ . Since  $\mu \in \mathcal{P}(T)$ , we also have  $\mu_n \in \mathcal{P}(T)$  a.s., and it follows that both

$q^*$  and  $q_n^*$  can be assumed to have all their codevectors inside  $B_T$ . Since both  $q^*$  and  $q_n^*$  were assumed to be nearest neighbor quantizers, we obtain  $q^*, q_n^* \in \mathcal{Q}_N(T)$ . This and bound (10) imply the lemma.  $\square$

The following result is the key to the finite sample bounds in this section.

**Theorem 3** *There is a constant  $C$ , depending only on  $d$ ,  $N$ , and  $T$ , such that for all  $n \geq 1$  and  $\mu \in \mathcal{P}(T)$ ,*

$$\mathbb{E} \left\{ \sup_{q \in \mathcal{Q}_N(T)} |D(\mu_n, q) - D(\mu, q)| \right\} \leq \frac{C}{\sqrt{n}}.$$

PROOF. There are several (not fundamentally different) ways to prove the bound of the theorem, resulting in somewhat different values for the constant  $C$ . These methods are all based on bounds on uniform deviations of empirical averages given in terms of combinatorial properties of certain classes of sets or functions. The proof given below is perhaps the simplest; it is based on a sharpened version of the basic Vapnik-Chervonenkis inequality. In what follows we will use notions and results of Vapnik-Chervonenkis theory given in Lugosi (2001) (see also Devroye, Györfi, and Lugosi (1996), Vapnik (1998), or Anthony and Bartlett (1999)).

For  $q \in \mathcal{Q}_N(T)$  define the distortion function

$$f_q(x) \triangleq \|x - q(x)\|^2.$$

Note that  $q \in \mathcal{Q}_N(T)$  implies  $0 \leq f_q(x) \leq 4T^2$  for all  $x \in B_T$ . Thus

$$D(\mu, q) = \mathbb{E} f_q(X) = \int_0^{4T^2} \mathbb{P}\{f_q(X) > u\} du$$

and

$$D(\mu_n, q) = \frac{1}{n} \sum_{k=1}^n f_q(X_k) = \int_0^{4T^2} \frac{1}{n} \sum_{k=1}^n I_{\{f_q(X_k) > u\}} du \quad \text{a.s.}$$

Hence

$$\begin{aligned}
& \sup_{q \in \mathcal{Q}_N(T)} |D(\mu_n, q) - D(\mu, q)| \\
&= \sup_{q \in \mathcal{Q}_N(T)} \left| \frac{1}{n} \sum_{k=1}^n f_q(X_k) - \mathbb{E} f_q(X) \right| \\
&= \sup_{q \in \mathcal{Q}_N(T)} \left| \int_0^{4T^2} \left( \frac{1}{n} \sum_{k=1}^n I_{\{f_q(X_k) > u\}} - \mathbb{P}\{f_q(X) > u\} \right) du \right| \\
&\leq 4T^2 \sup_{q \in \mathcal{Q}_N(T), u > 0} \left| \frac{1}{n} \sum_{k=1}^n I_{\{f_q(X_k) > u\}} - \mathbb{P}\{f_q(X) > u\} \right| \\
&= 4T^2 \sup_{A \in \mathcal{A}_N} |\mu_n(A) - \mu(A)| \quad \text{a.s.} \tag{11}
\end{aligned}$$

where  $\mathcal{A}_N$  is the family of sets in  $\mathbb{R}^d$  defined by

$$\mathcal{A}_N \triangleq \left\{ \{x : f_q(x) > u\} : q \in \mathcal{Q}_N(T), u > 0 \right\}.$$

Denoting by  $V(\mathcal{A}_N)$  the VC dimension of  $\mathcal{A}_N$ , the sharpened version of the Vapnik-Chervonenkis inequality in Section 4.5 of Lugosi (2001) gives

$$\mathbb{E} \left\{ \sup_{A \in \mathcal{A}_N} |\mu_n(A) - \mu(A)| \right\} \leq c \sqrt{\frac{V(\mathcal{A}_N)}{n}} \tag{12}$$

where  $c$  is a universal constant. Therefore

$$\mathbb{E} \left\{ \sup_{q \in \mathcal{Q}_N(T)} |D(\mu_n, q) - D(\mu, q)| \right\} \leq 4T^2 c \sqrt{\frac{V(\mathcal{A}_N)}{n}}. \tag{13}$$

Of course, the bound is meaningful only if  $V(\mathcal{A}_N) < \infty$ , which we show to be the case next.

For a nearest neighbor quantizer  $q$  with codevectors  $\{y_i\}_{i=1}^N$ ,  $f_q(x) > u$  if and only if  $\|x - y_i\| > \sqrt{u}$  for all  $i = 1, \dots, N$ . Thus each  $A \in \mathcal{A}_N$  is the intersection of the complements of  $N$  closed balls of equal radii in  $\mathbb{R}^d$ . Letting  $\mathcal{A}$  denote the collection of all complements of closed balls in  $\mathbb{R}^d$ , we therefore have

$$\mathcal{A}_N \subset \bar{\mathcal{A}}_N \triangleq \{A_1 \cap \dots \cap A_N : A_i \in \mathcal{A}, i = 1, \dots, N\}. \tag{14}$$

Note that  $\bar{\mathcal{A}}_1 = \mathcal{A}$ , and  $\bar{\mathcal{A}}_N = \{A \cap B : A \in \mathcal{A}, B \in \bar{\mathcal{A}}_{N-1}\}$  for  $N \geq 2$ . Thus Theorem 1.12(4) of Lugosi (2001) implies that for  $N \geq 2$  the  $m$ th shatter coefficient of  $\bar{\mathcal{A}}_N$  is upper bounded as

$$\mathbb{S}_{\bar{\mathcal{A}}_N}(m) \leq \mathbb{S}_{\mathcal{A}}(m) \mathbb{S}_{\bar{\mathcal{A}}_{N-1}}(m).$$

By induction on  $N$ , we obtain that for all  $N \geq 1$  and  $m \geq 1$ ,

$$\mathbb{S}_{\bar{\mathcal{A}}_N}(m) \leq \mathbb{S}_{\mathcal{A}}(m)^N. \quad (15)$$

Define  $\mathcal{B} = \{B^c : B \in \mathcal{A}\}$ . From Theorem 1.12(3) in Lugosi (2001), we have  $\mathbb{S}_{\mathcal{A}}(m) = \mathbb{S}_{\mathcal{B}}(m)$ . Hence (14) and (15) yield

$$\mathbb{S}_{\mathcal{A}_N}(m) \leq \mathbb{S}_{\bar{\mathcal{A}}_N}(m) \leq \mathbb{S}_{\mathcal{B}}(m)^N. \quad (16)$$

Since  $\mathcal{B}$  is the collection of all closed balls in  $\mathbb{R}^d$ , we have  $V(\mathcal{B}) = d + 1$  by the remark following Corollary 1.4 in Lugosi (2001). Thus a consequence of Sauer's lemma, Corollary 1.3 in Lugosi (2001), implies that for all  $m \geq d + 1$ ,

$$\mathbb{S}_{\mathcal{B}}(m) \leq \left( \frac{me}{d+1} \right)^{d+1}.$$

This and (16) imply that for all  $m \geq d + 1$ ,

$$\mathbb{S}_{\mathcal{A}_N}(m) \leq \left( \frac{me}{d+1} \right)^{N(d+1)}. \quad (17)$$

An upper bound to  $V(\mathcal{A}_N)$  can now be obtained by finding an  $m$  for which the right side is less than  $2^m$ . It is easy to check that if  $d \geq 2$ , then  $m = 4N(d+1) \ln(N(d+1))$  satisfies this requirement. Since for  $d = 1$  we obviously have  $V(\mathcal{A}_N) \leq 2N$ , we obtain that for all  $N, d \geq 1$ ,

$$V(\mathcal{A}_N) \leq 4N(d+1) \ln(N(d+1)).$$

This and the sharpened Vapnik-Chervonenkis inequality (13) imply the theorem with

$$C = 4T^2 c \sqrt{4N(d+1) \ln(N(d+1))}.$$

□

Combining Lemma 5 with the bound of Theorem 3, we obtain the main result of this section, a finite sample bound on the expected test distortion of the empirically optimal quantizer.

**Theorem 4** *There is a constant  $C_1$ , depending only on  $d$ ,  $N$ , and  $T$ , such that for all  $n \geq 1$  and  $\mu \in \mathcal{P}(T)$ ,*

$$\mathbb{E}D(\mu, q_n^*) - D_N^*(\mu) \leq \frac{C_1}{\sqrt{n}}.$$

We have seen in Section 3 that the training distortion of the empirically optimal quantizer is a strongly consistent, but optimistically biased estimate of the optimal distortion. Theorem 3 immediately provides an upper bound on the size of this bias.

**Theorem 5** For all  $n \geq 1$  and  $\mu \in \mathcal{P}(T)$ ,

$$D_N^*(\mu) - \mathbb{E}D(\mu_n, q_n^*) \leq \frac{C}{\sqrt{n}}.$$

where  $C$  is the constant in Theorem 3.

PROOF. As before, let  $q^*$  denote an optimal  $N$ -point quantizer for  $\mu$ . Then we have

$$\begin{aligned} D_N^*(\mu) - \mathbb{E}D(\mu_n, q_n^*) &= \mathbb{E}\left\{D(\mu, q^*) - D(\mu_n, q_n^*)\right\} \\ &\leq \mathbb{E}\left\{D(\mu, q_n^*) - D(\mu_n, q_n^*)\right\} \\ &\leq \mathbb{E}\left\{\sup_{q \in \mathcal{Q}_N(T)} |D(\mu_n, q) - D(\mu, q)|\right\} \end{aligned}$$

so the statement follows from Theorem 3.  $\square$

Combined with an appropriate concentration inequality (namely, the bounded difference inequality (see Section 1.3), Theorem 3 also provides a convergence rate in the consistency theorem for sources with bounded support.

**Theorem 6** For every  $\mu \in \mathcal{P}(T)$ , as  $n \rightarrow \infty$ ,

$$D(\mu, q_n^*) - D_N^*(\mu) = O\left(\sqrt{\frac{\ln n}{n}}\right) \quad a.s.$$

PROOF. Let

$$Y_n \triangleq \sup_{q \in \mathcal{Q}_N(T)} |D(\mu_n, q) - D(\mu, q)|.$$

From Lemma 5 and Theorem 3,

$$D(\mu, q_n^*) - D_N^*(\mu) \leq 2Y_n \leq 2(Y_n - \mathbb{E}Y_n) + O\left(\frac{1}{\sqrt{n}}\right). \quad (18)$$

We can view  $Y_n$  as a function  $Y_n = g(X_1, \dots, X_n)$  of the independent random vectors  $X_1, \dots, X_n$ . For any  $q \in \mathcal{Q}_N(T)$ , and any  $x_1, \dots, x_n \in B_T$ ,  $\hat{x}_1, \dots, \hat{x}_n \in B_T$  such that  $x_k = \hat{x}_k$  for all  $k \neq i$ ,

$$\begin{aligned} & \left| \frac{1}{n} \sum_{k=1}^n \|x_k - q(x_k)\|^2 - \frac{1}{n} \sum_{k=1}^n \|\hat{x}_k - q(\hat{x}_k)\|^2 \right| \\ &= \frac{1}{n} \left| \|x_i - q(x_i)\|^2 - \|\hat{x}_i - q(\hat{x}_i)\|^2 \right| \\ &\leq \frac{1}{n} 4T^2. \end{aligned}$$

It follows that  $g : (B_T)^n \rightarrow \mathbb{R}$  satisfies the assumptions of the bounded difference inequality (Theorem 1.8) with  $c_i = 4T^2/n$ , and we obtain for all  $t > 0$ ,

$$\mathbb{P}\{Y_n - \mathbb{E}Y_n \geq t\} \leq e^{-nt^2/(8T^4)}.$$

Choosing  $t = \sqrt{\hat{c} \ln n/n}$  with  $\hat{c} > 8T^4$ , the right side is summable in  $n$ , and so the Borel-Cantelli lemma implies  $Y_n - \mathbb{E}Y_n = O(\sqrt{\ln n/n})$  a.s. In view of (18) this proves the theorem.  $\square$

## Notes

Theorem 4 is due to Linder, Lugosi, and Zeger (1994). The constant can be improved by using covering numbers and metric entropy bounds from empirical process theory (see, e.g., Dudley (1978) or Pollard (1990)) instead of the Vapnik-Chervonenkis inequality; see Linder (2000). Related bounds are given in Linder, Lugosi, and Zeger (1997) for quantization of sources corrupted by noise, and for combined quantization and transmission over noisy channels. A generalization of Theorem 4 to dependent (mixing) training data is given in Zeevi (1998). Graf and Luschgy (1999) proved almost sure convergence rates for the training distortion. The sample behavior of the test distortion for a class of sources with smooth densities is given in Chou (1994). The dependence of the test distortion on the size of the training data was empirically investigated by Cosman *et al.* (1991) and Cohn, Riskin, and Ladner (1992) in the context of image coding.

