The Minimax Distortion Redundancy in Empirical Quantizer Design

Peter L. Bartlett, *Member, IEEE*, Tamás Linder, *Member, IEEE*, and Gábor Lugosi

***Abstract*—We obtain minimax lower and upper bounds for the expected distortion redundancy of empirically designed vector quantizers. We show that the mean-squared distortion of a vector quantizer designed from n independent and identically distributed (i.i.d.) data points using any design algorithm is at least $\Omega(n^{-1/2})$ away from the optimal distortion for some distribution on a bounded subset of $\mathcal{R}^d$. Together with existing upper bounds this result shows that the minimax distortion redundancy for empirical quantizer design, as a function of the size of the training data, is asymptotically on the order of $n^{-1/2}$. We also derive a new upper bound for the performance of the empirically optimal quantizer.**

***Index Terms*—Distortion redundancy, empirical quantizer design, lower bounds, minimax convergence rate, vector quantization.**

I. Introduction

ONE basic problem of data compression is the design of a vector quantizer without the knowledge of the source statistics. In this situation, a collection of sample vectors (called the training data) is given and the objective is to find a vector quantizer of a given rate whose average distortion on the source is as close as possible to the distortion of the optimal (i.e., minimum distortion) quantizer of the same rate.

Most existing design algorithms (see, e.g., [7], [9], [19], and [23]) attempt to implement, in various ways, the principle of empirical error minimization in the vector quantization context. According to this principle, a good quantizer can be found by searching for one that minimizes the distortion *over the training data*. If the training data represents the source well, this empirically optimal quantizer will hopefully perform near optimally also on the real source. The problem of quantifying how good empirically designed quantizers are compared to the truly optimal ones has been extensively studied for the case when the training data consists of n vectors independently drawn from the source distribution. It was shown by Pollard [16], [18] under general conditions that the method of empirical error minimization is consistent in the following sense. Let D_n be mean-squared error (MSE) of the empirically optimal quantizer, when measured on the real source, and let D^* be the minimum MSE achieved by an optimal quantizer. An empirically designed quantizer is consistent if the quantity $D_n - D^*$ (called the distortion redundancy) converges to zero as n tends to infinity.

Of course mere consistency does not give any indication of how large the training data should be so that the distortion of the designed quantizer is close to the optimum. This question can only be answered by analyzing the finite sample behavior of D_n. In this direction, it was shown in [10] and [15] that there exists a c such that $D_n - D^* \leq c\sqrt{\log n/n}$ for all sources over a bounded region. This result has since been extended to empirical quantizer design for vector quantizers operating on "noisy" sources and for vector quantizers for noisy channels [11]. An extension to unbounded sources is given in [13].

A deeper analysis of the method used to obtain the above upper bound shows that at the price of considerable technical difficulties, the $\sqrt{\log n}$ factor can be eliminated. Indeed, using a result of Alexander [1] the above upper bound can be sharpened to $O(1/\sqrt{n})$.

Two basic questions relating to the finite sample behavior of quantizer design algorithms have remained unanswered. The first is whether the $O(1/\sqrt{n})$ upper bound on the *distortion redundancy* $D_n - D^*$ is actually tight. The second, more general question is whether there exist methods, other than empirical error minimization, which provide smaller distortion redundancy (and thus use less training data to achieve the same distortion). The results of this paper answer both questions in a minimax sense.

There are indications that the upper bound can be tightened to $O(1/n)$. Indeed, for the special case of a one-codepoint scalar quantizer one can define the codepoint to be the average of the n independent and identically distributed (i.i.d.) training samples, a choice which actually minimizes the squared error on the training data. It is easy to see that $D_n - D^* = c/n$, where c is the variance of the source. Another indication that an $O(1/n)$ rate might be achieved comes from a result of Pollard [17]. He showed that for sources with some specially smooth and regular densities, the difference between the codepoints of the empirically designed quantizers and the codepoints of the optimal quantizer obeys a multidimensional

Manuscript received February 16, 1997; revised March 10, 1998. This work was supported in part by OTKA under Grant F 014174, under a DIST Bilateral Science and Technology Collaboration Grant, by DGES under Grant PB96-0300, and by the National Science Foundation. The material in this work was presented in part at the EUROCOLT'97, Jerusalem, Israel, 1997, and at the IEEE International Symposium on Information Theory, Ulm, Germany, June 1997.

P. L. Bartlett is with the Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra 0200, Australia (e-mail: Peter.Bartlett@anu.edu.au).

T. Linder was with the Department of Electrical and Computer Engineering, University of California, San Diego, CA, USA on leave from the Technical University of Budapest, Hungary. He is now with the Department of Mathematics and Statistics, Queen's University, Kingston, Ont., Canada K7L 3N6 (e-mail: linder@code.ucsd.edu).

G. Lugosi is with the Department of Economics, Pompeu Fabra University, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain (e-mail: lugosi@upf.es).

Publisher Item Identifier S 0018-9448(98)04787-7.

central limit theorem. As Chou [3] pointed out, this implies that within the class of sources in the scope of this result, the distortion redundancy decreases at a rate $O(1/n)$ *in probability*.

In the main result of this paper (Theorem 1) we show that despite these suggestive facts, the conjectured $O(1/n)$ distortion redundancy rate *does not* hold in the minimax sense. Let $B > 0$ and consider the class $\mathcal{B}$ of d-dimensional source distributions μ such that if X is distributed according to μ, then $(1/d)\|X\|^2 \leq B$ with probability one. We show that for any d-dimensional k-codepoint ($k > 2$) quantizer Q_n which is designed by *any* method from n independent training samples, there exists a distribution in $\mathcal{B}$ for which the per-dimension MSE of Q_n is bounded away from the optimal distortion by a constant times $\sqrt{\frac{k^{1-4/d}}{n}}$. Thus the gap between this lower bound and the existing upper bound is reduced to a constant factor, if the parameters k and d are kept constant.

In addition to this general lower bound, a new minimax upper bound for the empirically optimal quantizer is derived in Theorem 2. The bound is a constant times $\sqrt{\frac{k^{1-2/d}\log n}{n}}$. The main merit of this bound is that it partially explains the curious dependence of the lower bound on k: the bound *decreases* in k for very small values of d. Also, for realistic values of quantizer dimension and rate, it is tighter than the $O(1/\sqrt{n})$ bound obtained via Alexander's inequality, and yet its proof is rather elementary and accessible.

II. Main Results

A d-dimensional k-point quantizer Q is a mapping

$$Q(x) = y_i, \qquad \text{if } x \in B_i$$

where $B_1, \cdots, B_k$ form a measurable partition of $\mathcal{R}^d$, and $y_i \in \mathcal{R}^d$, $1 \leq i \leq k$. The y_i's are called codepoints, and the collection of codepoints $\{y_1, \cdots, y_k\}$ is the codebook. If μ is a probability measure on $\mathcal{R}^d$, the distortion of Q with respect to μ is

$$D(Q) = \int_{\mathcal{R}^d} \|x - Q(x)\|^2 \, \mu(dx)$$

where $\|x - Q(x)\|$ is the Euclidean distance between x and $Q(x)$.

An empirically designed k-point quantizer is a measurable function Q_n: $(\mathcal{R}^d)^{n+1} \rightarrow \mathcal{R}^d$ such that for each fixed $x_1, \cdots, x_n \in \mathcal{R}^d$, $Q_n(\cdot, x_1, \cdots, x_n)$ is a k-point quantizer. Thus an "empirically designed quantizer" consists of a family of quantizers and an "algorithm" which chooses one of them for each value of the training data $x_1, \cdots, x_n$.

In our investigation, $X, X_1, \cdots, X_n$ are i.i.d. random variables in $\mathcal{R}^d$ distributed according to some probability measure μ with $\mu(S(0, \sqrt{d})) = 1$, where $S(x, r) \subset \mathcal{R}^d$ denotes the closed ball of radius $r \geq 0$ centered at $x \in \mathcal{R}^d$. In other words, we assume that the normalized squared norm $(1/d)\|X\|^2$ of X is bounded by one with probability one. (By straightforward scaling one can generalize our results to cases with $\mu(S(0, \sqrt{d}B)) = 1$ for some fixed $B < \infty$.) The distortion of Q_n is the random variable

$$\begin{aligned} D(Q_n) &= \int_{\mathcal{R}^d} \|x - Q_n(x, X_1, \cdots, X_n)\|^2 \mu(dx) \\ &= \boldsymbol{E}[\|X - Q_n(X, X_1, \cdots, X_n)\|^2 | X_1, \cdots, X_n]. \end{aligned}$$

Let $D^*(k, \mu)$ be the minimum distortion achievable by the best k-point quantizer under the source distribution μ. That is,

$$D^*(k, \mu) = \min_Q \int_{\mathcal{R}^d} \|x - Q(x)\|^2 \mu(dx)$$

where the minimum is taken over all d-dimensional, k-point quantizers. The following quantity is in the focus of our attention:

$$J(Q_n, \mu) = \boldsymbol{E}D(Q_n) - D^*(k, \mu)$$

that is, the expected excess distortion of Q_n over the optimal quantizer for μ. In particular, we are interested in the *minimax expected distortion redundancy*, defined by

$$J^*(n, k, d) = \inf_{Q_n} \sup_{\mu} J(Q_n, \mu) \tag{1}$$

where the infimum is taken over all d-dimensional, k-point empirical quantizers trained on n samples, and the supremum is taken over all distributions over the ball $S(0, \sqrt{d})$ in $\mathcal{R}^d$. The minimax expected distortion redundancy expresses the minimal worst case excess distortion that an empirical quantizer can have.

