Improved Minimax Bounds on the Test and Training Distortion of Empirically Designed Vector Quantizers *

András Antos

Informatics Laboratory, Research Division, Computer and Automation Research Institute of the Hungarian Academy of Sciences, H-1518 Lágymányosi u.11, Budapest, Hungary. antos@szit.bme.hu

Abstract. It is shown by earlier results that the minimax expected (test) distortion redundancy of empirical vector quantizers with three or more levels designed from n independent and identically distributed data points is at least $\Omega(1/\sqrt{n})$ for the class of distributions on a bounded set. In this paper, a much simpler construction and proof for this are given with much better constants. There are similar bounds for the training distortion of the empirically optimal vector quantizer with three or more levels. These rates, however, do not hold for a one-level quantizer. Here the two-level quantizer case is clarified, showing that it already shares the behavior of the general case. Given that the minimax bounds are proved using a construction that involves discrete distributions, one suspects that for the class of distributions with uniformly bounded continuous densities, the expected distortion redundancy might decrease as $o(1/\sqrt{n})$ uniformly. It is shown as well that this is not so, proving that the lower bound for the expected test distortion remains true for these subclasses.

1 Introduction

Designing empirical vector quantizers is an important problem in data compression. In many practical situations we do not have a good source model in hand, but we are able to collect source samples, called also the training data, to get information on the source distribution. Here our aim is to design a quantizer with a given rate, based on these samples, whose expected distortion on the source distribution is as close to the distortion of an optimal quantizer (that is, one with minimum distortion) of the same rate as possible.

One approach to this problem is, for example, the empirical distortion minimization, supported by the idea that if the samples are from the real source, then a quantizer that performs well on the training data (i.e., that has small training distortion) should have a good performance on this source, as well. In fact, Pollard [1], [2] showed that this method is consistent under general conditions on the source μ when the training sample is n consecutive elements of a stationary and ergodic sequence drawn according to this source. He proved that

* This research was supported in part by the NATO Science Fellowship.

the mean-squared error (MSE) $D(\mu, q_n^*)$ of the empirically optimal quantizer q_n^* (when applied to the true source) tends almost surely (a.s.) to the optimum MSE $D^*(\mu)$ achieved by an optimal quantizer. It is also shown there that the training distortion $D(\mu_n, q_n^*)$ of q_n^* tends as well to $D^*(\mu)$ a.s., that is, $D(\mu_n, q_n^*)$ is a strongly consistent estimate of $D^*(\mu)$, however $\mathbf{E}D(\mu_n, q_n^*) \leq D^*(\mu)$.

These consistency results do not inform us about the number of training samples required to ensure that the test or training distortion of q_n^* is close to the optimum MSE. To answer this question one usually studies the rates of the convergences above, that is, gives upper bounds for the distortion redundancy $D(\mu, q_n^*) - D^*(\mu)$ and for $D^*(\mu) - D(\mu_n, q_n^*)$ for finite sample sizes. It was shown by Linder *et al.* [3] (see also [4],[5]) that the expected distortion redundancy $J(\mu, q_n^*) = \mathbf{E}D(\mu, q_n^*) - D^*(\mu)$ as well as $D^*(\mu) - \mathbf{E}D(\mu_n, q_n^*)$ can be bounded as $O(1/\sqrt{n})$ uniformly for all distributions over a given bounded region. These results have several extensions; see, for example, [6], [7], and [8].

In case of one-level quantizing, the only code point of q_n^* is the average of the samples, and it is easy to show that $J(\mu, q_n^*) = D^*(\mu) - \mathbf{E}D(\mu_n, q_n^*) = O(1/n)$ uniformly for all distributions over a given bounded region, that is, the convergences rate of both the test and the training distortion are much faster than the general $O(1/\sqrt{n})$ rate.

However, it is known that for three or more quantization levels, the $O(1/\sqrt{n})$ upper bounds on $J(\mu, q_n^*)$ and $D^*(\mu) - \mathbf{E}D(\mu_n, q_n^*)$ are tight in the minimax sense: It is shown in [9] for these cases that for any empirical quantizer design algorithm, that is, when the resulting quantizer q_n is an arbitrary function of the samples, and for any n large enough, there is a distribution in the class of distributions over a bounded region such that $J(\mu, q_n) = \mathbf{E}D(\mu, q_n) - D^*(\mu) \geq c_0/\sqrt{n}$. These "bad" distributions used in the proof are supported on finitely many atoms, but the proof is really complicated, which also leads to losses in the constants. As [9] words: "The constant c_0 of the theorem is rather small [...], and it can probably be improved upon at the expense of a more complicated analysis." However, in Theorem 1 of this paper, we give an even simpler and much shorter proof for this minimax lower bound with about 10^7-times better constant c_0.

For the training distortion of q_n^*, Linder [4] showed for three or more quantization levels that there is a distribution in the class of distributions over a bounded region such that for all sample size large enough $D^*(\mu) - \mathbf{E}D(\mu_n, q_n^*) = \Omega(1/\sqrt{n})$.

However, the above lower bounds in [9] and [4] leave open the case of two-level quantizers. We prove in Theorems 2 and 3 for the test and training distortion, respectively, that in this case we already get the same lower rates as in the general (at least three-level) case.

Based on certain considerations, one might think that for the class of distributions over a bounded region with uniformly bounded continuous densities, $J(\mu, q_n^*) = O(1/n^\alpha)$ uniformly, where $\alpha > 1/2$ constant. We also show that this is not so proving that the lower bound for the expected distortion redundancy remains true restricting ourselves to these type of distribution classes.