## 5 Minimax lower bounds

We showed in the previous section that for source distributions with bounded support, the expected test and training distortions of an empirically optimal quantizer trained on  $n$  data samples are bounded as

$$D_N^*(\mu) \leq \mathbb{E}D(\mu, q_n^*) \leq D_N^*(\mu) + \frac{C_1}{\sqrt{n}} \quad (19)$$

and

$$D_N^*(\mu) - \frac{C}{\sqrt{n}} \leq \mathbb{E}D(\mu_n, q_n^*) \leq D_N^*(\mu). \quad (20)$$

In these bounds the positive constants  $C_1$  and  $C$  depend only on the dimension, the number of codevectors, and the diameter of the support.

Unfortunately, the proofs give no indication whether the  $O(1/\sqrt{n})$  terms can be tightened. More generally, we don't know if there exists a method, perhaps different from empirical error minimization, which provides an empirically designed quantizer with substantially smaller test distortion.

Let us examine the simple case of quantizers with  $N = 1$  codevector. In this case, as the centroid condition of Lemma 1 implies, the optimal quantizer  $q^*$  has a unique codepoint  $y_1 = \mathbb{E}X$  and its distortion is

$$D_1^*(\mu) = \mathbb{E}\|X - \mathbb{E}X\|^2.$$

The empirically optimal 1-point quantizer  $q_n^*$  is also unique with codepoint  $y_1^{(n)} = \frac{1}{n} \sum_{j=1}^n X_j$ , and its expected test distortion is easily seen to be

$$\begin{aligned} \mathbb{E}D(\mu, q_n^*) &= \mathbb{E}\left\|X - \frac{1}{n} \sum_{k=1}^n X_k\right\|^2 \\ &= \left(1 + \frac{1}{n}\right) \mathbb{E}\|X - \mathbb{E}X\|^2 \\ &= D_1^*(\mu) + \frac{D_1^*(\mu)}{n}. \end{aligned}$$

Similarly, the expected training distortion of  $q_n^*$  is

$$\begin{aligned} \mathbb{E}D(\mu_n, q_n^*) &= \mathbb{E}\left\{\frac{1}{n} \sum_{k=1}^n \left\|X_k - \frac{1}{n} \sum_{j=1}^n X_j\right\|^2\right\} \\ &= \left(1 - \frac{1}{n}\right) \mathbb{E}\|X - \mathbb{E}X\|^2 \\ &= D_1^*(\mu) - \frac{D_1^*(\mu)}{n}. \end{aligned}$$

Thus the convergence rate in both cases is  $O(1/n)$ , which is substantially faster than the  $O(1/\sqrt{n})$  rate in (19) and (20). However, perhaps surprisingly, the main results of this section show that the case  $N = 1$  is something of an anomaly, and for  $N \geq 3$ , the  $O(1/\sqrt{n})$  convergence rate of Theorems 4 and 5 cannot be improved upon in the minimax sense.

To formalize the problem, recall that an empirically designed  $N$ -point quantizer is a (measurable) function  $q_n : (\mathbb{R}^d)^{n+1} \rightarrow \mathbb{R}^d$  such that for any fixed  $x_1, \dots, x_n \in \mathbb{R}^d$ ,  $q_n(\cdot, x_1, \dots, x_n)$  is an  $N$ -point quantizer. Thus an empirically designed quantizer consists of a family of quantizers and an “algorithm” which chooses one of these quantizers for each value of the training data  $x_1, \dots, x_n$ .

We are interested in the *minimax distortion redundancy* defined by

$$\inf_{q_n} \sup_{\mu \in \mathcal{P}(T)} \mathbb{E}D(\mu, q_n) - D_N^*(\mu)$$

where the supremum is taken over all source distributions supported in  $B_T$ , and the infimum is over all empirically designed  $N$ -point quantizers. As the next theorem shows, for all  $N \geq 3$  the minimax distortion redundancy is lower bounded by a quantity proportional to  $1/\sqrt{n}$ . This means that no matter what method of empirical design we use, there always exists a “bad” source distribution in  $\mathcal{P}(T)$  such that the expected test distortion exceeds the optimal distortion by constant times  $1/\sqrt{n}$ .

**Theorem 7** *If  $N \geq 3$ , then for any empirically designed  $N$ -point quantizer  $q_n$  trained on  $n \geq n_0$  samples,*

$$\sup_{\mu \in \mathcal{P}(T)} \mathbb{E}D(\mu, q_n) - D_N^*(\mu) \geq \frac{C_2}{\sqrt{n}}$$

where the threshold  $n_0$  depends only on  $N$ , and  $C_2$  is a positive constant that depends only on  $d$ ,  $N$ , and  $T$ .

The idea behind the proof is best illustrated by the special case  $d = 1$ ,  $N = 3$ . Assume that  $\mu$  is concentrated on four points:  $-1$ ,  $-1 + \Delta$ ,  $1 - \Delta$ , and  $1$ , such that either  $\mu(-1) = \mu(-1 + \Delta) = (1 - \delta)/4$  and  $\mu(1 - \Delta) = \mu(1) = (1 + \delta)/4$ , or  $\mu(-1) = \mu(-1 + \Delta) = (1 + \delta)/4$  and  $\mu(1 - \Delta) = \mu(1) = (1 - \delta)/4$ . Then if  $\Delta$  is sufficiently small, the codepoints of the 3-point optimal quantizer are  $-1 + \Delta/2$ ,  $1 - \Delta$ ,  $1$  in the first case, and  $-1$ ,  $-1 + \Delta$ ,  $1 - \Delta/2$  in the second case. Therefore, an empirical quantizer should “learn” from the data which of the two distributions generates the data. This leads to a hypothesis testing problem whose error may be estimated by appropriate inequalities for the binomial distribution. Proper choice of the parameters  $\Delta$  and  $\delta$  yields the desired  $1/\sqrt{n}$ -type lower bound for the minimax expected distortion redundancy.

The proof of the general case is more complicated, but the basic idea is the same. A complete proof is given only for the special case; the main

steps of the proof of the general case are sketched in the Appendix at the end of the section.

PROOF OF THEOREM 7 (CASE  $d = 1$ ,  $N = 3$ ).

For simplicity we assume  $T = 1$  (i.e., we consider distributions supported in the interval  $[-1, 1]$ ); the result is generalized for arbitrary  $T > 0$  by straightforward scaling. Define

$$J(\mu, q_n) \triangleq \mathbb{E}D(\mu, q_n) - D_N^*(\mu)$$

and note that

$$\sup_{\mu \in \mathcal{P}(1)} J(\mu, q_n) \geq \sup_{\mu \in \mathcal{D}} J(\mu, q_n)$$

for any restricted class of distributions  $\mathcal{D}$  supported in  $[-1, 1]$ . In particular, we let  $\mathcal{D}$  contain two discrete distributions  $\nu_1$  and  $\nu_2$  concentrated on four points  $\{-1, -1 + \Delta, 1 - \Delta, 1\}$  with probabilities

$$\nu_1(-1) = \nu_1(-1 + \Delta) = \frac{1 - \delta}{4}, \quad \nu_1(1 - \Delta) = \nu_1(1) = \frac{1 + \delta}{4}$$

and

$$\nu_2(-1) = \nu_2(-1 + \Delta) = \frac{1 + \delta}{4}, \quad \nu_2(1 - \Delta) = \nu_2(1) = \frac{1 - \delta}{4}$$

where the parameters  $0 < \delta < 1$  and  $0 < \Delta \leq 1/2$  are to be specified later.

The optimal 3-point quantizer  $q^{(i)}$  for  $\nu_i$  is easy to find. Since  $q^{(i)}$  is a nearest neighbor quantizer, its partition cells are intervals, and optimality requires that each of these three intervals contain at least one of the four atoms of  $\nu_i$ . There are only three such partitions of the four points, and the centroid condition (Lemma 1) will uniquely determine the codepoints in each case. Thus there are only three possible candidates for an optimal quantizer for  $\nu_i$ .

It is easy to check that if  $0 < \delta < 1$  and  $0 < \Delta \leq 1/2$ , then the unique codebooks  $\mathcal{C}_i$  of the optimal quantizers  $q^{(i)}$ ,  $i = 1, 2$ , are given by

$$\mathcal{C}_1 = \{-1 + \Delta/2, 1 - \Delta, 1\}, \quad \mathcal{C}_2 = \{-1, -1 + \Delta, 1 - \Delta/2\}$$

with equal minimum distortions

$$D_3^*(\nu_1) = D_3^*(\nu_2) = \frac{\Delta^2}{8}(1 - \delta).$$

It is also easy to check that if  $\delta \leq 1/2$  and  $\Delta \leq 1/2$ , then for any 3-point quantizer  $q$  there is a  $q' \in \{q^{(1)}, q^{(2)}\}$ , such that

$$D(\nu_i, q') \leq D(\nu_i, q), \quad i = 1, 2.$$

(Clearly, we need only to check that under both  $\nu_1$  and  $\nu_2$ , the distortion of say  $q^{(1)}$  is less than that of any  $q$  that partitions the four points as  $\{-1\}, \{-1 + \Delta, 1 - \Delta\}$  and  $\{1\}$ ).

Let  $\mathcal{Q}^{(n)}$  denote the family of empirically designed quantizers  $q_n$  such that for every  $x_1, \dots, x_n$ , we have  $q_n(\cdot, x_1, \dots, x_n) \in \{q^{(1)}, q^{(2)}\}$ . Then by the preceding discussion,

$$\begin{aligned} \inf_{q_n} \sup_{\mu \in \mathcal{P}(1)} J(\mu, q_n) &\geq \inf_{q_n} \max_{\mu \in \{\nu_1, \nu_2\}} J(\mu, q_n) \\ &= \inf_{q_n \in \mathcal{Q}^{(n)}} \max_{\mu \in \{\nu_1, \nu_2\}} J(\mu, q_n). \end{aligned} \quad (21)$$

The maximum can be lower bounded by an average: If  $Z$  is a random variable with distribution  $\mathbb{P}\{Z = 1\} = \mathbb{P}\{Z = 2\} = 1/2$ , then for any  $q_n$ ,

$$\max_{\mu \in \{\nu_1, \nu_2\}} J(\mu, q_n) \geq \frac{1}{2} \sum_{i=1}^2 J(\nu_i, q_n) = \mathbb{E}J(\nu_Z, q_n). \quad (22)$$

Define

$$M \triangleq |\{k : x_k \in \{-1, -1 + \Delta\}, k = 1 \dots n\}|$$

i.e.,  $M$  is the number of training samples falling in  $\{-1, -1 + \Delta\}$ , and let  $Q_n^* \in \mathcal{Q}^{(n)}$  be the ‘‘maximum likelihood’’ quantizer defined by

$$Q_n^*(\cdot, x_1, \dots, x_n) = \begin{cases} q^{(1)} & \text{if } M < n/2 \\ q^{(2)} & \text{otherwise.} \end{cases}$$

The idea is that  $M < n/2$  indicates that the training data was drawn from  $\nu_1$ , in which case  $q^{(1)}$  is better than  $q^{(2)}$ . Next we show that this is indeed the optimal strategy, i.e.,

$$\inf_{q_n \in \mathcal{Q}^{(n)}} \mathbb{E}J(\nu_Z, q_n) = \mathbb{E}J(\nu_Z, Q_n^*). \quad (23)$$