A quantizer Q is a *nearest neighbor quantizer* if for all x, $\|x - Q(x)\| \leq \|x - y_i\|$ for all codepoints y_i of Q. It is well known that for each quantizer Q and distribution μ there exists a nearest neighbor quantizer which has the same codebook as Q but less than or equal distortion. Therefore, when investigating the minimax distortion redundancy, it suffices to consider nearest neighbor quantizers.

The empirically optimal quantizer, denoted Q_n^*, is an empirically designed quantizer which minimizes the empirical error

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

over all k-point nearest neighbor quantizers Q.

The first result upper-bounding the minimax distortion redundancy was given in [10], where it was proved that for the empirically optimal quantizer

$$J(Q_n^*, \mu) \leq cd^{3/2}\sqrt{\frac{k \log n}{n}} \tag{2}$$

for all μ, where c is a universal constant. The main message of the above inequality is that there exists a sequence of empirical quantizers such that for all distributions supported on a given d-dimensional sphere the expected distortion redundancy decreases as $O(\sqrt{\log n/n})$. Another application of this result, which uses the dependence of this bound on k, was pointed out in [13] (see the discussion after Theorem 2).

With analysis based on sophisticated uniform large-deviation inequalities of Alexander [1] or Talagrand [21]

it is possible to get rid of the $\sqrt{\log n}$ factor. More precisely, one can prove that

$$J(Q_n^*, \mu) \le c' d^{3/2} \sqrt{\frac{k \log(kd)}{n}} \tag{3}$$

for all μ, where c' is another universal constant (see the discussion in [10] and [6, Problem 12.10]).

The theorem below—the main result of this paper—shows that for any empirical quantizer Q_n (i.e., for any design method whose input is $X_1, \cdots, X_n$ and output is a d-dimensional, k-codepoint quantizer Q_n) the excess distortion is as large as a constant times $d\sqrt{\frac{k^{1-4/d}}{n}}$ for some distribution. Let Φ denote the distribution function of a standard normal random variable.

Theorem 1: For any dimension d, number of codepoints $k \ge 3$, and sample size $n \ge 16k/(3\Phi(-2)^2)$, and for any empirically designed k-point quantizer Q_n, there exists a distribution μ on $S(0, \sqrt{d})$ such that

$$J(Q_n, \mu) \ge c_0 d \sqrt{\frac{k^{1-4/d}}{n}} \tag{4}$$

where c_0 is a universal constant which may be taken to be $c_0 = \Phi(-2)^4 2^{-12}/\sqrt{6}$.

The proof of the theorem is given in the next section.

Remarks:

i) In the proof of the theorem, for the sake of simplicity, we consider a family of distributions concentrated on a finite set of points in $S(0, \sqrt{d})$. It is then demonstrated that for each Q_n there exists a μ in this family for which (4) holds. Since these distributions can be arbitrarily well approximated (for our purposes) by distributions with smooth (say infinitely many times differentiable) densities, essentially the same argument shows that for each Q_n there exists a μ with a smooth density such that (4) holds.

ii) The constant c_0 of the theorem is rather small (note that $\Phi(-2) \approx 0.0228$), and it can probably be improved upon at the expense of a more complicated analysis.

The above theorem, together with (3), essentially describes the convergence rate of the minimax expected distortion redundancy in terms of the sample size n. Using definition (1) we obtain that

$$\limsup_{n\to\infty} \sqrt{n} J^*(n, k, d) \le c_1$$

and

$$\liminf_{n\to\infty} \sqrt{n} J^*(n, k, d) \ge c_2$$

for some constants $c_1, c_2 > 0$ depending on d and k. However, there is still a gap if the bounds are viewed in terms of the number of codepoints k. For large d the difference is small. In fact, if, according to the usual information-theoretic asymptotic view, the number of codepoints is set as $k = 2^{Rd}$ for some constant rate $R > 0$, then the difference between the upper and lower bounds is asymptotically negligible in an exponential sense. Indeed, (3) and Theorem 1 imply that for large d, the per-dimension minimax distortion redundancy is sandwiched as

$$\sqrt{\frac{2^{d(R-O(d^{-1}))}}{n}} \le d^{-1} J^*(n, k, d) \le \sqrt{\frac{2^{d(R+O(d^{-1}\log d))}}{n}}.$$

The difference is more essential for small d, not only because of the difference in the exponents of k in the two bounds, but also because the constant c' in (3) is large (it is of the order of 10^3), a price paid for eliminating the $\sqrt{\log n}$ factor in (2). For this reason, we now present a new minimax upper bound on the distortion redundancy of empirically optimal quantizers.

Theorem 2: For the class of sources considered in Theorem 1, if $n \ge k^{4/d}$, $\sqrt{dk^{1-2/d} \log n} \ge 15$, $kd \ge 8$, $n \ge 8d$, and $n/\log n \ge dk^{1+2/d}$, then

$$J(Q_n^*, \mu) \le 32 d^{3/2} \sqrt{\frac{k^{1-2/d} \log n}{n}}$$

where Q_n^* is the empirically optimal quantizer.

Just like the lower bound of Theorem 1, the new upper bound is also a decreasing function of the number of codepoints k if $d = 1$. Comparing the two bounds leads to the conjecture that for very small values of d (i.e., for $d = 1$ and perhaps for $d = 2, 3, 4$) the minimax distortion redundancy is a decreasing function of k, while for large values of d it is an increasing function of k. We cannot prove this conclusion because of the gap between the upper and lower bounds, but for $d = 1$ it is possible to show values of $k_1 < k_2$ and n such that the minimax distortion redundancy for k_1 codepoints is larger than that for k_2 codepoints. Intuitively, one might expect the minimax distortion redundancy to increase with k since the number of unknown parameters (i.e., kd) is increasing with k. On the other hand, the distortion of an optimal quantizer becomes small as k increases, and "smaller" quantities can be estimated with smaller variance. (The effect is the same as encountered in estimating the parameter p based on n Bernoulli (p) random variables, where the MSE of the best unbiased estimate is $p(1-p)/n$.) Since the distortion of a vector quantizer decreases with k typically as $O(k^{-2/d})$, this effect becomes negligible for large d. This might explain why our upper bound is decreasing in k for $d = 1$ but is increasing in k for $d > 2$. The proof of Theorem 2 provides further insight. The exact dependence of the minimax distortion redundancy on k and d is still a challenging open problem.

The relatively simple proof of this result is given in Section III-B. Note that this upper bound is always better than (3) if

$$k^{2/d} > \log_2 n$$

or

$$n < 2^{2^{2R}}$$

where R is the rate of the quantizer defined by $R = (1/d) \log_2 k$. For practical values of the training set size, this condition is satisfied for medium bit rates. For example, for $n = 10^6$, the new upper bound is smaller than (3) if $R \ge 2.16$.

In recent work, Merhav and Ziv [13] studied a problem closely related to quantizer design. In their setup, the "design algorithm" is given N bits of information (called side

information bits) about the source. The question is how many side information bits are necessary and sufficient to obtain a d-dimensional rate R quantizer ($R = (1/d)\log k$, where k is the number of codepoints) whose distortion is close to the optimum. Their main result gives the answer $N = 2^{dR}$ in an exponential sense, if d is large. The sufficiency part of this statement was proved using (2). Note that this problem is more general than the problem we consider. The N information bits are allowed to represent an arbitrary description of the source, of which discretized independent training samples are a special case. While the necessity part of this result does not translate directly to a lower bound on the convergence rate we study, it does have implications on how the minimax bounds can depend on the rate R and dimension d. For example, it is not hard to see that the fact that $N = 2^{d(R-\epsilon)}$ side information bits are not enough implies that the minimax distortion redundancy convergence rate cannot be upper-bounded in the form $c(2^{d(R-\epsilon)}/n)^{\delta}$ for any constants $c, \epsilon, \delta > 0$.

Our setting is slightly different from that studied in [13]. While Merhav and Ziv concentrated on stationary and ergodic sources, we only restrict the distribution to have support in a bounded subset of $\mathcal{R}^d$. It is not hard to see that in general there does not exist a real stationary process whose d-dimensional marginals have exactly our counterexample distribution. We presently do not see a way of constructing stationary and ergodic sources (as was done in [13] for determining the number of necessary side information bits) whose d-dimensional marginals approximate the counterexample distributions well enough so that the rather fine analysis of the lower bound carries over without destroying the $n^{-1/2}$ rate.

Finally, we would like to point out that our formulation of minimax redundancy has close connections with universal lossy coding. In particular, following Davisson's [5] definitions of various types of universality for lossless coding, Neuhoff *et al.* [14] defined three main types of universality in fixed-rate universal lossy coding. Of these three definitions, the one called strong minimax universality parallels our minimax redundancy formulation. A sequence of fixed-rate block codes is called *strongly minimax universal* with respect to a given class of sources if the distortion and rate of the codes converge with increasing blocklength to their respective OPTA (optimal performance theoretically attainable) functions *uniformly* over the source class. Thus by choosing sufficiently large blocklength for a strongly minimax universal code, one can achieve a preassigned level of performance regardless of which source in the class is encoded. In our case, the minimax distortion redundancy $J^*(n, k, d)$ converges to zero with increasing n if and only if there exists a sequence of empirically designed quantizers Q_n such that $J(Q_n, \mu)$ converges to zero uniformly over all μ in the given source class. The implication is similar to the universal coding case; by choosing the number of training samples large enough, the distortion redundancy of the empirically designed quantizer will be arbitrarily small for all sources in the class.

Neuhoff *et al.* [14] also defined a weaker notion of universality. In this definition, a sequence of codes with increasing blocklength is *weakly minimax universal* with respect to a class of sources if the rate and distortion converge (not necessarily uniformly) to their OPTA functions for each source in the class. Refining this definition, Shields [20] defined the notion of weak minimax convergence rates in universal coding. Using Shield's formulation, we can define weak minimax convergence rates in empirical quantizer design in the following way.

A nondecreasing positive function $n \to f(n)$ is called a *weak rate* for empirical quantizer design for a class of d-dimensional sources $\mathcal{P}$ if the following simultaneously hold.

i) There exists a sequence of k-point empirical quantizers $\{Q_n\}$ such that for each $\mu \in \mathcal{P}$ there is a finite number $M(\mu)$ for which

$$J(Q_n, \mu) \le M(\mu) f(n), \qquad \text{for all } n \ge 1. \tag{5}$$

ii) For any sequence of k-point empirical quantizers $\{Q_n\}$ and function $g(n) = o(f(n))$, there exists a source $\mu \in \mathcal{P}$ such that $J(Q_n, \mu)/g(n)$ is unbounded as $n \to \infty$.

Note that the constant $M(\mu)$ in (5) can depend on the source distribution μ. For this reason, the minimax lower bound in Theorem 1 *does not* imply that the weak rate for the class of sources over $S(0, \sqrt{d})$ cannot be less than $n^{-1/2}$. It is an interesting and challenging problem to find the weak rate for this source class.

III. Proofs

A. Proof of Theorem 1

The basic idea of the proof may be illustrated by the following simple example: let $d = 1$, $k = 3$, and assume that μ is concentrated on four points: $0, \epsilon, 1-\epsilon$, and 1, such that either $\mu(0) = \mu(\epsilon) = 1/4 + \delta$ and $\mu(1) = \mu(1-\epsilon) = 1/4 - \delta$, or $\mu(0) = \mu(\epsilon) = 1/4 - \delta$ and $\mu(1) = \mu(1-\epsilon) = 1/4 + \delta$. Then if ϵ is sufficiently small, the codepoints of the optimal quantizer are $0, \epsilon, 1 - \epsilon/2$ in the first case, and $\epsilon/2, 1-\epsilon, 1$ 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 ϵ, δ yields the desired $\Omega(n^{-1/2})$ lower bound for the minimax expected distortion redundancy. The general, $d > 1$, $k > 3$, case is more complicated, but the basic idea is the same.

We present the proof in several steps. Some of the technical details are given in the Appendix.

Step 1: First observe that we can restrict our attention to nearest neighbor quantizers, that is, to Q_n's with the property that for all $x_1, \cdots, x_n$, the corresponding quantizer is a nearest neighbor quantizer. This follows from the fact that for any Q_n not satisfying this property, we can find a nearest neighbor quantizer Q_n' such that for *all* μ, $J(Q_n', \mu) \le J(Q_n, \mu)$.

Step 2: Clearly,

$$\sup_{\mu} J(Q_n, \mu) \ge \sup_{\mu \in \mathcal{D}} J(Q_n, \mu)$$

where $\mathcal{D}$ is any restricted class of distributions on $S(0, \sqrt{d})$. We define $\mathcal{D}$ as follows: each member of $\mathcal{D}$ is concentrated on the set of $2m = 4k/3$ fixed points $\{z_i, z_i+w: i = 1, \cdots, m\}$, where $w = (\Delta, 0, 0, \cdots, 0)$ is a fixed d-vector, and Δ is a small positive number to be determined later. The positions of $z_1, \cdots, z_m \in S(0, \sqrt{d})$ satisfy the property that the distance between any two of them is greater than $A\Delta$, where the value of A is determined in Step 5 below. For the sake of simplicity, we assume that k is divisible by 3. (This assumption is clearly insignificant.) Let $\delta \leq 1/2$ be a positive number. For each $1 \leq i \leq m$, set

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

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 $M = \binom{m}{m/2}$. Denote the members of $\mathcal{D}$ by $\mu_1, \mu_2, \cdots, \mu_M$.

Step 3: Let $\mathcal{Q}$ denote the collection of k-point quantizers $Q \in \mathcal{Q}$ such that for $m/2$ values of $i \in \{1, \cdots, m\}$, 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 $z_i + w/2$. If $A \geq \sqrt{2/(1-\delta)} + 1$, then for any k-point quantizer Q there exists a $\tilde{Q}$ in $\mathcal{Q}$ such that, for all μ in $\mathcal{D}$, $D(\tilde{Q}) \leq D(Q)$. The proof of this is given in the Appendix.

Step 4: Consider a distribution $\mu_j \in \mathcal{D}$ and the corresponding optimal quantizer $Q^{(j)}$. Clearly, from Step 3, if $A \geq \sqrt{2/(1-\delta)} + 1$, then for the $m/2$ values of i in $\{1, \cdots, m\}$ that have $\mu_j(\{z_i, z_i + w\}) = (1 + \delta)/m$, $Q^{(j)}$ has codepoints at both z_i and $z_i + w$. For the remaining $m/2$ values of i there is a single codepoint at $z_i + w/2$.

For any distribution in $\mathcal{D}$ and any quantizer in $\mathcal{Q}$, it is easy to see that the distortion of the quantizer is between $(1-\delta)\Delta^2/8$ and $(1 + \delta)\Delta^2/8$.

Step 5: Let $\mathcal{Q}_n$ denote the family of empirically designed quantizers such that for every fixed $x_1, \cdots, x_n$, we have $Q(\cdot, x_1, \cdots, x_n) \in \mathcal{Q}$. Since $\delta \leq 1/2$, the property of the optimal quantizer described in Step 4 is always satisfied if we take $A = 3$. In particular, if $A = 3$, we have

$$\inf_{Q_n} \max_{\mu \in \mathcal{D}} J(Q_n, \mu) = \min_{Q_n \in \mathcal{Q}_n} \max_{\mu \in \mathcal{D}} J(Q_n, \mu)$$

and it suffices to lower-bound the quantity on the right-hand side.

Step 6: Let Z be a random variable which is uniformly distributed on the set of integers $\{1, 2, \cdots, M\}$. Then, for any Q_n, we obviously have

$$\max_{\mu \in \mathcal{D}} J(Q_n, \mu) \geq \boldsymbol{E} J(Q_n, \mu_Z) = \frac{1}{M} \sum_{i=1}^{M} J(Q_n, \mu_i).$$

Step 7:

$$\min_{Q_n \in \mathcal{Q}_n} \boldsymbol{E} J(Q_n, \mu_Z) = \boldsymbol{E} J(Q_n^*, \mu_Z) \tag{6}$$

where Q_n^* is the "empirically optimal" (or "maximum-likelihood") quantizer from $\mathcal{Q}$, that is, if N_i denotes the number of X_i's falling in $\{z_i, z_i + w\}$, then Q_n^* has a codepoint at both z_i and $z_i + w$ if the corresponding N_i is one of the $m/2$ largest values. For the other i's (i.e., those with the $m/2$ smallest N_i's) Q_n^* has a codepoint at $z_i + w/2$.

The proof is given in the Appendix.

Step 8: By symmetry, we have

$$\boldsymbol{E} J(Q_n^*, \mu_Z) = J(Q_n^*, \mu_1).$$

The rest of the proof involves bounding $J(Q_n^*, \mu_1)$ from below, where Q_n^* is the empirically optimal quantizer.