To prove (23), note that

$$\begin{aligned} \mathbb{E}J(\nu_Z, q_n) &= \mathbb{E}\|Y - q_n(Y, Y_1, \dots, Y_n)\|^2 - \mathbb{E}D_N^*(\mu_Z) \\ &= \mathbb{E}\{\mathbb{E}(\|Y - q_n(Y, Y_1, \dots, Y_n)\|^2 | Y_1, \dots, Y_n)\} - \mathbb{E}D_N^*(\mu_Z) \end{aligned}$$

where, under the condition  $Z = i$ , the sequence  $Y, Y_1, \dots, Y_n$  is conditionally i.i.d. with  $Y \sim \nu_i$ . Thus for any given  $x_1, \dots, x_n$ , a  $q_n$  achieving the infimum on the left side of (23) must pick a quantizer that is optimal for the conditional distribution of  $Y$  given  $Y_i = x_i$ ,  $i = 1, \dots, n$ . Note that by conditional independence,

$$\begin{aligned} & \mathbb{P}\{Y = x | Y_1 = x_1, \dots, Y_n = x_n\} \\ &= \sum_{i=1}^2 \mathbb{P}\{Y = x | Y_1 = x_1, \dots, Y_n = x_n, Z = i\} \mathbb{P}\{Z = i | Y_1 = x_1, \dots, Y_n = x_n\} \\ &= \sum_{i=1}^2 \mathbb{P}\{Y = x | Z = i\} \mathbb{P}\{Z = i | Y_1 = x_1, \dots, Y_n = x_n\} \end{aligned} \quad (24)$$

for all  $x, x_1, \dots, x_n \in \{-1, -1 + \Delta, 1 - \Delta, 1\}$ . Since  $Z$  is uniformly distributed,

$$\mathbb{P}\{Z = 1 | Y_1 = x_1, \dots, Y_n = x_n\} > \mathbb{P}\{Z = 2 | Y_1 = x_1, \dots, Y_n = x_n\} \quad (25)$$

if and only if

$$\mathbb{P}\{Y_1 = x_1, \dots, Y_n = x_n | Z = 1\} > \mathbb{P}\{Y_1 = x_1, \dots, Y_n = x_n | Z = 2\}. \quad (26)$$

We have

$$\mathbb{P}\{Y_1 = x_1, \dots, Y_n = x_n | Z = i\} = \left(\frac{1 - \delta_i}{4}\right)^M \left(\frac{1 + \delta_i}{4}\right)^{n-M} \quad (27)$$

where  $\delta_1 = \delta$  and  $\delta_2 = -\delta$ . Thus (25) holds if and only if  $M < n/2$ . Introducing the notation

$$p_n(x) \triangleq \mathbb{P}\{Y = x | Y_1 = x_1, \dots, Y_n = x_n\}$$

and noting that  $\mathbb{P}\{Y = x | Z = i\} = \nu_i(x)$ , we obtain from (24)-(27) that

$$p_n(-1) = p_n(-1 + \Delta) < p_n(1 - \Delta) = p_n(1)$$

if  $M < n/2$ , and

$$p_n(-1) = p_n(-1 + \Delta) \geq p_n(1 - \Delta) = p_n(1)$$

otherwise. To avoid the asymmetry caused by tie-breaking if  $n$  is even, we assume here that  $n$  is odd (this assumption is clearly insignificant). It

follows that the optimal  $q_n$  must pick  $q^{(1)}$  if  $M < n/2$ , and  $q^{(2)}$  otherwise, i.e.,  $q_n = Q_n^*$  as claimed.

By symmetry we have

$$\mathbb{E}J(\nu_Z, Q_n^*) = J(\nu_1, Q_n^*). \quad (28)$$

With a slight abuse of notation, let now  $M$  denote the number of training samples, drawn independently under  $\nu_1$ , falling in  $\{-1, -1 + \Delta\}$ . Since  $D(\nu_1, q^{(1)}) = \Delta^2(1 - \delta)/8$  and  $D(\nu_1, q^{(2)}) = \Delta^2(1 + \delta)/8$ , it is easy to see that

$$\begin{aligned} J(\nu_1, Q_n^*) &= \frac{\Delta^2}{8}(1 - \delta)\mathbb{P}\{M < n/2\} + \frac{\Delta^2}{8}(1 + \delta)\mathbb{P}\{M \geq n/2\} - \frac{\Delta^2}{8}(1 - \delta) \\ &= \frac{\Delta^2}{4}\delta\mathbb{P}\{M \geq n/2\}. \end{aligned}$$

Note that  $M$  has binomial distribution with parameters  $n$  and  $p = (1 - \delta)/2$ . From (21), (22), (23), and (28) we conclude that for all  $\Delta, \delta \leq 1/2$ ,

$$\inf_{q_n} \sup_{\mu \in \mathcal{P}(1)} J(\mu, q_n) \geq \frac{\Delta^2}{4}\delta\mathbb{P}\{M \geq n/2\}. \quad (29)$$

The above binomial probability can be lower bounded via the standard method of normal approximation. However, it is more convenient to use a non-asymptotic inequality by Slud (1977) which states that for all  $np \leq k \leq n(1 - p)$ ,

$$\mathbb{P}\{M \geq k\} \geq \Phi\left(-\frac{k - np}{\sqrt{np(1 - p)}}\right)$$

where  $\Phi$  is the standard normal distribution function. Letting

$$\delta = \frac{1}{\sqrt{n}}$$

for  $n \geq 4$ , the choice  $k = \lceil n/2 \rceil$  satisfies the conditions of Slud's inequality, and we obtain

$$\begin{aligned} \mathbb{P}\{M \geq n/2\} &\geq \Phi\left(-\frac{n/2 + 1 - n(1 - 1/\sqrt{n})/2}{\sqrt{n(1 - 1/\sqrt{n})(1 + 1/\sqrt{n})/4}}\right) \\ &\geq \Phi(-2) \end{aligned}$$

where the second inequality holds for all  $n \geq 6$ . Combining this with (29) and setting  $\Delta = 1/2$ , we obtain for all  $n \geq 6$ ,

$$\inf_{q_n} \sup_{\mu \in \mathcal{P}(1)} J(\mu, q_n) \geq \frac{\Phi(-2)/16}{\sqrt{n}}.$$

If we consider  $\mathcal{P}(T)$  rather than  $\mathcal{P}(1)$ , then  $\Delta$  becomes  $T/2$ , and so the theorem holds with  $C_2 = T^2\Phi(-2)/16$  and  $n_0 = 6$ .  $\square$

The next result provides a similar lower bound on the bias of the training distortion of empirically optimal quantizers in estimating the optimal distortion. The theorem shows that there exist “bad” distributions for which the bias is on the order of  $1/\sqrt{n}$ . This lower bound matches, at least in order of magnitude, the upper bound of Theorem 5. Again, we prove only the special case  $d = 1$ ,  $N = 3$ . In the general case the argument is similar, but the details are more involved.

**Theorem 8** *If  $N \geq 3$ , then for the empirically optimal  $N$ -point quantizer  $q_n^*$  trained on  $n \geq n_0$  samples,*

$$\sup_{\mu \in \mathcal{P}(T)} D_N^*(\mu) - \mathbb{E}D(\mu_n, q_n^*) \geq \frac{C_3}{\sqrt{n}}$$

where  $n_0$  depends only on  $N$ , and  $C_3$  is a positive constant that depends only on  $d$ ,  $N$ , and  $T$ .

PROOF (case  $d = 1$ ,  $N = 3$ ). Suppose  $T = 1$  and let the discrete random variable  $X \sim \mu$  be uniformly distributed on  $\{-1, -1 + \Delta, 1 - \Delta, 1\}$ . It is easy to see that if  $0 < \Delta < 2/3$ , then there are exactly two optimal 3-point quantizers for  $X$ ; one with codebook

$$\mathcal{C}_1 = \{-1 + \Delta/2, 1 - \Delta, 1\}$$

the other with codebook

$$\mathcal{C}_2 = \{-1, -1 + \Delta, 1 - \Delta/2\}.$$

Let  $q^*$  be the nearest neighbor quantizer with codebook  $\mathcal{C}_1$ . The training data  $X_1, \dots, X_n$  consist of i.i.d. copies of  $X$ . Let  $M$  be the number of  $X_k$ ,  $k = 1, \dots, n$  such that  $X_k \in \{-1, -1 + \Delta\}$ . Then

$$\begin{aligned} D_3^*(\mu) &= \mathbb{E}(X - q^*(X))^2 = \mathbb{E} \left\{ \frac{1}{n} \sum_{k=1}^n (X_k - q^*(X_k))^2 \right\} \\ &= \frac{1}{n} \frac{\Delta^2}{4} \mathbb{E}M \end{aligned} \tag{30}$$

where the second equality holds because  $(q^*(X_k) - X_k)^2 = \Delta^2/4$  if  $X_k$  takes value in  $\{-1, -1 + \Delta\}$  and  $(q^*(X_k) - X_k)^2 = 0$  otherwise.

Let the training set dependent nearest neighbor quantizer  $q_n$  be the following: the codebook of  $q_n$  is  $\mathcal{C}_1$  if  $M < n/2$  and the codebook is  $\mathcal{C}_2$  if  $M \geq n/2$ . Then the expected training distortion of  $q_n$  is

$$\begin{aligned} \mathbb{E}D(\mu_n, q_n) &= \mathbb{E} \left\{ \frac{1}{n} \sum_{k=1}^n (X_k - q_n(X_k))^2 \right\} \\ &= \frac{1}{n} \frac{\Delta^2}{4} \mathbb{E} \min(M, n - M). \end{aligned} \quad (31)$$

Since the empirically optimal 3-point quantizer  $q_n^*$  minimizes the training distortion, we have

$$\mathbb{E}D(\mu_n, q_n) \geq \mathbb{E}D(\mu_n, q_n^*).$$

Hence (30) and (31) imply

$$D_3^*(\mu) - \mathbb{E}D(\mu_n, q_n^*) \geq \frac{1}{n} \frac{\Delta^2}{4} \mathbb{E}\{M - \min(M, n - M)\}.$$

Since  $M$  is binomial with parameters  $(n, 1/2)$ , its distribution is symmetric about  $n/2$ . Thus

$$\begin{aligned} \mathbb{E}\{M - \min(M, n - M)\} &= \mathbb{E}\{(2M - n)^+\} \\ &= \mathbb{E}\left\{2\left(M - \frac{n}{2}\right)^+\right\} \\ &= \mathbb{E}\left|M - \frac{n}{2}\right|. \end{aligned}$$

Now we can apply a special case of Khintchine's inequality (see Szarek (1976)) stating that if  $M$  has binomial distribution with parameters  $(n, 1/2)$ , then

$$\mathbb{E}\left|M - \frac{n}{2}\right| \geq \sqrt{\frac{n}{8}}.$$

We conclude that for all  $\Delta < 2/3$  and  $n \geq 1$ ,

$$\sup_{\mu \in \mathcal{P}(1)} D_3^*(\mu) - \mathbb{E}D(\mu, q_n^*) \geq \frac{\Delta^2}{8\sqrt{2}} \frac{1}{\sqrt{n}}.$$

If we consider distributions in  $\mathcal{P}(T)$  rather than in  $\mathcal{P}(1)$ , then the constraint on  $\Delta$  becomes  $\Delta < 2T/3$ , and we obtain

$$\sup_{\mu \in \mathcal{P}(T)} D_3^*(\mu) - \mathbb{E}D(\mu, q_n^*) \geq \frac{T^2}{18\sqrt{2}} \frac{1}{\sqrt{n}}.$$

□

## Notes

Theorem 7 is from Bartlett, Linder, and Lugosi (1998). A related lower bound in an information-theoretic setting was proved by Merhav and Ziv (1997). Theorem 8 is based on Linder (2000).

## Appendix

SKETCH OF PROOF OF THEOREM 7 (CASE  $d \geq 1$ ,  $N \geq 3$ ).

STEP 1: Define the restricted class of distributions  $\mathcal{D}$  as follows: each member of  $\mathcal{D}$  is concentrated on the set of  $2m = 4N/3$  fixed points  $\{z_i, z_i + w : i = 1, \dots, m\}$ , where  $w = (\Delta, 0, 0, \dots, 0)$  is a fixed  $d$ -vector, and  $\Delta$  is a small positive number to be determined later. The positions of  $z_1, \dots, z_m \in B_T$  satisfy the property that the distance between any two of them is greater than  $3\Delta$ . For the sake of simplicity, we assume that  $N$  is divisible by 3. (This assumption is clearly insignificant.) For  $0 < \delta \leq 1/2$  and  $i = 1, \dots, m$ , set

$$\nu(z_i) = \nu(z_i + w) = \begin{cases} \text{either} & \frac{1 - \delta}{2m} \\ \text{or} & \frac{1 + \delta}{2m} \end{cases}$$

such that exactly half of the pairs  $\{z_i, z_i + w\}$  have mass  $(1 - \delta)/m$ , and the other half of the pairs have mass  $(1 + \delta)/m$ , so that the total mass adds up to one. Let  $\mathcal{D}$  contain all such distributions. The cardinality of  $\mathcal{D}$  is  $K = \binom{m}{m/2}$ . Denote the members of  $\mathcal{D}$  by  $\nu_1, \dots, \nu_K$ .

STEP 2: Let  $\mathcal{Q}$  denote the collection of  $N$ -point quantizers  $q$  such that for  $m/2$  values of  $i \in \{1, \dots, m\}$ ,  $q$  has codepoints at both  $z_i$  and  $z_i + w$ , and for the remaining  $m/2$  values of  $i$ ,  $q$  has a single codepoint at  $z_i + w/2$ . Then for each  $\nu_i$  the unique optimal  $N$ -point quantizer is in  $\mathcal{Q}$ . It is easy to see that for all  $i$ ,

$$D_N^*(\nu_i) = \min_{q \in \mathcal{Q}} D(\nu_i, q) = \frac{\Delta^2}{8}(1 - \delta).$$

STEP 3: One can show that for any  $N$ -point quantizer  $q$  there exists a  $q' \in \mathcal{Q}$  such that, for all  $\nu$  in  $\mathcal{D}$ ,  $D(\nu, q') \leq D(\nu, q)$ . Thus if  $\mathcal{Q}^{(n)}$  denotes the family of empirically designed quantizers  $q_n$  such that for every fixed

$x_1, \dots, x_n$ , we have  $q_n(\cdot, x_1, \dots, x_n) \in \mathcal{Q}$ , then

$$\begin{aligned} \inf_{q_n} \sup_{\mu \in \mathcal{P}(T)} J(\mu, q_n) &\geq \inf_{q_n} \max_{\mu \in \mathcal{D}} J(\mu, q_n) \\ &= \inf_{q_n \in \mathcal{Q}^{(n)}} \max_{\mu \in \mathcal{D}} J(\mu, q_n). \end{aligned}$$

STEP 4: Let  $Z$  be a random variable which is uniformly distributed on the set of integers  $\{1, 2, \dots, K\}$ . Then, for any  $q_n$ , we have

$$\max_{\nu \in \mathcal{D}} J(\nu, q_n) \geq \frac{1}{K} \sum_{i=1}^K J(\nu_i, q_n) = \mathbb{E}J(\nu_Z, q_n).$$

Let  $Q_n^*$  denote the “maximum-likelihood” quantizer from  $\mathcal{Q}$ , that is, if  $M_i$  denotes the number of training samples falling in  $\{z_i, z_i + w\}$ , then  $Q_n^*$  has a codepoint at both  $z_i$  and  $z_i + w$  if the corresponding  $M_i$  is one of the  $m/2$  largest values. For the other  $i$ ’s (i.e., those with the  $m/2$  smallest  $M_i$ ’s)  $Q_n^*$  has a codepoint at  $z_i + w/2$ . Then it can be proved that

$$\inf_{q_n \in \mathcal{Q}^{(n)}} \mathbb{E}J(\nu_Z, q_n) = \mathbb{E}J(\nu_Z, Q_n^*).$$

STEP 5: By symmetry

$$\mathbb{E}J(\nu_Z, Q_n^*) = J(\nu_1, Q_n^*).$$

Under  $\nu_1$ , the vector of random integers  $(M_1, \dots, M_m)$  is multinomially distributed with parameters  $(n, p_1, \dots, p_m)$ , where  $p_1 = p_2 = \dots = p_{m/2} = (1 - \delta)/m$ , and  $p_{m/2+1} = \dots = p_m = (1 + \delta)/m$ . Let  $M_{\sigma(1)}, \dots, M_{\sigma(m)}$  be a reordering of the  $M_i$ ’s such that  $M_{\sigma(1)} \leq M_{\sigma(2)} \leq \dots \leq M_{\sigma(m)}$ . (In case of equal values, break ties using a random ordering.) Let  $P_j$ ,  $j = 1, \dots, m/2$  be the probability of the event that among  $M_{\sigma(1)}, \dots, M_{\sigma(m/2)}$  there are exactly  $j$  of the  $M_i$ ’s with  $i \geq m/2$  (i.e., the “maximum likelihood” estimate makes  $j$  mistakes). Then it is easy to see that

$$J(\nu_1, Q_n^*) = \frac{\Delta^2 \delta}{2m} \sum_{j=1}^{m/2} j P_j$$

since one “mistake” increases the distortion by  $\Delta^2 \delta / (2m)$ .

STEP 6: Using properties of the multinomial distribution, the above sum can be lower bounded in terms of the binomial probability  $\mathbb{P}\{M_1 > n/m\}$ .

Using normal approximation to binomials and choosing  $\delta = \sqrt{m/n}$ , we obtain for all  $n \geq 8m/\Phi(-2)^2$ ,

$$J(\nu_1, Q_n^*) \geq \frac{\Delta^2 \Phi(-2)^4}{512} \sqrt{\frac{m}{n}}.$$

STEP 7: It remains to choose  $\Delta$  as large as possible such that  $m$  pairs of points  $\{z_i, z_i + w\}$  can be placed in  $B_T$  so that the distance between any two of the  $z_i$ 's is at least  $3\Delta$ . A standard argument relating packings and coverings shows that  $\Delta = T/(4m^{1/d})$  satisfies this property, and so we obtain for all  $n \geq n_0 = 16N/(3\Phi(-2)^2)$ ,

$$\inf_{q_n} \sup_{\mu \in \mathcal{P}(T)} J(\mu, q_n) \geq \frac{C_2}{\sqrt{n}}$$

where  $C_2 = T^2 \Phi(-2)^4 2^{-13} \sqrt{(2N/3)^{1-4/d}}$ . □

## 6 Fundamentals of variable-rate quantization

In fixed-rate quantization, the possible outputs of a quantizer are represented by distinct binary strings of equal length. One can make a quantizer more efficient by using a variable-length representation. The idea is to assign shorter binary strings to quantizer outputs that occur more frequently (i.e., have higher probability), and longer binary strings to outputs that occur less frequently (i.e., have lower probability). This way the average rate, defined as the expected number of bits per quantizer output, can be substantially reduced. In this case, the number of codevectors no longer determines the rate, and in fact the average rate can be finite even if there are infinitely many codevectors.

Formally, a *variable-rate vector quantizer*  $q$  is described by an *encoder*  $\alpha : \mathbb{R}^d \rightarrow \mathcal{I}$ , where  $\mathcal{I}$  is a countable index set, a *decoder*  $\beta : \mathcal{I} \rightarrow \mathbb{R}^d$ , and an *index coder*  $\psi : \mathcal{I} \rightarrow \{0, 1\}^*$ , where  $\{0, 1\}^*$  denotes the collection of all finite-length binary strings. If  $\mathcal{I}$  is finite with  $N$  elements, without loss of generality we always take  $\mathcal{I} = \{1, \dots, N\}$ ; otherwise  $\mathcal{I}$  is taken to be the set of all positive integers.

In variable-rate quantization, an input  $x \in \mathbb{R}^d$  is encoded into an index  $i = \alpha(x)$ , which is represented by the binary string  $\psi(i)$  for purposes of storage or transmission. We require that  $\psi$  be invertible (in fact, as explained below, we require more) and so  $i$  can be recovered from  $\psi(i)$ , and the decoder

can output  $\beta(i)$ . Thus  $q$  maps any point  $x \in \mathbb{R}^d$  into one of the codevectors in the codebook  $\{\beta(i); i \in \mathcal{I}\}$ , via the rule

$$q(x) = \beta(\alpha(x)).$$

Letting  $S_i = \{x : \alpha(x) = i\}$  and  $y_i = \beta(i)$  for all  $i \in \mathcal{I}$ , we have, just as in the fixed-rate case,

$$q(x) = y_i \quad \text{if and only if} \quad x \in S_i.$$

As before, we assume that the  $d$ -dimensional random vector  $X \sim \mu$  has finite second moment  $\mathbb{E}\|X\|^2 < \infty$ , and define the distortion of  $q$  in the usual way:

$$D(\mu, q) = \mathbb{E}\|X - q(X)\|^2.$$

We require that the index coder  $\psi$  have the *prefix-free* property: If  $i \neq j$ , then the string  $\psi(i)$  is not a prefix of the string  $\psi(j)$ . In particular, the prefix-free property implies that the binary *codewords*  $\psi(i)$ ,  $i \in \mathcal{I}$  are all distinct, so  $\psi$  is invertible. More importantly, the prefix-free property also makes sure that if  $q$  is successively applied to a sequence of source outputs  $x_1, x_2, \dots, x_k$  of arbitrary length  $k$ , then from the binary string

$$\psi(\alpha(x_1))\psi(\alpha(x_2)) \dots \psi(\alpha(x_k))$$

obtained by concatenating the codewords  $\psi(\alpha(x_j))$ ,  $j = 1, \dots, k$ , one can uniquely recover the sequence of indices  $\alpha(x_1), \alpha(x_2), \dots, \alpha(x_k)$  and thus the quantizer outputs  $q(x_1), q(x_2), \dots, q(x_k)$ .

The *length function*  $\ell : \mathcal{I} \rightarrow \{0, 1, 2, \dots\}$  associates with each index  $i$  the length of the corresponding codeword  $\psi(i)$ , i.e.,  $\ell(i) = \text{length}(\psi(i))$ . The rate of the variable-rate quantizer  $q$  is defined as the expected codeword length:

$$r(\mu, q) \triangleq \mathbb{E} \ell(\alpha(X)) = \sum_{i \in \mathcal{I}} \ell(i) \mathbb{P}\{q(X) = y_i\}.$$

The following fundamental lemma gives a characterization of the set of codelengths for index coders that have the prefix-free property. The proof can be found, for example, in Cover and Thomas (1991).

**Lemma 6 (KRAFT'S INEQUALITY)** *If the binary codewords  $\psi(i)$ ,  $i \in \mathcal{I}$  have the prefix-free property, then their lengths  $\ell(i)$ ,  $i \in \mathcal{I}$  must satisfy the inequality*

$$\sum_{i \in \mathcal{I}} 2^{-\ell(i)} \leq 1.$$

Conversely, if the nonnegative integers  $\ell(i)$ ,  $i \in \mathcal{I}$ , satisfy this inequality, then there exists a set of codewords with these lengths that has the prefix-free property.

A length function  $\ell$  is called *admissible* if it satisfies Kraft's inequality. Note that the distortion of  $q$  depends only on  $\alpha$  and  $\beta$ , and that the rate of  $q$  depends on  $\psi$  only through  $\ell$ . By Kraft's inequality, a prefix-free index coder  $\psi$  exists for a given length function  $\ell$  if and only if  $\ell$  is admissible. Thus for our purposes, it is enough to specify a quantizer  $q$  by its encoder  $\alpha$ , decoder  $\beta$ , and admissible length function  $\ell$ . In this case, we write  $q \equiv (\alpha, \beta, \ell)$ .

We are interested in the minimum distortion achievable for a given rate  $R \geq 0$ :

$$\delta_R^*(\mu) \triangleq \inf_{q:r(\mu,q) \leq R} D(\mu, q).$$

To give some insight into the advantage of variable-rate quantization, assume  $\log_2 N$  is an integer and  $q^*$  is an optimal  $N$ -level fixed-rate quantizer. Then the lengths  $\ell(i) = \log_2 N$ ,  $i = 1, \dots, N$ , are clearly admissible, so we can view  $q^*$  as a special variable-rate quantizer with (constant) length function  $\ell$  and rate  $r(\mu, q^*) = \log_2 N$ . Hence,

$$D(\mu, q^*) \geq \inf_{q:r(\mu,q) \leq \log_2 N} D(\mu, q)$$

i.e., variable-rate quantizers always perform at least as well as fixed-rate ones.