Step 9: Recall that the vector of random integers $(N_1, \cdots, N_m)$ is multinomially distributed with parameters $(n, q_1, \cdots, q_m)$, where

$$q_1 = q_2 = \cdots = q_{m/2} = (1 - \delta)/m$$

and

$$q_{m/2+1} = \cdots = q_m = (1 + \delta)/m.$$

Let $N_{\sigma(1)}, \cdots, N_{\sigma(m)}$ be a reordering of the N_i's such that $N_{\sigma(1)} \leq N_{\sigma(2)} \leq \cdots \leq N_{\sigma(m)}$. (In case of equal values, break ties according to indices.) Let p_j $(j = 1, \cdots, m/2)$ be the probability of the event that among $N_{\sigma(1)}, \cdots, N_{\sigma(m/2)}$, there are exactly j of the N_i's with $i > m/2$ (i.e., the "maximum-likelihood" estimate makes j mistakes). Then it is easy to see that

$$J(Q_n^*, \mu_1) = \frac{\Delta^2 \delta}{2m} \sum_{j=1}^{m/2} j p_j$$

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

Step 10: From now on, we investigate the quantity

$$\sum_{j=1}^{m/2} j p_j$$

that is, the expected number of mistakes. First we use the trivial bound

$$\sum_{j=1}^{m/2} j p_j \geq j_0 \sum_{j=j_0}^{m/2} p_j$$

with j_0 to be chosen later. $\sum_{j=j_0}^{m/2} p_j$ is the probability that the maximum-likelihood decision makes at least j_0 mistakes. The key observation is that this probability may be bounded below by the probability that at least $2j_0$ of the events $A_1, \cdots, A_{m/2}$ hold, where

$$A_i = \{N_i > N_{m/2+i}\}.$$

In other words,

$$\sum_{j=j_0}^{m/2} p_j \geq \boldsymbol{P}\left\{ \sum_{j=1}^{m/2} I_{A_i} \geq 2j_0 \right\}.$$

Proof: Define the following sets of indices:

$$S_1 = \{i: \sigma(i) \le m/2, \quad i \ge m/2+1\},$$
$$S_2 = \{i: \sigma(i) \le m/2, \quad i \le m/2\}.$$

Then the maximum-likelihood decision makes $|S_1|$ mistakes. If $i \in S_2$ and $N_i > N_{m/2+i}$, then $m/2+i \in S_1$. Thus the number of indices i for which $N_i > N_{m/2+i}$ is bounded from above by $|S_1|+m/2-|S_2|=2|S_1|$, since $|S_2|=m/2-|S_1|$.□

Step 11: Thus we need a lower bound on the tail of the distribution of the random variable $\sum_{j=1}^{m/2} I_{A_i}$. First we obtain a suitable lower bound for its expected value.

$$\boldsymbol{E}\left[\sum_{j=1}^{m/2} I_{A_i}\right] = \frac{m}{2}\boldsymbol{P}\{A_1\}. \tag{7}$$

Now, bounding $\boldsymbol{P}\{A_1\}$ conservatively, we have

$$\begin{aligned}\boldsymbol{P}\{A_1\} &= \boldsymbol{P}\{N_1 > N_{m/2+1}\}\\ &\ge \boldsymbol{P}\{N_1 > n/m \text{ and } N_{m/2+1} \le n/m\}\\ &= \boldsymbol{P}\{N_1 > n/m\} - \boldsymbol{P}\{N_1 > n/m \text{ and } N_{m/2+1} > n/m\}\\ &\ge \boldsymbol{P}\{N_1 > n/m\} - \boldsymbol{P}\{N_1 > n/m\}\\ &\quad \cdot \boldsymbol{P}\{N_{m/2+1} > n/m\}\\ &= \boldsymbol{P}\{N_1 > n/m\}\boldsymbol{P}\{N_{m/2+1} \le n/m\}.\end{aligned}$$

The last inequality follows by Mallows' inequality (see Mallows [12]) which states that if $(N_1, \cdots, N_m)$ are multinomially distributed, then

$$\boldsymbol{P}\{N_1 > t_1, N_2 > t_2, \cdots, N_m > t_m\} \le \prod_{i=1}^{m} \boldsymbol{P}\{N_i > t_i\}.$$

Finally, we approximate the last two binomial probabilities by normals. To this end, we use the Berry–Esséen inequality (see, e.g., Chow and Teicher [4]), which states that if $Z_1, \cdots, Z_n$ are i.i.d. random variables with $\boldsymbol{E}Z_1 = 0$, $\boldsymbol{E}[Z_1^2] = \sigma^2$, and $\boldsymbol{E}[|Z_1|^3] = \gamma$, then

$$\left|\boldsymbol{P}\left\{\sum_{i=1}^{n} Z_i < x\sigma\sqrt{n}\right\} - \Phi(x)\right| \le \frac{\gamma}{\sigma^3\sqrt{n}}$$

where Φ is the distribution function of a standard normal random variable. Choose $\delta = \sqrt{m/n}$. Observe that N_1 is the sum of n i.i.d. Bernoulli$((1-\delta)/m)$ random variables. Then the Berry–Esséen inequality implies that if $n \ge 8m/\Phi(-2)^2$, then

$$\boldsymbol{P}\{N_1 > n/m\} \ge \Phi(-2)/2$$

and similarly

$$\boldsymbol{P}\{N_{m/2+1} \le n/m\} \ge \Phi(-2)/2.$$

.

Therefore, by (7) we get

$$\boldsymbol{E}\left[\sum_{j=1}^{m/2} I_{A_i}\right] \ge \frac{m\Phi(-2)^2}{8}. \tag{8}$$

Step 12: To obtain the desired lower bound for

$$\boldsymbol{P}\left\{\sum_{j=1}^{m/2} I_{A_i} \ge 2j_0\right\}$$

we use the following elementary inequality: if the random variable Z satisfies $\boldsymbol{P}\{Z \in [0, B]\} = 1$, then

$$\boldsymbol{P}\left\{Z \ge \frac{\boldsymbol{E}Z}{2}\right\} \ge \frac{\boldsymbol{E}Z}{2B}. \tag{9}$$

To see this, notice that for α in $[0, B]$

$$\boldsymbol{E}Z \le \alpha + B\boldsymbol{P}\{Z \ge \alpha\}$$

and substitute $\alpha = \boldsymbol{E}Z/2$.

Step 13: To apply this inequality, choose $j_0 = m\Phi(-2)^2/32$. Then (8) implies that

$$2j_0 \le (1/2)\boldsymbol{E}\left[\sum_{j=1}^{m/2} I_{A_i}\right]$$

and, therefore,

$$\begin{aligned}\boldsymbol{P}\left\{\sum_{j=1}^{m/2} I_{A_i} \ge 2j_0\right\} &\ge \boldsymbol{P}\left\{\sum_{j=1}^{m/2} I_{A_i} \ge \tfrac{1}{2}\boldsymbol{E}\left[\sum_{j=1}^{m/2} I_{A_i}\right]\right\}\\ &\ge \frac{1}{m}\boldsymbol{E}\left[\sum_{j=1}^{m/2} I_{A_i}\right]\\ &\ge \frac{\Phi(-2)^2}{8}\end{aligned}$$

where the second inequality follows from (9) and the last inequality follows from (8).

Step 14: Collecting everything, we have that

$$\inf_{Q_n}\sup_{\mu} J(Q_n, \mu) \ge \frac{\Delta^2\Phi(-2)^4}{512}\sqrt{\frac{m}{n}}$$

where Δ is any positive number with the property that m pairs of points $\{z_i, z_i + w\}$ can be placed in $S(0, \sqrt{d})$ such that the distance between any two of the z_i's is at least 3Δ. In other words, to make Δ large, we need find a (desirably large) Δ such that m points $z_1, \cdots, z_m$ can be packed into the ball $S(0, \sqrt{d}-\Delta)$. (We decrease the radius of the ball by Δ to make sure that the $(z_i + w)$'s also fall in the ball $S(0, \sqrt{d})$.) Thus we need a good lower bound for the cardinality of the maximal 3Δ-packing of $S(0, \sqrt{d}-\Delta)$. It is well known (see Kolmogorov and Tikhomirov [8]) that the cardinality of the maximal packing is lower-bounded by the cardinality of the minimal covering, that is, by the minimal number of balls of radius 3Δ whose union covers $S(0, \sqrt{d}-\Delta)$. But this number is clearly bounded from below by the ratio of the volume of $S(0, \sqrt{d}-\Delta)$, and that of $S(0, 3\Delta)$. Therefore, m points can certainly be packed in $S(0, \sqrt{d}-\Delta)$ as long as

$$m \le \left(\frac{\sqrt{d}-\Delta}{3\Delta}\right)^d.$$

If $\Delta \le \sqrt{d}/4$ (which is satisfied by our choice of Δ below), the above inequality holds if

$$m \le \left(\frac{\sqrt{d}}{4\Delta}\right)^d.$$

Thus the choice

$$\Delta = \frac{\sqrt{d}}{4m^{1/d}}$$

satisfies the required property. Resubstitution of this value proves the theorem. □

B. Proof of Theorem 2

The first step in the analysis of the performance of the empirical quantizer Q_n^* is the following lemma.