To assess the advantage quantitatively, suppose  $\{S_i\}_{i=1}^N$  is the partition of  $q^*$  and denote  $p_i = \mathbb{P}\{X \in S_i\}$ . The *entropy* of the discrete random variable  $q^*(X)$  is the nonnegative quantity

$$H(q^*(X)) \triangleq - \sum_{i=1}^N p_i \log_2 p_i.$$

(Here we use the convention that  $0 \log_2 0 = 0$ .) Using the inequality  $\log_2 t \leq (t-1) \log_2 e$ , valid for all  $t > 0$ , we see that

$$\begin{aligned} H(q^*(X)) - \log_2 N &= \sum_{i=1}^N p_i \log_2 \frac{1/N}{p_i} \\ &\leq \log_2 e \sum_{i:p_i > 0} p_i \left( \frac{1}{N p_i} - 1 \right) \\ &\leq 0 \end{aligned}$$

that is,

$$H(q^*(X)) \leq \log_2 N. \quad (32)$$

It is easy to see that the inequality is strict unless  $p_i = 1/N$  for all  $i$ . It can be shown in a similar manner that for any admissible  $\ell$ ,

$$\sum_{i \in \mathcal{I}} \ell(i) p_i \geq H(q^*(X)).$$

Hence the entropy of a variable-rate quantizer is an absolute lower bound on its rate. This lower bound can be approached by the *Shannon-Fano codelengths*

$$\ell^*(i) \triangleq \lceil -\log_2 p_i \rceil, \quad i = 1, \dots, N$$

( $\ell^*$  is admissible since  $\sum_{i=1}^N 2^{-\lceil -\log_2 p_i \rceil} \leq \sum_{i=1}^N 2^{\log_2 p_i} = 1$ ). Then  $\ell^*$  achieves the lower bound within one bit since

$$\sum_{i=1}^N p_i \ell^*(i) \leq \sum_{i=1}^N p_i (-\log_2 p_i + 1) = H(q^*(X)) + 1. \quad (33)$$

When the probabilities  $p_i$ ,  $i = 1, \dots, N$ , are highly nonuniform, the entropy  $H(q^*(X))$  can be much less than  $\log_2 N$ . In this case, as (32) and (33) show, the rate of the variable-rate quantizer obtained from  $q^*$  using the Shannon-Fano codelengths can be significantly less than the rate  $R(q^*) = \log_2 N$  of the original fixed-rate quantizer.

The discussion above illustrates how the performance of an optimal fixed-rate quantizer can be improved by appropriate variable-rate encoding. Note, however, that even when  $q^*$  is equipped with a length function  $\ell$  that minimizes the expected codeword length, the resulting quantizer is not necessarily an optimal variable-rate quantizer. In general, optimality issues in variable-rate quantization are more difficult than in the fixed-rate case.

## Notes

Fundamentals of the theory of lossless coding are given in Cover and Thomas (1991). Gersho and Gray (1992) and Sayood (2000) discuss several methods of variable-length lossless coding used in lossy data compression. To facilitate analyses, it is customary in the quantization literature to approximate the average rate of a variable-rate quantizer by the entropy of the quantizer output. In this “entropy-constrained” setting the optimal quantizer performance is still hard to determine analytically (nontrivial examples are known only in the scalar  $d = 1$  case; see Berger (1972) and György and Linder (2000)), but this approach makes it possible to find approximations

to the optimal performance that become tight as the rate increases; see, e.g., Zador (1966) and (1982), Gish and Pierce (1968), and Gray, Linder, and Li (2001). For a survey of results in this area, see Gray and Neuhoff (1998).

## 7 The Lagrangian formulation

At the core of our results on learning fixed-rate quantizers from empirical data was the observation that we could restrict attention to the parametric class of  $N$ -point nearest neighbor quantizers instead of having to deal with the much larger class of all  $N$ -point quantizers. Unfortunately, for  $d \geq 2$  there is very little known concerning the structure of variable-rate quantizers achieving minimum distortion  $\delta_R^*(\mu)$  under a rate constraint  $R$ ; nor is it known whether an optimal variable-rate quantizer always exists.

In this section we recast the problem of optimal distortion-rate tradeoff for variable-rate quantization in a Lagrangian formulation that will resolve most of these difficulties. For a variable-rate quantizer  $q \equiv (\alpha, \beta, \ell)$ , and for  $\lambda > 0$ , define the *Lagrangian distortion* by

$$\Delta_\lambda(\mu, q) \triangleq D(\mu, q) + \lambda r(\mu, q) = \mathbb{E}\{\|X - q(X)\|^2 + \lambda \ell(\alpha(X))\}$$

and the optimal Lagrangian performance by

$$\Delta_\lambda^*(\mu) \triangleq \inf_q \Delta_\lambda(\mu, q)$$

where the infimum is taken over all variable-rate quantizers.

To see the connection between the original and the Lagrangian formulation of optimal variable-rate quantization, suppose that  $q_\lambda^*$  achieves  $\Delta_\lambda^*(\mu)$ , i.e.,

$$\Delta_\lambda(\mu, q_\lambda^*) = \inf_q D(\mu, q) + \lambda r(\mu, q).$$

Consider any quantizer  $q'$  with rate  $r(\mu, q') \leq r(\mu, q_\lambda^*)$ . Since  $D(\mu, q') + \lambda r(\mu, q') \geq D(\mu, q_\lambda^*) + \lambda r(\mu, q_\lambda^*)$ , we have

$$\begin{aligned} D(\mu, q') &\geq D(\mu, q_\lambda^*) + \lambda(r(\mu, q_\lambda^*) - r(\mu, q')) \\ &\geq D(\mu, q_\lambda^*). \end{aligned}$$

Thus  $q_\lambda^*$  is an optimal variable-rate quantizer for the rate constraint  $R = r(\mu, q_\lambda^*)$ , i.e.,

$$D(\mu, q_\lambda^*) = \inf_{q: r(\mu, q) \leq R} D(\mu, q), \quad r(\mu, q_\lambda^*) \leq R.$$

Unfortunately, the converse statement does not hold in general: For a given  $R$  there may not exist  $\lambda > 0$  such that  $q_\lambda^*$  achieves  $\delta_R^*(\mu)$ . In other words, we may not be able to find an optimal quantizer for an arbitrary rate constraint  $R$  by minimizing  $\Delta_\lambda(\mu, q)$  for some value of  $\lambda$ .

One can characterize the rates for which an optimal variable-rate quantizer can be obtained by the Lagrangian design by considering the *convex hull* of  $\delta_R^*(\mu)$ , defined as the largest convex function  $\hat{\delta}_R^*(\mu)$ ,  $R \geq 0$ , which is majorized by  $\delta_R^*(\mu)$  (see Rockafellar (1970)). One can show that for a given rate  $R$ ,  $\delta_R^*(\mu)$  is achievable by  $q_\lambda^*$  for some  $\lambda > 0$  if and only if  $\delta_R^*(\mu)$  coincides with its convex hull at this rate, i.e.,  $\delta_R^*(\mu) = \hat{\delta}_R^*(\mu)$ . (Here the Lagrange multiplier  $\lambda$  is the slope of the line supporting the convex hull.) Thus by minimizing  $\Delta_\lambda(\mu, q)$  for all values of  $\lambda$ , one can obtain all variable-rate quantizers that achieve the convex hull of  $\delta_R^*(\mu)$ . For values of  $R$  such that  $\hat{\delta}_R^*(\mu)$  is strictly less than  $\delta_R^*(\mu)$  (such  $R$  exist if and only if  $\delta_R^*(\mu)$  is not convex), optimal variable-rate quantizers cannot be obtained by the Lagrangian method. However, this is not a serious limitation in practical applications since any rate and distortion pair  $(R, \hat{\delta}_R^*(\mu))$  on the convex hull can be achieved by “timesharing” between two quantizers that achieve the convex hull, i.e., two quantizers that can be obtained by Lagrangian minimization.

The Lagrangian formulation yields a set of useful necessary conditions for quantizer optimality. The following result is the variable-rate counterpart of Lemma 1.

**Lemma 7** *Suppose  $q \equiv (\alpha, \beta, \ell)$  is an arbitrary variable-rate quantizer. Then in each of the following three cases the variable-rate quantizer  $q'$  defined there satisfies*

$$\Delta_\lambda(\mu, q') \leq \Delta_\lambda(\mu, q).$$

(a)  $q' \equiv (\alpha', \beta, \ell)$ , where the encoder  $\alpha'$  is defined by

$$\alpha'(x) = \arg \min_{i \in \mathcal{I}} (\|x - \beta(i)\|^2 + \lambda \ell(i)), \quad x \in \mathbb{R}^d$$

(ties are broken arbitrarily).

(b)  $q' \equiv (\alpha, \beta', \ell)$ , where the decoder  $\beta'$  is defined by

$$\beta'(i) = \arg \min_{y \in \mathbb{R}^d} \mathbb{E}[\|X - y\|^2 | \alpha(X) = i] = \mathbb{E}[X | \alpha(X) = i], \quad i \in \mathcal{I}.$$

(c)  $q' \equiv (\alpha, \beta, \ell')$ , where the codelength function  $\ell'$  minimizes

$$\sum_{i \in \mathcal{I}} \ell(i) \mathbb{P}\{\alpha(X) = i\}$$

over all admissible codelengths  $\ell$ .

PROOF. To prove (a), let  $S_i = \{x : \alpha(x) = i\}$ ,  $i \in \mathcal{I}$  denote the partition cells of  $q$ . Note that the equation defining  $\alpha'$  is equivalent to

$$\|x - \beta(\alpha'(x))\|^2 + \lambda \ell(\alpha'(x)) = \min_{i \in \mathcal{I}} (\|x - \beta(i)\|^2 + \lambda \ell(i)). \quad (34)$$

To see that the minimum (and so  $\alpha'$ ) is well defined for all  $x$  even when  $\mathcal{I}$  is not finite, note that in this case the admissibility of  $\ell$  implies that  $\lim_{i \rightarrow \infty} \ell(i) = \infty$ . Thus for each  $x$  it suffices to take the minimum over a finite subset of  $\mathcal{I}$ . Hence (34) is always well defined, and we obtain

$$\begin{aligned} \Delta_\lambda(\mu, q) &= \sum_{j \in \mathcal{I}} \int_{S_j} (\|x - \beta(j)\|^2 + \lambda \ell(j)) \mu(dx) \\ &\geq \sum_{j \in \mathcal{I}} \int_{S_j} \min_{i \in \mathcal{I}} (\|x - \beta(i)\|^2 + \lambda \ell(i)) \mu(dx) \\ &= \Delta_\lambda(\mu, q'). \end{aligned}$$

To prove (b), notice that the choice of the decoder only affects the term  $\|x - q(x)\|^2$  in the Lagrangian expression. Therefore (b) follows directly from the centroid condition of Lemma 1.

Finally, write

$$\Delta_\lambda(\mu, q) = \mathbb{E} \|X - \beta(\alpha(X))\|^2 + \lambda \sum_{i \in \mathcal{I}} \ell(i) p_i$$

where  $p_i = \mathbb{P}\{\alpha(X) = i\}$ . Now (c) follows since for  $\alpha$  and  $\beta$  fixed, the admissible length function that minimizes  $\sum_{i \in \mathcal{I}} \ell(i) p_i$  is the one that minimizes the overall Lagrangian distortion.  $\square$

**Remarks** (i) Lemma 7 (a) is analogous to the nearest neighbor condition of fixed-rate quantization. An optimal  $\alpha'$  for a given  $\beta$  and  $\ell$  is called a *modified nearest neighbor encoder*.

(ii) For a finite index set  $\mathcal{I}$ , optimal codelengths in part (c) of the lemma can be obtained, for example, as the codelengths of the binary Huffman code for the probabilities  $p_i = \mathbb{P}\{\alpha(X) = i\}$ ,  $i \in \mathcal{I}$  (see, e.g., Cover and

Thomas (1991)). For infinite index sets the existence of an optimal prefix-free code and an associated optimal admissible length function is shown in Linder, Tarokh, and Zeger (1997) under the condition that the entropy  $-\sum_{i \in \mathcal{I}} p_i \log_2 p_i$  is finite.

The next result shows the existence of quantizers minimizing the Lagrangian distortion. The proof, which we omit here, relies on the optimality criteria of Lemma 7, but is somewhat more involved than the proof of the analogous Theorem 1. The difficulty is posed by the fact that optimal variable-rate quantizers can have an infinite number of codevectors. Of course, if a quantizer minimizes the Lagrangian distortion, it can be assumed to have a modified nearest neighbor encoder by Lemma 7.

**Theorem 9** *For any  $\mu$  with finite second moment and  $\lambda > 0$  there is a variable-rate quantizer  $q_\lambda^*$  with a modified nearest neighbor encoder such that*

$$\Delta_\lambda(\mu, q_\lambda^*) = \Delta_\lambda^*(\mu).$$

**Remark** The necessary conditions for optimality in Lemma 7 suggest an iterative algorithm for designing variable-rate quantizers in a manner analogous to the fixed-rate case. Let  $\Delta_\lambda(\alpha, \beta, \ell)$  stand for  $\Delta_\lambda(\mu, q)$  if  $q \equiv (\alpha, \beta, \ell)$ . Start with an arbitrary quantizer  $q_0 \equiv (\alpha^{(0)}, \beta^{(0)}, \ell^{(0)})$  with a finite index set. In the  $m$ th iteration ( $m = 1, 2, \dots$ ), first choose  $\alpha^{(m)}$  to minimize  $\Delta_\lambda(\alpha, \beta^{(m-1)}, \ell^{(m-1)})$  for fixed  $\beta^{(m-1)}$  and  $\ell^{(m-1)}$ , then choose  $\beta^{(m)}$  to minimize  $\Delta_\lambda(\alpha^{(m)}, \beta, \ell^{(m-1)})$  for fixed  $\alpha^{(m)}$  and  $\ell^{(m-1)}$ , and then choose an admissible  $\ell^{(m)}$  to minimize  $\Delta_\lambda(\alpha^{(m)}, \beta^{(m)}, \ell)$  for fixed  $\alpha^{(m)}$  and  $\beta^{(m)}$ . Since the Lagrangian distortion is decreasing (or at least not increasing) in each step, setting

$$q^{(m)} \equiv (\alpha^{(m)}, \beta^{(m)}, \ell^{(m)})$$

we obtain

$$\Delta_\lambda(\mu, q^{(m)}) \leq \Delta_\lambda(\mu, q^{(m-1)})$$

so  $\lim_{m \rightarrow \infty} \Delta_\lambda(\mu, q^{(m-1)}) - \Delta_\lambda(\mu, q^{(m)}) = 0$ . The algorithm stops (after a finite number of iterations) when the drop in distortion falls below a given threshold. It may be necessary to repeat this procedure several times with different values of  $\lambda$  to obtain a quantizer with rate that is close enough to the desired rate.

As in the fixed-rate case, the algorithm is most often used with the empirical distribution  $\mu_n$  in place of  $\mu$ . Although the sequence of distortions converges as  $m \rightarrow \infty$ , there is no guarantee that the limit is the optimum

distortion. However, the quantizers designed using this algorithm have very favorable performance in general.

### Notes

The optimality conditions of Lemma 7 are due to Chou, Lookabaugh, and Gray (1989) who also introduced the Lagrangian formulation of variable-rate vector quantization discussed in this section. Theorem 9 is proved in György and Linder (2001b) (see György and Linder (2001a) for an existence result that does not assume the Lagrangian formulation). Gray, Linder, and Li (2001) used the Lagrangian formulation for the asymptotic (high-rate) analysis of optimal entropy-constrained vector quantizer performance. The algorithm sketched above is the well-known entropy-constrained vector quantizer design algorithm of Chou *et al.* (1989).

## 8 Consistency of Lagrangian empirical design

We are interested in the performance of quantizers learned from a finite training sequence. As before, let  $X_1^n = X_1, \dots, X_n$  be i.i.d. copies of  $X$  such that  $X_1^n$  and  $X$  are also independent, and let  $\mu_n$  denote the empirical distribution of  $X_1^n$ . Fix  $\lambda > 0$  and assume that the empirically optimal variable-rate quantizer  $q_n^*$  is one that minimizes the empirical Lagrangian distortion:

$$\Delta_\lambda(\mu_n, q_n^*) = \inf_q \Delta_\lambda(\mu_n, q) = \Delta_\lambda^*(\mu_n)$$

i.e.,

$$q_n^* = \arg \min_{q \equiv (\alpha, \beta, \ell)} \frac{1}{n} \sum_{k=1}^n \|X_k - \beta(\alpha(X_k))\|^2 + \lambda \ell(\alpha(X_k)).$$

We will always assume that  $q_n^*$  has a modified nearest neighbor encoder (see Theorem 9).

As before, the performance of the empirically optimal variable-rate quantizer  $q_n^* = (\alpha_n^*, \beta_n^*, \ell_n^*)$  is measured by its Lagrangian test distortion, given by

$$\Delta_\lambda(\mu, q_n^*) = \mathbb{E}[\|X - \beta_n^*(\alpha_n^*(X))\|^2 + \lambda \ell_n^*(\alpha_n^*(X)) | X_1^n].$$

The following theorem, the variable-rate counterpart of Theorem 2, shows that the design based on the minimization of the empirical Lagrangian distortion is consistent.

**Theorem 10 (CONSISTENCY OF LAGRANGIAN EMPIRICAL DESIGN)** *For any  $\lambda > 0$  the sequence of variable-rate quantizers  $q_n^*$ ,  $n = 1, 2, \dots$ , minimizing*

the empirical Lagrangian distortion satisfies

$$\lim_{n \rightarrow \infty} \Delta_\lambda(\mu, q_n^*) = \Delta_\lambda^*(\mu) \quad a.s.$$

The proof is based on the properties of the metric  $\rho(\mu, \nu)$  introduced in Section 3, but we need some additional definitions and auxiliary results.

Let  $\mathcal{D}$  denote the set of all discrete distributions on  $\mathbb{R}^d$  with finite second moment and finite entropy. That is,  $\nu \in \mathcal{D}$  if and only if  $\nu$  is concentrated on a finite or countably infinite set  $\{x_i; i \in \mathcal{I}_\nu\} \subset \mathbb{R}^d$ , and satisfies

$$\sum_{i \in \mathcal{I}_\nu} \|x_i\|^2 \nu(x_i) < \infty, \quad - \sum_{i \in \mathcal{I}_\nu} \nu(x_i) \log_2 \nu(x_i) < \infty.$$

For any  $\nu \in \mathcal{D}$  let  $L_\nu$  denote the minimum expected codelength over all admissible codelength functions  $\ell : \mathcal{I}_\nu \rightarrow \{0, 1, \dots\}$ ,

$$L_\nu = \min_{\ell} \sum_{i \in \mathcal{I}_\nu} \ell(i) \nu(x_i). \quad (35)$$

Note that by the remark after Lemma 7, a minimizing admissible  $\ell$  always exists, and that  $L_\nu < \infty$  by the Shannon-Fano bound (33).

For  $\lambda > 0$ ,  $\mu$  with finite second moment, and  $\nu \in \mathcal{D}$  define

$$\rho_\lambda(\mu, \nu) \triangleq (\rho(\mu, \nu)^2 + \lambda L_\nu)^{1/2}.$$

To interpret  $\rho_\lambda(\mu, \nu)$ , suppose  $X \sim \mu$  and  $Y \sim \nu$  achieve  $\rho(\mu, \nu)$ , and let  $\ell_\nu$  be an admissible codelength achieving  $L_\nu$  in (35). Then  $Y$  can be viewed as the output of a variable-rate “random quantizer” that, to each  $x$ , assigns the reproduction vector  $x_i$  and a binary codeword of length  $\ell_\nu(i)$  with probability  $\mathbb{P}\{Y = x_i | X = x\}$ . The quantity  $\rho_\lambda(\mu, \nu)^2$  is the Lagrangian distortion of this random quantizer.

In this interpretation, the next lemma states that deterministic quantizers always outperform random quantizers in the Lagrangian sense.

**Lemma 8**

$$\Delta_\lambda^*(\mu) = \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu, \nu)^2.$$

PROOF. Suppose  $q \equiv (\alpha, \beta, \ell)$  is a variable-rate quantizer such that  $\Delta_\lambda(\mu, q) < \infty$ . Let  $\nu_q$  denote the distribution of the discrete random variable  $q(X)$  and note that  $\nu_q \in \mathcal{D}$ . Since  $X \sim \mu$  and  $q(X) \sim \nu_q$ ,

$$\begin{aligned} \Delta_\lambda(\mu, q) &= \mathbb{E}\|X - q(X)\|^2 + \lambda \mathbb{E} \ell(\alpha(X)) \\ &\geq \rho(\mu, \nu_q)^2 + \lambda L_{\nu_q} \\ &= \rho_\lambda(\mu, \nu_q)^2 \end{aligned}$$

and hence

$$\Delta_\lambda^*(\mu) = \inf_q \Delta_\lambda(\mu, q) \geq \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu, \nu)^2.$$

To show the reverse inequality, assume  $X \sim \mu$  and  $Y \sim \nu$  achieve  $\rho(\mu, \nu)$ , where  $\nu \in \mathcal{D}$  is concentrated on a countable set of points  $\{y_i; i \in \mathcal{I}_\nu\}$ . Define the variable-rate quantizer  $q$  with index set  $\mathcal{I}_\nu$  to have decoder  $\beta(i) = y_i$ ,  $i \in \mathcal{I}_\nu$ , codelength  $\ell_\nu$  such that

$$L_\nu = \sum_{i \in \mathcal{I}_\nu} \ell_\nu(i) \nu(y_i)$$

and encoder  $\alpha$  that is optimized for  $\beta$  and  $\ell_\nu$ , i.e.,

$$\alpha(x) = \arg \min_{i \in \mathcal{I}_\nu} (\|x - y_i\|^2 + \lambda \ell_\nu(i)).$$

Then

$$\Delta_\lambda(\mu, q) = \mathbb{E} \min_{i \in \mathcal{I}_\nu} (\|X - y_i\|^2 + \lambda \ell_\nu(i)).$$

Since  $X$  and  $Y$  achieve  $\rho(\mu, \nu)$ , and  $Y$  takes values in  $\{y_i; i \in \mathcal{I}_\nu\}$ ,

$$\begin{aligned} \rho_\lambda(\mu, \nu)^2 &= \mathbb{E} \{ \|X - Y\|^2 + \lambda L_\nu \} \\ &= \int_{\mathbb{R}^d} \sum_{i \in \mathcal{I}_\nu} (\|x - y_i\|^2 + \lambda \ell_\nu(i)) \mathbb{P}\{Y = y_i | X = x\} \mu(dx) \\ &\geq \int_{\mathbb{R}^d} \min_{i \in \mathcal{I}_\nu} (\|x - y_i\|^2 + \lambda \ell_\nu(i)) \mu(dx) \\ &= \Delta_\lambda(\mu, q) \end{aligned}$$

and so we obtain

$$\inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu, \nu)^2 \geq \inf_q \Delta_\lambda(\mu, q).$$

□

As a consequence of the previous lemma, we obtain a stability result for the optimal Lagrangian performance which is the variable-rate counterpart of Lemma 3.

**Lemma 9** *For any  $\mu$  and  $\mu'$  with finite second moment,*

$$|\Delta_\lambda^*(\mu)^{1/2} - \Delta_\lambda^*(\mu')^{1/2}| \leq \rho(\mu, \mu').$$

PROOF. Assume  $\Delta_\lambda^*(\mu) \geq \Delta_\lambda^*(\mu')$ . Fix  $\epsilon > 0$  and let  $\nu' \in \mathcal{D}$  be such that

$$\rho_\lambda(\mu', \nu') \leq \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu', \nu) + \epsilon.$$

Then by Lemma 8,

$$\begin{aligned} \Delta_\lambda^*(\mu)^{1/2} - \Delta_\lambda^*(\mu')^{1/2} &= \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu, \nu) - \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu', \nu) \\ &\leq \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu, \nu) - \rho_\lambda(\mu', \nu') + \epsilon \\ &\leq \rho_\lambda(\mu, \nu') - \rho_\lambda(\mu', \nu') + \epsilon \\ &= (\rho(\mu, \nu')^2 + \lambda L_{\nu'})^{1/2} - (\rho(\mu', \nu')^2 + \lambda L_{\nu'})^{1/2} + \epsilon \\ &\leq |\rho(\mu, \nu') - \rho(\mu', \nu')| + \epsilon \\ &\leq \rho(\mu, \mu') + \epsilon \end{aligned}$$

where the third inequality holds because  $(a+c)^{1/2} - (b+c)^{1/2} \leq a^{1/2} - b^{1/2}$  for all  $a \geq b \geq 0$ ,  $c \geq 0$  by the concavity of the square root, and the last inequality follows from the triangle inequality since  $\rho$  is a metric (see Lemma 2). Since  $\epsilon > 0$  was arbitrary, we obtain  $\Delta_\lambda^*(\mu)^{1/2} - \Delta_\lambda^*(\mu')^{1/2} \leq \rho(\mu, \mu')$ . The case  $\Delta_\lambda^*(\mu') \geq \Delta_\lambda^*(\mu)$  is handled similarly.  $\square$

The preceding lemma immediately implies that the Lagrangian training distortion  $\Delta_\lambda(\mu_n, q_n^*)$  is a strongly consistent estimate of the optimal Lagrangian distortion  $\Delta_\lambda^*(\mu)$ .