Lemma 1: Let $S(x, r)$ denote the closed d-dimensional sphere of radius r centered at x. Let $\rho > 0$ and let $N(\rho)$ denote the cardinality of the minimum ρ covering of $S(0, r)$, that is, $N(\rho)$ is the smallest integer N such that there exist points $\{y_1, \cdots, y_N\} \subset S(0, r)$ with the property

$$\max_{x \in S(0,r)} \min_{1 \le i \le N} \|x - y_i\| \le \rho. \tag{10}$$

Then, for all $\rho \le 2r$ we have

$$N(\rho) \le \left(\frac{4r}{\rho}\right)^d.$$

Proof: By a classical observation of Kolmogorov and Tikhomirov [8] the covering (10) exists if it is impossible to construct another set $\{z_1, \cdots, z_{N+1}\} \subset S(0, r)$ which is ρ-separated, that is,

$$\min_{\substack{i \ne j \\ 1 \le i, j \le N+1}} \|z_i - z_j\| \ge \rho. \tag{11}$$

Let us now consider an arbitrary ρ-separated set of cardinality $N+1$. Then the open balls of radius $\rho/2$ centered at the z_i are disjoint and their union is included in $S(0, r + \rho/2)$. Also, if $\rho/2 \le r$, then $S(0, r + \rho/2) \subset S(0, 2r)$. Thus such a separating set cannot exist as long as $N + 1$ is greater than the ratio of the volumes of $S(0, 2r)$ and $S(0, \rho/2)$, that is,

$$N > \left(\frac{4r}{\rho}\right)^d - 1.$$

Since there exists an integer $N \le (4r/\rho)^d$ which satisfies the above inequality, the lemma is proved. □

Corollary 1: Let $0 < \epsilon \le 8d$. There exists a finite collection of k-point quantizers $\mathcal{Q}_\epsilon$ such that

i) the cardinality of $\mathcal{Q}_\epsilon$ is bounded as

$$|\mathcal{Q}_\epsilon| \le \left(\frac{16d}{\epsilon}\right)^{kd};$$

ii) all quantizers in $\mathcal{Q}_\epsilon$ have their codepoints inside $S(0, \sqrt{d})$;

iii) for any k-point nearest neighbor quantizer Q whose codepoints are contained in $S(0, \sqrt{d})$, there exists a $Q' \in \mathcal{Q}_\epsilon$ such that for all $x \in S(0, \sqrt{d})$,

$$| \|x - Q'(x)\|^2 - \|x - Q(x)\|^2 | \le \epsilon.$$

Proof: Let $\rho = \epsilon/(4\sqrt{d})$. Then $0 < \rho \le 2\sqrt{d}$, and by Lemma 1 there exists a ρ-covering set of points $\{y_1, \cdots, y_N\} \subset S(0, \sqrt{d})$ if $N \le (4\sqrt{d}/\rho)^d$. Define $\mathcal{Q}_\epsilon$ as the collection of all k-point nearest neighbor quantizers whose codepoints are from the covering set $\{y_1, \cdots, y_N\}$. Then

$$|\mathcal{Q}_\epsilon| \le \left(\frac{4\sqrt{d}}{\rho}\right)^{kd} = \left(\frac{16d}{\epsilon}\right)^{kd}.$$

If $\{x_1, \cdots, x_k\}$ are the codepoints of Q, then there exists a quantizer $Q' \in \mathcal{Q}_\epsilon$ with codepoints $\{x_1', \cdots, x_k'\}$ such that $\|x_i - x_i'\| \le \rho$ for all i. If $Q(x) = x_j$, we have by the nearest neighbor property that

$$\begin{aligned} \|x - Q'(x)\|^2 - \|x - Q(x)\|^2 &\le \|x - x_j'\|^2 - \|x - x_j\|^2 \\ &\le 4\sqrt{d}\|x_j' - x_j\| \\ &\le \epsilon. \end{aligned}$$

The inequality $\|x - Q(x)\|^2 - \|x - Q'(x)\|^2 \le \epsilon$ may be proved similarly. □

Corollary 2: For all distributions such that $\boldsymbol{P}\{\|X\| \le \sqrt{d}\} = 1$, there exists a k-point quantizer $(k \ge 1)$ whose codepoints are contained in $S(0, \sqrt{d})$ and whose distortion satisfies

$$D(Q) \le 16\, dk^{-2/d}.$$

Proof: If $k \le 2^d$, then the statement trivially holds for the quantizer having one codepoint at the origin. Otherwise, let $\rho = 4\sqrt{d}k^{-1/d}$. Then $\rho \le 2\sqrt{d}$ and by Lemma 1 there exists a set of points $\{y_1, \cdots, y_k\} \subset S(0, \sqrt{d})$ that ρ-covers $S(0, \sqrt{d})$. Letting Q be the nearest neighbor quantizer with these codepoints, we get $D(Q) \le \rho^2 = 16\, dk^{-2/d}$. □

Let $0 < \epsilon \le 8d$, and let $\mathcal{Q}_\epsilon$ be a set of quantizers satisfying properties i), ii), and iii) of Corollary 1. Let $\hat{Q} \in \mathcal{Q}_\epsilon$ denote a quantizer whose distortion is minimal in $\mathcal{Q}_\epsilon$, that is,

$$D(\hat{Q}) \le D(Q), \qquad \text{for all } Q \in \mathcal{Q}_\epsilon.$$

Then it is clear that $D(\hat{Q}) \le D^* + \epsilon$, where D^* denotes the minimum distortion achievable by any quantizer. Let Q_n be a quantizer in $\mathcal{Q}_\epsilon$ such that for all $x \in S(0, \sqrt{d})$

$$\|x - Q_n(x)\|^2 \le \|x - Q_n^*(x)\|^2 + \epsilon.$$

Such a quantizer exists by Corollary 1. Then clearly, by the definition of the empirically optimal quantizer Q_n^*

$$D_n(Q_n) \le D_n(Q) + \epsilon, \qquad \text{for all } Q \in \mathcal{Q}_\epsilon.$$

The next lemma is based on ideas of Vapnik and Chervonenkis [22].

Lemma 2: For all $\delta > \epsilon$, we have

$$\begin{aligned}&\boldsymbol{P}\{D(Q_n) - D(\hat{Q}) > 2\delta\}\\ &\quad \le \boldsymbol{P}\{D_n(\hat{Q}) - D(\hat{Q}) > \delta - \epsilon\}\\ &\qquad + \boldsymbol{P}\left\{\max_{Q\in\mathcal{Q}_\epsilon} \frac{D(Q) - D_n(Q)}{\sqrt{D(Q)}} > \frac{\delta}{\sqrt{D(\hat{Q}) + 2\delta}}\right\}.\end{aligned}$$

Proof: If

$$\max_{Q\in\mathcal{Q}_\epsilon} \frac{D(Q) - D_n(Q)}{\sqrt{D(Q)}} \le \frac{\delta}{\sqrt{D(\hat{Q}) + 2\delta}}$$

then for each $Q \in \mathcal{Q}_\epsilon$

$$D_n(Q) \ge D(Q) - \delta\sqrt{\frac{D(Q)}{D(\hat{Q}) + 2\delta}}.$$

If, in addition, Q is such that $D(Q) > D(\hat{Q}) + 2\delta$, then by the monotonicity of the function $x - c\sqrt{x}$ (for $c > 0$ and $x > c^2/4$)

$$D_n(Q) > D(\hat{Q}) + 2\delta - \delta\sqrt{\frac{D(\hat{Q}) + 2\delta}{D(\hat{Q}) + 2\delta}} = D(\hat{Q}) + \delta.$$

Therefore,

$$\begin{aligned}&\boldsymbol{P}\left\{\min_{Q:\, D(Q) > D(\hat{Q}) + 2\delta} D_n(Q) \le D(\hat{Q}) + \delta\right\}\\ &\qquad \le \boldsymbol{P}\left\{\max_{Q\in\mathcal{Q}_\epsilon} \frac{D(Q) - D_n(Q)}{\sqrt{D(Q)}} > \frac{\delta}{\sqrt{D(\hat{Q}) + 2\delta}}\right\}.\end{aligned}$$

But if $D(Q_n) - D(\hat{Q}) > 2\delta$, then there exists an $Q \in \mathcal{Q}_\epsilon$ such that $D(Q) > D(\hat{Q}) + 2\delta$ and $D_n(Q) \le D_n(\hat{Q}) + \epsilon$. Thus

$$\begin{aligned}&\boldsymbol{P}\{D(Q_n) - D(\hat{Q}) > 2\delta\}\\ &\quad \le \boldsymbol{P}\left\{\min_{Q:\, D(Q) > D(\hat{Q}) + 2\delta} D_n(Q) \le D_n(\hat{Q}) + \epsilon\right\}\\ &\quad \le \boldsymbol{P}\left\{\min_{Q:\, D(Q) > D(\hat{Q}) + 2\delta} D_n(Q) \le D(\hat{Q}) + \delta\right\}\\ &\qquad + \boldsymbol{P}\{D_n(\hat{Q}) > D(\hat{Q}) + \delta - \epsilon\}\\ &\quad \le \boldsymbol{P}\left\{\max_{Q\in\mathcal{Q}_\epsilon} \frac{D(Q) - D_n(Q)}{\sqrt{D(Q)}} > \frac{\delta}{\sqrt{D(\hat{Q}) + 2\delta}}\right\}\\ &\qquad + \boldsymbol{P}\{D_n(\hat{Q}) - D(\hat{Q}) > \delta - \epsilon\}.\end{aligned}$$

□

Lemma 3: Let $Q \in \mathcal{Q}_\epsilon$. Then for all $\gamma > 0$

$$\boldsymbol{P}\left\{\frac{D(Q) - D_n(Q)}{\sqrt{D(Q)}} > \gamma\right\} \le e^{-3n\gamma^2/(32d)}.$$