### Theorem 11

$$\lim_{n \rightarrow \infty} \Delta_\lambda(\mu_n, q_n^*) = \Delta_\lambda^*(\mu) \quad a.s.$$

PROOF. Since  $\Delta_\lambda(\mu_n, q_n^*) = \Delta_\lambda^*(\mu_n)$ , Lemma 9 with  $\mu' = \mu_n$  gives

$$|\Delta_\lambda^*(\mu)^{1/2} - \Delta_\lambda^*(\mu_n)^{1/2}| \leq \rho(\mu, \mu_n).$$

The statement follows since we know from the proof of Theorem 2 that

$$\lim_{n \rightarrow \infty} \rho(\mu, \mu_n) = 0 \quad a.s. \quad (36)$$

$\square$

Now we are ready to prove the consistency theorem.

PROOF OF THEOREM 10. We have

$$\begin{aligned} \Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu) &= \Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu_n) + \Delta_\lambda^*(\mu_n) - \Delta_\lambda^*(\mu). \end{aligned} \quad (37)$$

The second difference on the right side converges to zero a.s. by Theorem 11.

To bound the first difference, recall that by assumption, the encoder of  $q_n^* \equiv (\alpha_n^*, \beta_n^*, \ell_n^*)$  uses the modified nearest neighbor rule. Thus for any  $x, y \in \mathbb{R}^d$ ,

$$\|x - q_n^*(x)\|^2 + \lambda \ell_n^*(\alpha_n^*(x)) \leq \|x - q_n^*(y)\|^2 + \lambda \ell_n^*(\alpha_n^*(y)). \quad (38)$$

Now let  $X \sim \mu$  and  $Y \sim \mu_n$  achieve  $\rho(\mu, \mu_n)$ , and in addition suppose that the pair  $X, Y$  is independent of the training sequence  $X_1^n$ . Letting  $\mathbb{E}_n$  denote conditional expectation with respect to  $X_1^n$ , (38) implies

$$\begin{aligned} \Delta_\lambda(\mu, q_n^*) &= \mathbb{E}_n \{ \|X - q_n^*(X)\|^2 + \lambda \ell_n^*(\alpha_n^*(X)) \} \\ &\leq \mathbb{E}_n \{ \|X - q_n^*(Y)\|^2 + \lambda \ell_n^*(\alpha_n^*(Y)) \} \\ &\leq \mathbb{E}_n \|X - Y\|^2 + \mathbb{E}_n \{ \|Y - q_n^*(Y)\|^2 + \lambda \ell_n^*(\alpha_n^*(Y)) \} \\ &\quad + 2\mathbb{E}_n \{ \|X - Y\| \|Y - q_n^*(Y)\| \} \\ &= \rho(\mu, \mu_n)^2 + \Delta_\lambda(\mu_n, q_n^*) + 2\mathbb{E}_n \{ \|X - Y\| \|Y - q_n^*(Y)\| \} \\ &\leq \rho(\mu, \mu_n)^2 + \Delta_\lambda(\mu_n, q_n^*) + 2(\mathbb{E}_n \|X - Y\|^2)^{1/2} (\mathbb{E}_n \|Y - q_n^*(Y)\|^2)^{1/2} \\ &= \rho(\mu, \mu_n)^2 + \Delta_\lambda^*(\mu_n) + 2\rho(\mu, \mu_n) D(\mu_n, q_n^*)^{1/2} \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Since  $D(\mu_n, q_n^*) \leq \Delta_\lambda(\mu_n, q_n^*) = \Delta_\lambda^*(\mu_n)$ , using Lemma 9 again, we obtain

$$D(\mu_n, q_n^*)^{1/2} \leq \Delta_\lambda^*(\mu)^{1/2} + \rho(\mu, \mu_n)$$

and so

$$\Delta_\lambda(\mu, q_n^*) \leq \rho(\mu, \mu_n)^2 + \Delta_\lambda^*(\mu_n) + 2\rho(\mu, \mu_n) (\Delta_\lambda^*(\mu)^{1/2} + \rho(\mu, \mu_n))$$

or

$$\Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu_n) \leq \rho(\mu, \mu_n)^2 + 2\rho(\mu, \mu_n) (\Delta_\lambda^*(\mu)^{1/2} + \rho(\mu, \mu_n)).$$

Therefore, by (36), the first difference in (37) converges to zero a.s. as  $n \rightarrow \infty$ , completing the proof of the theorem.  $\square$

## Notes

The material of this section is new. Lemma 8 has a counterpart in fixed-rate quantization; see Pollard (1982a). Zador (1966) was probably the first to use random quantization as a tool for analyzing the performance of optimal variable-rate quantizers.

## 9 Finite sample bounds in Lagrangian design

In analogy to the fixed-rate case, we would like to establish a connection between the performance of the empirically designed variable-rate quantizer and the number of training samples used in the design. The methods used in proving such results in the fixed-rate case will turn out to be applicable once we have established a variable-rate counterpart of Lemma 5.

As we did in Section 4, we assume here that the source distribution is an element of  $\mathcal{P}(T)$ , the set of probability distributions on  $\mathbb{R}^d$  supported on the closed ball  $B_T$  of radius  $T$  centered at the origin.

In the next lemma we show that for sources with a given bounded support, variable-rate quantizers that are optimal in the Lagrangian sense can be assumed not to have too many codevectors or very large codeword lengths. The strength of the Lagrangian approach is evident here; no such general result is known for variable-rate quantizers that minimize the distortion for a given rate constraint.

For  $T > 0$  and positive integers  $N$  and  $L$ , let  $\mathcal{Q}_{N,L}(T)$  denote the collection of all variable-rate quantizers  $q \equiv (\alpha, \beta, \ell)$  with index set  $\mathcal{I}$  such that

- (i)  $\|\beta(i)\| \leq T$  for all  $i \in \mathcal{I}$ ;
- (ii)  $\alpha$  is a modified nearest neighbor encoder;
- (iii)  $\ell(i) \leq L$  for all  $i \in \mathcal{I}$ , and  $\mathcal{I}$  is finite with cardinality  $|\mathcal{I}| \leq N$ .

**Lemma 10** *For any  $\mu \in \mathcal{P}(T)$  and  $\lambda > 0$ ,*

$$\min_q \Delta_\lambda(\nu, q) = \min_{q \in \mathcal{Q}_{N,L}(T)} \Delta_\lambda(\nu, q) \quad (39)$$

where  $N = \lfloor 2^{5T^2/\lambda} \rfloor$  and  $L = \lfloor 5T^2/\lambda \rfloor$ . Thus there exists  $q_n^* \in \mathcal{Q}_{N,L}(T)$ , and for this  $q_n^*$  we have

$$\Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu) \leq 2 \sup_{q \in \mathcal{Q}_{N,L}(T)} |\Delta_\lambda(\mu_n, q) - \Delta_\lambda(\mu, q)| \quad (40)$$

PROOF. The second statement is an easy consequence of the first one. Let  $q^*$  denote a variable-rate quantizer achieving the minimum Lagrangian distortion  $\Delta_\lambda^*(\mu)$ . Using the same argument as in Lemma 5, we obtain the basic inequality

$$\begin{aligned} \Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu) \\ \leq \Delta_\lambda(\mu, q_n^*) - \Delta_\lambda(\mu_n, q_n^*) + \Delta_\lambda(\mu_n, q^*) - \Delta_\lambda(\mu, q^*). \end{aligned}$$

Since  $\mu, \mu_n \in \mathcal{P}(T)$ , (39) implies that there exist  $q^*, q_n^* \in \mathcal{Q}_{N,L}(T)$ , and we obtain (40).

To prove (39), suppose  $q \equiv (\alpha, \beta, \ell)$  with index set  $\mathcal{I}$  achieves  $\Delta_\lambda^*(\nu)$ . Using the argument of Lemma 5, we see that any codevector of  $q$  outside  $B_T$  can be replaced by its projection to the surface of  $B_T$  without increasing the Lagrangian distortion, and so we can assume that  $\|\beta(i)\| \leq T$  for all  $i \in \mathcal{I}$ .

Next note that by Lemma 7 (a) we can assume that  $\alpha$  is a modified nearest neighbor encoder. Also, we can assume that for each  $S_i = \{x : \alpha(x) = i\}$ ,  $i \in \mathcal{I}$ , we have  $S_i \cap B_T \neq \emptyset$ ; otherwise, since  $\nu \in \mathcal{P}(T)$ , we can discard  $i$  from  $\mathcal{I}$  without affecting the performance. Let  $i_0 \in \mathcal{I}$  be an index with minimum codeword length, i.e.,

$$\ell(i_0) = \min_{i \in \mathcal{I}} \ell(i).$$

Since  $\alpha$  is a modified nearest neighbor encoder, for any  $i \in \mathcal{I}$  and  $x \in S_i$ ,

$$\|x - \beta(i)\|^2 + \lambda \ell(i) \leq \|x - \beta(i_0)\|^2 + \lambda \ell(i_0).$$

Since  $\|\beta(i_0)\| \leq T$ , we have  $\|x - \beta(i_0)\|^2 \leq 4T^2$  for all  $x \in B_T$ , and since  $S_i \cap B_T$  is nonempty, the previous inequality implies that for all  $i \in \mathcal{I}$ ,

$$\ell(i) \leq \frac{4T^2}{\lambda} + \ell(i_0). \quad (41)$$

Now let  $q_1$  denote the quantizer with a single codepoint  $y = 0$  and rate  $r(\nu, q_1) = 0$  (formally, the single binary codeword of  $q_1$  is the empty string of length zero). Then since  $\nu \in \mathcal{P}(T)$ ,

$$\Delta_\lambda(\nu, q_1) = D(\nu, q_1) + \lambda r(\nu, q_1) \leq T^2. \quad (42)$$

On the other hand,

$$\Delta_\lambda(\nu, q) \geq \lambda r(\nu, q) \geq \lambda \ell(i_0)$$

which, together with (42) and the fact that  $\Delta_\lambda(\nu, q) \leq \Delta_\lambda(\nu, q_1)$  (since  $q$  minimizes the Lagrangian distortion for  $\nu$ ), implies that

$$\ell(i_0) \leq \frac{T^2}{\lambda}.$$

Hence by (41), we have for all  $i \in \mathcal{I}$ ,

$$\ell(i) \leq \frac{5T^2}{\lambda}.$$

The admissibility of  $\ell$  then yields

$$1 \geq \sum_{i \in \mathcal{I}} 2^{-\ell(i)} \geq |\mathcal{I}| 2^{-5T^2/\lambda}$$

and so  $|\mathcal{I}| \leq 2^{5T^2/\lambda}$ . Setting  $N = \lfloor 2^{5T^2/\lambda} \rfloor$  and  $L = \lfloor 5T^2/\lambda \rfloor$ , we obtain  $q \in \mathcal{Q}_{N,L}(T)$ , which completes the proof.  $\square$

Lemma 10 allows us to adapt the proof of Theorem 3 to the Lagrangian case. The following finite sample bound on the Lagrangian performance on an empirically optimal variable-rate quantizer is the main result of this section. In this result we assume (as we may by Lemma 10) that  $q_n^* \in \mathcal{Q}_{N,L}(T)$ .

**Theorem 12** *There is a constant  $C_4$ , depending only on  $d$ ,  $\lambda$ , and  $T$ , such that for all  $n \geq 1$  and  $\mu \in \mathcal{P}(T)$ ,*

$$\mathbb{E} \Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu) \leq \frac{C_4}{\sqrt{n}}.$$

PROOF. By Lemma 10, it suffices to give an appropriate upper bound on the expected value of  $\sup_{q \in \mathcal{Q}_{N,L}(T)} |\Delta_\lambda(\mu_n, q) - \Delta_\lambda(\mu, q)|$ .

For  $N$  and  $L$  as in Lemma 10 and  $q \equiv (\alpha, \beta, \ell) \in \mathcal{Q}_{N,L}(T)$ , define the distortion function

$$f_{\lambda,q}(x) \triangleq \|x - q(x)\|^2 + \lambda \ell(\alpha(x)).$$

Then for all  $x \in B_T$ ,

$$0 \leq f_{\lambda,q}(x) \leq 4T^2 + \lambda L \leq 4T^2 + \lambda \lfloor 5T^2/\lambda \rfloor = 9T^2.$$

Hence, we can repeat the steps leading to (11) in the proof of Theorem 3 to obtain

$$\begin{aligned} & \sup_{q \in \mathcal{Q}_{N,L}(T)} |\Delta_\lambda(\mu_n, q) - \Delta_\lambda(\mu, q)| \\ & \leq 9T^2 \sup_{q \in \mathcal{Q}_{N,L}(T), u > 0} \left| \frac{1}{n} \sum_{k=1}^n I_{\{f_{\lambda,q}(X_k) > u\}} - \mathbb{P}\{f_{\lambda,q}(X) > u\} \right| \\ & = 9T^2 \sup_{A \in \tilde{\mathcal{A}}_N} |\mu_n(A) - \mu(A)| \quad \text{a.s.} \end{aligned}$$

where now  $\tilde{\mathcal{A}}_N$  is the family of subsets of  $\mathbb{R}^d$  defined by

$$\tilde{\mathcal{A}}_N \triangleq \left\{ \{x : f_{\lambda,q}(x) > u\} : q \in \mathcal{Q}_{N,L}(T), u > 0 \right\}.$$

Denoting by  $V(\tilde{\mathcal{A}}_N)$  the VC dimension of  $\tilde{\mathcal{A}}_N$  and using the sharpened version of the Vapnik-Chervonenkis inequality as in the proof of Theorem 3, we obtain

$$\mathbb{E} \left\{ \sup_{q \in \mathcal{Q}_{N,L}(T)} |\Delta_\lambda(\mu_n, q) - \Delta_\lambda(\mu, q)| \right\} \leq 9T^2 c \sqrt{\frac{V(\tilde{\mathcal{A}}_N)}{n}}. \quad (43)$$

Since the encoder of each  $q \in \mathcal{Q}_{N,L}(T)$  uses the modified nearest neighbor rule,  $f_{\lambda,q}(x) > u$  if and only if

$$\|x - \beta(i)\|^2 + \lambda\ell(i) > u \quad \text{for all } i \in \mathcal{I}$$

i.e.,

$$\{x : f_{\lambda,q}(x) > u\} = \bigcap_{i \in \mathcal{I}} \{x : \|x - \beta(i)\|^2 > u - \lambda\ell(i)\}.$$

Since  $|\mathcal{I}| \leq N$ , we obtain that either  $\{x : f_{\lambda,q}(x) > u\}$  is an intersection of the complements of at most  $N$  closed balls of possibly different radii in  $\mathbb{R}^d$  (if  $u - \lambda\ell(i) \geq 0$  for some  $i$ ), or  $\{x : f_{\lambda,q}(x) > u\} = \mathbb{R}^d$  (this is the case if  $u - \lambda\ell(i) < 0$  for all  $i$ ). As in the proof of Theorem 3, let  $\bar{\mathcal{A}}_N$  denote the family of all intersections of complements of  $N$  closed balls in  $\mathbb{R}^d$ . Then we have  $\tilde{\mathcal{A}}_N \subset \bar{\mathcal{A}}_N \cup \{\mathbb{R}^d\}$ , and so  $V(\tilde{\mathcal{A}}_N) \leq V(\bar{\mathcal{A}}_N)$ . Thus we can use the upper bound  $V(\bar{\mathcal{A}}_N) \leq 4N(d+1) \ln(N(d+1))$  derived in the proof of Theorem 3 to obtain

$$V(\tilde{\mathcal{A}}_N) \leq 4N(d+1) \ln(N(d+1)).$$

Thus (43) and Lemma 10 imply the theorem with

$$C_4 = 18T^2 c \sqrt{4N(d+1) \ln(N(d+1))}$$

where  $N = \lfloor 2^{5T^2/\lambda} \rfloor$ . □

**Remark** A variable-rate versions of Theorem 5 and Theorem 6 can also be derived from Lemma 10 and the inequality (43). In particular, the variable-rate counterpart of Theorem 5 states that for every  $\mu \in \mathcal{P}(T)$  the bias of the expected Lagrangian training distortion is upper bounded as

$$\Delta_\lambda^*(\mu) - \mathbb{E} \Delta_\lambda(\mu_n, q_n^*) \leq \frac{C_4/2}{\sqrt{n}}$$

where  $C_4$  is the constant in Theorem 12. Also, the bounded difference inequality can be used to show that for every  $\mu \in \mathcal{P}(T)$ ,

$$\Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu) = O\left(\sqrt{\frac{\ln n}{n}}\right) \quad \text{a.s.}$$

which is the variable-rate counterpart of Theorem 6. The details are left as an exercise.

## Notes

The proof of Lemma 10 is based on an idea of Chou and Betts (1998). The rest of the material in this section is new.

## References

- [1] E. A. Abaya and G. L. Wise. On the existence of optimal quantizers. *IEEE Trans. Inform. Theory*, 28:937 – 940, Nov. 1982.
- [2] E. A. Abaya and G. L. Wise. Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal on Applied Mathematics*, 44:183–189, 1984.
- [3] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
- [4] R. B. Ash. *Probability and Measure Theory*. Academic Press, New York, 2000.
- [5] P. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory*, IT-44(5):1802–1813, Sep. 1998.
- [6] T. Berger. Optimum quantizers and permutation codes. *IEEE Trans. Inform. Theory*, IT-18:759–765, Nov. 1972.
- [7] P. A. Chou. The distortion of vector quantizers trained on  $n$  vectors decreases to the optimum as  $O_p(1/n)$ . in *Proc. IEEE Int. Symp. Information Theory* (Trondheim, Norway, Jun. 27-Jul. 1, 1994), p. 457.
- [8] P. A. Chou and B. J. Betts. When optimal entropy-constrained quantizers have only a finite number of codewords. in *Proc. IEEE Int. Symp. Information Theory* (Cambridge, MA, USA, Aug. 16-21, 1998), p. 97.
- [9] P. A. Chou, T. Lookabaugh, and R. M. Gray. Entropy-constrained vector quantization. *IEEE Trans. Acoust. Speech, Signal Processing*, ASSP-37:31–42, Jan. 1989.
- [10] D. Cohn, E. Riskin, and R. Ladner. Theory and practice of vector quantizers trained on small training sets. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16:54–65, Jan. 1994.

- [11] P. C. Cosman, K. O. Perlmutter, S. M. Perlmutter, R. A. Olshen, and R. M. Gray. Training sequence size and vector quantizer performance. In *Proceedings of Asilomar Conference on Signals, Systems, and Computers*, pages 434–438, 1991.
- [12] T. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
- [13] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.
- [14] R. M. Dudley. *Real Analysis and Probability*. Chapman & Hall, New York, 1989.
- [15] R.M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.
- [16] A Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer, Boston, 1992.
- [17] H. Gish and J. N. Pierce. Asymptotically efficient quantizing. *IEEE Trans. Inform. Theory*, IT-14:676–683, Sep. 1968.
- [18] S. Graf and H. Luschgy. Consistent estimation in the quantization problem for random vectors. In *Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 84–87, 1994.
- [19] S. Graf and H. Luschgy. Rates of convergence for the empirical quantization error. preprint, 1999.
- [20] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer Verlag, Berlin, Heidelberg, 2000.
- [21] R. M. Gray. *Source Coding Theory*. Kluwer, Boston, 1990.
- [22] R. M. Gray and L. D. Davisson. Quantizer mismatch. *IEEE Trans. Communications*, 23:439–443, 1975.
- [23] R. M. Gray, T. Linder, and J. Li. A Lagrangian formulation of Zador’s entropy-constrained quantization theorem. *IEEE Trans. Inform. Theory*, 2001 (to appear).
- [24] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Trans. Inform. Theory*, (Special Commemorative Issue), IT-44(6):2325–2383, Oct. 1998.

- [25] R. M. Gray, D. L. Neuhoff, and P. C. Shields. A generalization of Orstein's  $\bar{d}$ -distance with applications to information theory. *Annals of Probability*, 3:315–328, 1975.
- [26] A. György and T. Linder. Optimal entropy-constrained scalar quantization of a uniform source. *IEEE Trans. Inform. Theory*, IT-46:pp. 2704–2711, Nov. 2000.
- [27] A. György and T. Linder. On the structure of optimal entropy-constrained scalar quantizers. *IEEE Trans. Inform. Theory*, 2001 (to appear).
- [28] A. György and T. Linder. On optimal Lagrangian entropy-constrained vector quantization. preprint, 2001.
- [29] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, COM-28:84–95, Jan. 1980.
- [30] T. Linder. On the training distortion of vector quantizers. *IEEE Trans. Inform. Theory*, IT-46:1617–1623, Jul. 2000.
- [31] T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Trans. Inform. Theory*, 40:1728–1740, Nov. 1994.
- [32] T. Linder, G. Lugosi, and K. Zeger. Empirical quantizer design in the presence of source noise or channel noise. *IEEE Trans. Inform. Theory*, IT-43:612–623, Mar. 1997.
- [33] T. Linder, V. Tarokh, and K. Zeger. Existence of optimal prefix codes for infinite source alphabets. *IEEE Trans. Inform. Theory*, IT-43:2026–2028, Nov. 1997.
- [34] S. P. Lloyd. Least squared quantization in PCM. unpublished memorandum, Bell Lab., 1957; Also, *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 129-137., Mar. 1982.
- [35] G. Lugosi. Pattern classification and learning theory. Lecture notes for the *Advanced School on the Principles of Nonparametric Learning*, Udine, Italy, July 9-13, 2001.
- [36] J. MacQueen. Some methods for classification and analysis of multivariate observations. in *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, pp. 281–296, 1967.

- [37] N. Merhav and J. Ziv. On the amount of side information required for lossy data compression. *IEEE Trans. Inform. Theory*, IT-43:1112–1121, July 1997.
- [38] D. Pollard. Strong consistency of  $k$ -means clustering. *Annals of Statistics*, 9, no. 1:135–140, 1981.
- [39] D. Pollard. Quantization and the method of  $k$ -means. *IEEE Trans. Inform. Theory*, IT-28:199–205, Mar. 1982.
- [40] D. Pollard. A central limit theorem for  $k$ -means clustering. *Annals of Probability*, vol. 10, no. 4:919–926, 1982.
- [41] D. Pollard. *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, CA, 1990.
- [42] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.
- [43] M. J. Sabin. *Global convergence and empirical consistency of the generalized Lloyd algorithm*. PhD thesis, Stanford Univ., 1984.
- [44] M. J. Sabin and R. M. Gray. Global convergence and empirical consistency of the generalized Lloyd algorithm. *IEEE Trans. Inform. Theory*, IT-32:148–155, Mar. 1986.
- [45] K. Sayood. *Introduction to Data Compression*. Morgan Kaufmann Publishers, San Francisco, 2nd edition, 2001.
- [46] E.V. Slud. Distribution inequalities for the binomial law. *Annals of Probability*, 5:404–412, 1977.
- [47] H. Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, IV:801–804, May 1956.
- [48] S. Szarek. On the best constants in the Khintchine inequality. *Studia Mathematica*, 63:197–208, 1976.
- [49] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- [50] P. Zador. Topics in the asymptotic quantization of continuous random variables. unpublished memorandum, Bell Laboratories, Murray Hill, NJ, Feb. 1966.

- [51] P. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans. Inform. Theory*, IT-28:139–149, Mar. 1982.
- [52] A. J. Zeevi. On the performance of vector quantizers empirically designed from dependent sources. in *Proceedings of Data Compression Conference, DCC'98*, (J. Storer, M. Cohn, ed.) pp. 73–82, IEEE Computer Society Press, Los Alamitos, California, 1998.