Proof: The probability is clearly zero if $\gamma > \sqrt{D(Q)}$. For $\gamma \le \sqrt{D(Q)}$, we may use Bernstein's inequality [2]

$$\begin{aligned}&\boldsymbol{P}\left\{D(Q) - D_n(Q) > \gamma\sqrt{D(Q)}\right\}\\ &\qquad \le e^{-[n\gamma^2 D(Q)/2\sigma^2 + (2/3)4d\gamma\sqrt{D(Q)}]}\end{aligned}$$

where $\sigma^2 = \mathbf{var}\,(\|X - Q(X)\|^2)$. But observe that

$$\|X - Q(X)\|^2 \le 4d$$

with probability one, and, therefore, $\sigma^2 \le 4dD(Q)$, and the statement follows. □

Corollary 3: For all $\delta > \epsilon$

$$\begin{aligned}&\boldsymbol{P}\{D(Q_n) - D(\hat{Q}) > 2\delta\}\\ &\qquad \le (|\mathcal{Q}_\epsilon| + 1)e^{-3n(\delta-\epsilon)^2/(32d(D(\hat{Q}) + 2(\delta-\epsilon)))}.\end{aligned}$$

Proof: By Lemma 3 we have

$$\begin{aligned}&\boldsymbol{P}\left\{\max_{Q\in\mathcal{Q}_\epsilon} \frac{D(Q) - D_n(Q)}{\sqrt{D(Q)}} > \frac{\delta}{\sqrt{D(\hat{Q}) + 2\delta}}\right\}\\ &\quad \le |\mathcal{Q}_\epsilon| \max_{Q\in\mathcal{Q}_\epsilon} \boldsymbol{P}\left\{\frac{D(Q) - D_n(Q)}{\sqrt{D(Q)}} > \frac{\delta}{\sqrt{D(\hat{Q}) + 2\delta}}\right\}\\ &\quad \le |\mathcal{Q}_\epsilon| e^{-3n\delta^2/(32d(D(\hat{Q}) + 2\delta))}\\ &\quad \le |\mathcal{Q}_\epsilon| e^{-3n(\delta-\epsilon)^2/(32d(D(\hat{Q}) + 2(\delta-\epsilon)))}.\end{aligned}$$

On the other hand, by Bernstein's inequality

$$\boldsymbol{P}\{D_n(\hat{Q}) - D(\hat{Q}) > \delta - \epsilon\} \le e^{-n(\delta-\epsilon)^2/(8dD(\hat{Q}) + 8d(\delta-\epsilon)/3)}$$

and applying Lemma 2 finishes the proof. □

Proof of Theorem 2: Since the distribution of X is supported on $S(0, \sqrt{d})$, we have that with probability one, $D(Q_n) - D(\hat{Q}) \le 4d$, hence for every $u > 0$

$$\boldsymbol{E}D(Q_n) - D(\hat{Q}) \le u + 4d\boldsymbol{P}\{D(Q_n) - D(\hat{Q}) > u\}.$$

Thus it follows from Corollary 3 that for any $u > \epsilon$

$$\boldsymbol{E}D(Q_n) - D(\hat{Q}) \le u + 8d|\mathcal{Q}_\epsilon| e^{-3n(u-\epsilon)^2/(32d(D(\hat{Q}) + 2(u-\epsilon)))}.$$

If $D(\hat{Q}) \ge (32d\,\log(8d|\mathcal{Q}_\epsilon|\sqrt{n})/n)$, then with

$$u = \sqrt{\frac{32dD(\hat{Q})\,\log(8d|\mathcal{Q}_\epsilon|\sqrt{n})}{n}} + \epsilon$$

we have $u - \epsilon \le D(\hat{Q})$. In such a case

$$\begin{aligned}\boldsymbol{E}D(Q_n) - D(\hat{Q}) &\le u + 8d|\mathcal{Q}_\epsilon| e^{-n(u-\epsilon)^2/(32dD(\hat{Q}))}\\ &= \sqrt{\frac{32dD(\hat{Q})\,\log(8d|\mathcal{Q}_\epsilon|\sqrt{n})}{n}} + \frac{1}{\sqrt{n}} + \epsilon.\end{aligned}$$

On the other hand, if $D(\hat{Q}) < 32d\,\log(8d|\mathcal{Q}_\epsilon|\sqrt{n})/n$, then take

$$u = \frac{32d\,\log(8d|\mathcal{Q}_\epsilon|n)}{n} + \epsilon.$$

Then $D(\hat{Q}) < u - \epsilon$, and, therefore,

$$\begin{aligned}\boldsymbol{E}D(Q_n) - D(\hat{Q}) &\le u + 8d|\mathcal{Q}_\epsilon|e^{-n(u-\epsilon)/(32d)} \\ &= \frac{32d\log(8d|\mathcal{Q}_\epsilon|n)+1}{n} + \epsilon.\end{aligned}$$

Noting that $\boldsymbol{E}D(Q_n^*) \le \boldsymbol{E}D(Q_n) + \epsilon$ and $D(\hat{Q}) - D^* \le \epsilon$, we obtain

$$\boldsymbol{E}D(Q_n^*) - D^* \le 3\epsilon + \max\left(\sqrt{\frac{32dD(\hat{Q})\log(8d|\mathcal{Q}_\epsilon|\sqrt{n})}{n}} + \frac{1}{\sqrt{n}}, \frac{32d\log(8d|\mathcal{Q}_\epsilon|n)+1}{n}\right).$$

Take $\epsilon = 16dn^{-1/2}$, and also recall that by Corollary 2

$$D(\hat{Q}) \le D^* + \epsilon \le 16dk^{-2/d} + \frac{16d}{\sqrt{n}} \le 32dk^{-2/d}$$

whenever $n \ge k^{4/d}$. Substituting these values into the above inequality, we obtain the inequalities shown at the bottom of this page, if $\sqrt{dk^{1-2/d}\log n} \ge 15$, $kd \ge 8$, and $n \ge 8d$. In particular, if $n/\log n \ge dk^{1+2/d}$, then

$$J(Q_n^*, \mu) \le 32d^{3/2}\sqrt{\frac{k^{1-2/d}\log n}{n}}. \qquad \square$$

IV. Concluding Remarks

The main results of the paper are new upper and lower bounds for the minimax expected distortion redundancy of empirical quantizers. Combining these with previously known bounds we see that for some universal constants $c_0, c_1 > 0$

$$\begin{aligned}c_0 d\sqrt{\frac{k^{1-4/d}}{n}} &\le J^*(n, k, d) \\ &\le c_1 d^{3/2}\sqrt{\frac{k^{1-2/d}}{n}} \\ &\quad \cdot \min\left(\sqrt{\log n}, \sqrt{k^{2/d}\log(kd)}\right).\end{aligned}$$

For most practical values of the dimension d, the number of codepoints k, and the number of training vectors n, the two bounds are fairly close to each other, essentially describing the behavior of the minimax distortion. For example, it follows that the minimax distortion redundancy, as a function of the number of training samples n, is on the order of $n^{-1/2}$. Also, if $k = 2^{dR}$ for a constant rate R, we obtain that the *per-dimension* minimax distortion redundancy is approximately

$$\sqrt{\frac{2^{dR}}{n}}$$

for large d and n.

However, some interesting questions remain unanswered. We conjecture that the factor of $\sqrt{\log n}$ in the upper bound of Theorem 2 might be eliminated, and the minimax expected distortion redundancy is some constant times

$$d^a\sqrt{\frac{k^{1-b/d}}{n}}$$

for some values of $a \in [1, 3/2]$ and $b \in [2, 4]$.

Another challenging problem is to find (or give bounds on) the weak minimax convergence rate defined at the end of Section II. In particular, Pollard's result [16] suggests that the weak minimax rate can still be $O(1/n)$ for a class of sources with sufficiently regular and smooth densities. We have no conjecture at present, however, as to what the weak rate might be for the class of all sources concentrated on $S(0, \sqrt{d})$.

Appendix

Proof of Step 3: Let $C = \{y_1, \cdots, y_k\}$ be the codebook of Q. Consider the Voronoi partition of $\mathcal{R}^d$ induced by the set of points $\{z_i, z_i + w;\ 1 \le i \le m\}$ and for each i define V_i as the union of the two Voronoi cells belonging to z_i and $z_i + w$. Furthermore, let m_i be the cardinality of $C \cap V_i$. A new nearest neighbor quantizer $\hat{Q}$ with codebook $\hat{C}$ is constructed as follows. Start with $\hat{C}$ empty. For all i

- if $m_i \ge 2$, put z_i and $z_i + w$ into $\hat{C}$,
- if $m_i = 1$ or $m_i = 0$, put $z_i + w/2$ into $\hat{C}$.

Note that $\hat{C}$ may contain more than k codepoints, but this will be fixed later. Define

$$\begin{aligned}D_i(Q) = &\|z_i - Q(z_i)\|^2\mu(\{z_i\}) \\ &+ \|z_i + w - Q(z_i + w)\|^2\mu(\{z_i + w\}).\end{aligned}$$

Then we have the following:

- if $m_i \ge 2$, then $D_i(\hat{Q}) = 0$ so that $D_i(Q) \ge D_i(\hat{Q})$;
- if $m_i = 1$, then there are two cases:
  1) $Q(z_i) = Q(z_i + w) \in V_i$. Then $D_i(Q) \ge D_i(\hat{Q})$ since $\hat{Q}(z_i) = \hat{Q}(z_i + w) = z_i + w/2$ is the optimal choice with the condition that both z_i and $z_i + w$ are mapped into the same codepoint;

$$\begin{aligned}\boldsymbol{E}D(Q_n) - D^* &\le \frac{48d}{\sqrt{n}} + \max\left(\sqrt{\frac{16kd^2D(\hat{Q})\log n + 16dD(\hat{Q})\log n + 32dD(\hat{Q})\log(8d)}{n}} + \frac{1}{\sqrt{n}}, \right. \\ &\qquad \left. \frac{16kd^2\log n + 32d\log n + 32d\log(8d) + 1}{n}\right) \\ &\le \max\left(32d^{3/2}\sqrt{\frac{k^{1-2/d}\log n}{n}}, \frac{32kd^2\log n}{n}\right)\end{aligned}$$

2) either z_i or $z_i + w$ is mapped by Q to a codepoint outside V_i. Say $Q(z_i) \notin V_i$. Then

$$D_i(Q) \geq \frac{1 \pm \delta}{2m} \|Q(z_i) - z_i\|^2 \geq \frac{1 \pm \delta}{2m} \left(\frac{(A-1)\Delta}{2} \right)^2$$

where the second inequality follows by the triangle inequality. (Here $\pm$ means $+$ if μ puts mass $(1+\delta)/m$ on $\{z_i, z_i + w\}$, and $-$ otherwise.) On the other hand, $D_i(\hat{Q}) = (1 \pm \delta)\Delta^2/(4m)$ so that $D_i(Q) \geq D_i(\hat{Q})$ if $A \geq \sqrt{2} + 1$;

- if $m_i = 0$, then both $Q(z_i)$ and $Q(z_i + w)$ are outside V_i. Thus

$$D_i(Q) \geq \frac{1 \pm \delta}{m} \left(\frac{(A-1)\Delta}{2} \right)^2$$

which implies

$$D_i(Q) \geq D_i(\hat{Q}) + \frac{1 \pm \delta}{m} \left(\frac{(A-1)\Delta}{2} \right)^2 - \frac{1 \pm \delta}{m} \frac{\Delta^2}{4} \tag{12}$$

so that $D_i(Q) \geq D_i(\hat{Q})$ if $A \geq 2$.

Thus we conclude that $D(Q) \geq D(\hat{Q})$, and we are done if $\hat{C}$ has no more than k codepoints. If $\hat{C}$ contains $\hat{k} > k$ codepoints, pick $\hat{k} - k$ arbitrary pairs $\{z_i, z_i + w\} \in \hat{C}$ and replace them with the corresponding codepoint $z_i + w/2$. We thus obtain a nearest neighbor quantizer $\tilde{Q}$. Each such replacement increases the distortion by no more than $(1 + \delta)\Delta^2/(4m)$, so that

$$D(\tilde{Q}) \leq D(\hat{Q}) + (\hat{k} - k) \frac{(1+\delta)\Delta^2}{4m}.$$

On the other hand, there must be $\hat{k} - k$ indices i for which $m_i = 0$. For each of these (12) holds, so that

$$D(\hat{Q}) \leq D(Q) - (\hat{k} - k) \frac{1 - \delta}{m} \frac{\Delta^2}{4} ((A-1)^2 - 1).$$

Therefore,

$$D(\tilde{Q}) \leq D(Q) + (\hat{k} - k) \frac{\Delta^2}{4m} ((1+\delta) - (1-\delta)((A-1)^2 - 1))$$

and this is no more than $D(Q)$ if $A \geq \sqrt{2/(1-\delta)} + 1$. □

Proof of Step 7: Let $(Y, Y_1, \cdots, Y_n)$ be jointly distributed as the mixture

$$\frac{1}{M} \sum_{i=1}^{M} \mu_i^{n+1}$$

where μ_i^{n+1} is the $(n+1)$-fold product of μ_i. Then for any Q_n

$$\boldsymbol{E}J(Q_n, \mu_Z) = \boldsymbol{E}(\|Y - Q_n(Y, Y_1, \cdots, Y_n)\|^2) - \frac{(1-\delta)\Delta^2}{8}.$$

Since $Y, Y_1, \cdots, Y_n$ are exchangeable random variables, the distribution of Y given $(Y_1, \cdots, Y_n)$ depends only on the empirical counts $(N_1, \cdots, N_k)$. It follows that the empirical quantizer Q_n achieving the minimum in (6) chooses its codebook as a function of the vector $(N_1, \cdots, N_m)$. Thus it suffices to restrict our attention to empirical quantizers that choose their codebook only as a function of $(N_1, \cdots, N_m)$. Recall that each quantizer in $\mathcal{Q}$ is such that for each i it either has one codepoint at $z_i + w/2$ or has codepoints at both z_i and $z_i + w$. Since $k = 3m/2$, there must be $m/2$ codepoints of the first kind, and m of the second.

We will represent the distribution μ_Z as an m-vector, $\gamma = (\gamma_1, \cdots, \gamma_m) \in \Gamma_m \subset \{-1, 1\}^m$, with

$$\mu_Z(\{z_i, z_i + w\}) = (1 + \gamma_i \delta)/m$$

where

$$\Gamma_m = \left\{ \gamma \in \{-1, 1\}^m : \sum_{i=1}^{m} \gamma_i = 0 \right\}.$$

We write $\boldsymbol{P}_{\gamma, n}(E)$ to denote the probability of the event E under the multinomial distribution with parameters $(n, q_1, \cdots, q_m)$ where

$$q_i = \frac{1 + \gamma_i \delta}{\sum_{j=1}^{m} (1 + \gamma_j \delta)}.$$

We will represent a quantizer's choice of the codebook as a vector $\alpha = (\alpha_1, \cdots, \alpha_m) \in \Gamma_m$, with $\alpha_i = -1$ indicating one codepoint at $z_i + w/2$ and $\alpha_i = 1$ indicating codepoints at both z_i and $z_i + w$.

Represent the quantizer $Q_n^*(\cdot, X_1, \cdots, X_n)$ by

$$\alpha^*(N_1, \cdots, N_m) \in \Gamma_m$$

for the corresponding values of N_i. Define α similarly in terms of Q_n. Then it suffices to show that (with suitable abuse of notation)

$$\sum_{\gamma \in \Gamma_m} (D(\alpha(n_1, \cdots, n_m)) - D(\alpha^*(n_1, \cdots, n_m))) \boldsymbol{P}_{\gamma, n}(\forall i,\ N_i = n_i) \geq 0$$

for all m-tuples of nonnegative integers $(n_1, \cdots, n_m)$ that sum to n and for all functions α.

For the numbers $n_1, \cdots, n_m$, let $\alpha = \alpha(n_1, \cdots, n_m)$ and $\alpha^* = \alpha^*(n_1, \cdots, n_m)$. Define $\beta \in \{-1, 0, 1\}^m$ by $\beta_i = (\alpha_i^* - \alpha_i)/2$. Note that $\sum_i \beta_i = 0$. It is easy to see that

$$D(\alpha) = \left(m - \delta \sum_{j=1}^{m} \gamma_j \alpha_j \right) \Delta^2/(8m)$$

hence the difference $D(\alpha) - D(\alpha^*)$ is some positive constant times $\sum_j \beta_j \gamma_j$, and so it suffices to show that

$$\sum_{\gamma \in \Gamma_m} \boldsymbol{P}_{\gamma, n}(\forall i,\ N_i = n_i) \sum_{j=1}^{m} \beta_j \gamma_j \geq 0.$$

To prove this inequality, we shall split the outer sum into several parts, and show that each part is nonnegative. Each part corresponds to a set of distributions that satisfy a convenient symmetry property. First, divide the components of β into

$m/2$ pairs (i, j), with $\beta_i = -\beta_j$. Without loss of generality, suppose

$$\left.\begin{aligned}&\beta_{2i-1} = -\beta_{2i},\\ &\beta_{2i-1} \le 0, \text{ and}\\ &\beta_{2i} \ge 0\end{aligned}\right\}, \qquad \text{for all } 1 \le i \le m/2. \tag{13}$$

Then for $\tilde{\gamma} \in \{-1, 1\}^m$, let $S(\tilde{\gamma})$ denote the set of all permuted versions of $\tilde{\gamma}$ obtained by swapping the components $\tilde{\gamma}_{2i-1}$ and $\tilde{\gamma}_{2i}$, for all i in some subset of $\{1, \cdots, m/2\}$. Clearly, it suffices to show that for all $\tilde{\gamma} \in \Gamma_m$

$$\sum_{\gamma \in S(\tilde{\gamma})} \boldsymbol{P}_{\gamma, n}(\forall i,\ N_i = n_i) \sum_{j=1}^{m} \beta_j \gamma_j \ge 0.$$

But we have

$$\begin{aligned}&\sum_{\gamma \in S(\tilde{\gamma})} \boldsymbol{P}_{\gamma, n}(\forall i,\ N_i = n_i) \sum_{j=1}^{m} \beta_j \gamma_j\\ &\quad = \sum_{\gamma \in S(\tilde{\gamma})} \boldsymbol{P}_{\gamma, n}(\forall i,\ N_i = n_i | \forall i,\ N_{2i-1} + N_{2i} = n_{2i-1} + n_{2i})\\ &\qquad \times \boldsymbol{P}_{\gamma, n}(\forall i,\ N_{2i-1} + N_{2i} = n_{2i-1} + n_{2i}) \sum_{j=1}^{m} \beta_j \gamma_j\\ &\quad = \boldsymbol{P}_{\tilde{\gamma}, n}(\forall i,\ N_{2i-1} + N_{2i} = n_{2i-1} + n_{2i})\\ &\qquad \times \sum_{\gamma \in S(\tilde{\gamma})} \boldsymbol{P}_{\gamma, n}(\forall i,\ N_i = n_i | \forall i,\\ &\qquad\qquad N_{2i-1} + N_{2i} = n_{2i-1} + n_{2i}) \sum_{j=1}^{m} \beta_j \gamma_j.\end{aligned}$$

We can ignore the nonnegative constant factor, and the other probabilities are of independent events, so we can write

$$\begin{aligned}&\sum_{\gamma \in S(\tilde{\gamma})} \boldsymbol{P}_{\gamma, n}(\forall i,\ N_i = n_i | \forall i,\ N_{2i-1} + N_{2i} = n_{2i-1} + n_{2i})\\ &\qquad \cdot \sum_{j=1}^{m} \beta_j \gamma_j\\ &\quad = \sum_{\gamma \in S(\tilde{\gamma})} \prod_{i=1}^{m/2} \boldsymbol{P}_{(\gamma_{2i-1}, \gamma_{2i}), n_{2i-1} + n_{2i}}\\ &\qquad (N_{2i-1} = n_{2i-1}, N_{2i} = n_{2i}) \sum_{j=1}^{m} \beta_j \gamma_j.\end{aligned}$$

So it suffices to show that for all $\tilde{\gamma} \in \{-1, 1\}^m$, all $n_1, \cdots, n_m$ summing to n, and all $\beta \in \{-1, 0, 1\}^m$ satisfying (13), we have

$$\sum_{\gamma \in S(\tilde{\gamma})} \prod_{i=1}^{m/2} \boldsymbol{P}_{(\gamma_{2i-1}, \gamma_{2i}), n_{2i-1} + n_{2i}} (N_{2i-1} = n_{2i-1}, N_{2i} = n_{2i}) \sum_{j=1}^{m} \beta_j \gamma_j \ge 0. \tag{14}$$

Without loss of generality, we can assume that $\beta_j \neq 0$ for all j. Indeed, suppose that $\beta_{2i-1} = \beta_{2i} = 0$ for some i. Then we can split the sum over γ in (14) into a sum over the pair $(\gamma_{2i-1}, \gamma_{2i})$ and a sum over the other components of γ, and the corresponding factors in the product can be taken outside the outermost sum, since $\sum_{j=1}^{m} \beta_j \gamma_j$ is identical for both values of the pair $(\gamma_{2i-1}, \gamma_{2i})$.

Now, $\beta_{2i-1} = -1$ and $\beta_{2i} = 1$ imply that $n_{2i-1} \le n_{2i}$. So to show that (14) holds for the cases of interest, it suffices to show that for all even m, for all $n_1, \cdots, n_m$ satisfying $n_{2i-1} \le n_{2i}$, and all $\tilde{b} \in \{-1, 1\}^m$, we have

$$\sum_{b \in S(\tilde{b})} \sum_{j=1}^{m} (-1)^j b_j \prod_{i=1}^{m/2} P_i \ge 0$$

where

$$P_i = \boldsymbol{P}_{(b_{2i-1}, b_{2i}), n_{2i-1} + n_{2i}} (N_{2i-1} = n_{2i-1}, N_{2i} = n_{2i}).$$

First suppose $m = 2$. If $\tilde{b}_1 = \tilde{b}_2$, the expression is clearly zero. Otherwise, it is equal to

$$\begin{aligned}&2(\boldsymbol{P}_{(-1, 1), n_1 + n_2}(N_1 = n_1, N_2 = n_2)\\ &\qquad - \boldsymbol{P}_{(1, -1), n_1 + n_2}(N_1 = n_1, N_2 = n_2))\\ &\quad = 2(\boldsymbol{P}_{(-1, 1), n_1 + n_2}(N_1 = n_1, N_2 = n_2)\\ &\qquad - \boldsymbol{P}_{(-1, 1), n_1 + n_2}(N_1 = n_2, N_2 = n_1))\end{aligned}$$

which is clearly nonnegative, since $n_2 \ge n_1$. Next, suppose the expression is nonnegative up to some even number m. Let $\tilde{b} \in \{-1, 1\}^{m+2}$. Then

$$\begin{aligned}&\sum_{b \in S(\tilde{b})} \sum_{j=1}^{m+2} (-1)^j b_j \prod_{i=1}^{m/2+1} P_i\\ &\quad = \sum_{b_1, \cdots, b_m} \sum_{b_{m+1}, b_{m+2}} \left(\sum_{j=1}^{m} (-1)^j b_j + \sum_{j=m+1}^{m+2} (-1)^j b_j \right)\\ &\qquad \cdot \left(\prod_{i=1}^{m/2} P_i \right) P_{m/2+1}\\ &\quad = \sum_{b_{m+1}, b_{m+2}} P_{m/2+1} \left(\sum_{b_1, \cdots, b_m} \sum_{j=1}^{m} (-1)^j b_j \prod_{i=1}^{m/2} P_i \right)\\ &\qquad + \sum_{b_1, \cdots, b_m} \prod_{i=1}^{m/2} P_i \left(\sum_{b_{m+1}, b_{m+2}} \sum_{j=m+1}^{m+2} (-1)^j b_j P_{m/2+1} \right)\end{aligned}$$

and both of these terms are nonnegative, since the expressions in parentheses are nonnegative by the inductive hypothesis. □

REFERENCES

[1] K. Alexander, "Probability inequalities for empirical processes and a law of the iterated logarithm," *Ann. Prob.*, vol. 4, pp. 1041–1067, 1984.
[2] S. N. Bernstein, *The Theory of Probabilities.* Moscow, USSR: Gastehizdat, 1946.

[3] 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, 1994).
[4] Y. S. Chow and H. Teicher, *Probability Theory, Independence, Interchangeability, Martingales.* New York: Springer-Verlag, 1978.
[5] L. D. Davisson, "Universal lossless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.
[6] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition.* New York: Springer-Verlag, 1996.
[7] R. M. Gray, J. C. Kieffer, and Y. Linde, "Locally optimum block quantizer design," *Inform. Contr.*, vol. 45, pp. 178–198, 1980.
[8] A. N. Kolmogorov and V. M. Tikhomirov, "ϵ-entropy and ϵ-capacity of sets in function spaces," *Transl. Amer. Math. Soc.*, vol. 17, pp. 277–364, 1961.
[9] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, 1980.
[10] 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*, vol. 40, pp. 1728–1740, Nov. 1994.
[11] ———, "Empirical quantizer design in the presence of source noise or channel noise," *IEEE Trans. Inform. Theory*, vol. 43, pp. 612–623, Mar. 1997.
[12] C. L. Mallows, "An inequality involving multinomial probabilities," *Biometrika*, vol. 55, pp. 422–424, 1968.
[13] N. Merhav and J. Ziv, "On the amount of side information required for lossy data compression," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1112–1121, July 1997.
[14] D. L. Neuhoff, R. M. Gray, and L. D. Davisson, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 511–523, Sept. 1975.
[15] A. Nobel and R. Olshen, personal communication.
[16] D. Pollard, "Strong consistency of k-means clustering," *Ann. Statist.*, vol. 9, pp. 135–140, 1981.
[17] ———, "A central limit theorem for k-means clustering," *Ann. Prob.*, vol. 10, pp. 919–926, 1982.
[18] ———, "Quantization and the method of k-means," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 199–205, 1982.
[19] K. Rose, E. Gurewitz, and G. C. Fox, "Vector quantization by deterministic annealing," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1249–1257, July 1992.
[20] P. C. Shields, "When is the weak rate equal to the strong rate?" in *Proc. 1994 IEEE-IMS Workshop on Information Theory and Statistics*. New York: IEEE, 1994, p. 16.
[21] M. Talagrand, "Sharper bounds for Gaussian and empirical processes," *Ann. Prob.*, vol. 22, pp. 28–76, 1994.
[22] V. N. Vapnik and A. Ya. Chervonenkis, *Theory of Pattern Recognition.* Moscow, USSR: Nauka, 1974, in Russian. (German translation: *Theorie der Zeichenerkennung.* Berlin: Akademie Verlag, 1979.)
[23] E. Yair, K. Zeger, and A. Gersho, "Competitive learning and soft competition for vector quantizer design," *IEEE Trans. Signal Processing*, vol. 40, pp. 294–309, Feb. 1992.