The rest of the paper is organized as follows. In Sect. 2 we give the necessary definitions and earlier results in details. In Sect. 3 we present our lower bound results. Section 4 contains the proofs of some theorems.

2 Empirical Vector Quantizer Design

Nearest Neighbor Quantizers. A d-dimensional N*-level vector quantizer* ($N \geq 1$ is an integer) is a measurable mapping $q : \mathcal{R}^d \to \mathcal{C}$, where the *codebook* $\mathcal{C} = \{y_1, \ldots, y_N\} \subset \mathcal{R}^d$ is a collection of N distinct d-vectors, called the *code points*. The quantizer is completely characterized by its codebook and the sets $S_i = \{x \in \mathcal{R}^d : q(x) = y_i\}$, $i = 1, \ldots, N$ called the *partition cells* (as they form a partition of $\mathcal{R}^d$) via the rule $q(x) = y_i$, if $x \in S_i$. The set $\{S_1, \ldots, S_N\}$ is called the partition of q.

The performance of the quantizer q, when quantizing a random vector $X \in \mathcal{R}^d$ with distribution μ called the *source*, is measured by the *expected distortion* (or mean-squared error)

$$D(\mu, q) \stackrel{\text{def}}{=} \mathbf{E}\{\|X - q(X)\|^2\}$$

where $\|\cdot\|$ denotes the Euclidean norm. Throughout this paper we assume $\mathbf{E}\{\|X\|^2\} < \infty$, thus the distortion is finite. The goal of quantization is to find quantizers with a given number of code points that have minimum distortion. We define the optimum distortion by

$$D^*(\mu) \stackrel{\text{def}}{=} \inf_{q \in \mathcal{Q}_N} D(\mu, q)$$

where $\mathcal{Q}_N$ denotes the set of all N-level quantizers. If $D(\mu, q^*) = D^*(\mu)$ then $q^* \in \mathcal{Q}_N$ is called *optimal*.

A quantizer q is called a *nearest neighbor* quantizer, if it satisfies

$$\|x - q(x)\| = \min_{y_i \in \mathcal{C}} \|x - y_i\|, \qquad \text{for all } x \in \mathcal{R}^d \ .$$

It is well known (see, e.g., [2]) that it suffices to consider nearest neighbor quantizers when searching for an optimal quantizer. Thus any optimal quantizer can be assumed to be nearest neighbor, and so finding an optimal quantizer is equivalent to finding its codebook. Using this, it is proved in [2] that there always exists an optimal quantizer.

Empirical Design. In many situations, the distribution μ is unknown, and the only available information about it is in the form of *training data*, that is, a sequence $X_1^n \stackrel{\text{def}}{=} X_1, \ldots, X_n$ 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 *empirically designed* quantizer $q_n(\cdot) = q_n(\cdot, X_1, \ldots, X_n)$, which is a random function depending on the training sample. The goal is to produce such quantizers with performance approaching $D^*(\mu)$. The performance of q_n in quantizing X is measured by the *test distortion*

$$D(\mu, q_n) = \mathbf{E}\{\|X - q_n(X)\|^2 | X_1^n\} = \int_{\mathcal{R}^d} \|x - q_n(x)\|^2 \, \mu(dx) \ .$$

Note that $D(\mu, q_n)$ is a random variable.

Also of interest is the *training distortion* (or empirical distortion) of any q, defined as its distortion in quantizing the training data

$$D(\mu_n, q) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - q(X_i)\|^2$$

where μ_n is the *empirical distribution* corresponding to the training data defined by

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

for every Borel set $A \subset \mathcal{R}^d$, where I_E denotes the indicator function of the event E. Note that although q is a deterministic mapping, the training distortion $D(\mu_n, q)$ is also a random variable depending on the training data X_1^n.

We define an *empirically optimal vector quantizer* (which is a specific empirically designed quantizer) as a quantizer q_n^* that minimizes the training distortion:

$$q_n^* \stackrel{\text{def}}{=} \arg\min_{q \in \mathcal{Q}_N} D(\mu_n, q) \ .$$

q_n^* is an optimal quantizer for μ_n. Note that q_n^* always exists (although it is not necessarily unique), and we can assume that it is a nearest neighbor quantizer.

It is well known that $\{q_n^*; n = 1, 2, \ldots\}$ is consistent in the sense that $\lim_{n\to\infty} D(\mu, q_n^*) = D^*(\mu)$ a.s. for any $N \geq 1$ [1],[2]. Similarly, it is also shown that $D(\mu_n, q_n^*)$ is a strongly consistent estimate of $D^*(\mu)$, that is, $\lim_{n\to\infty} D(\mu_n, q_n^*) = D^*(\mu)$ a.s. However the training distortion is always optimistically biased, that is, $\mathbf{E}D(\mu_n, q_n^*) < D^*(\mu)$ (unless $\mathbf{E}D(\mu_n, q_n^*) = D^*(\mu) = 0$).

Finite Sample Upper Bounds. To determine the number of samples needed to achieve a given level of distortion, the finite sample behavior of the *expected distortion redundancy*

$$J(\mu, q_n) \stackrel{\text{def}}{=} \mathbf{E}D(\mu, q_n) - D^*(\mu)$$

has to be analyzed.

Denote the closed ball of radius r centered at $x \in \mathcal{R}^d$ by $S(x, r)$ and, in particular, let $S = S(0, 1)$. Let $\mathcal{P}(1)$ denote the family of distributions supported on S. From this point, we assume that $\mu \in \mathcal{P}(1)$, hence μ is supported on

S, that is, $\mathbf{P}\{\|X\| \leq 1\} = 1$. (One can generalize the results to cases when $\mathbf{P}\{\|X\| \leq T\} = 1$ for some fixed $T < \infty$ by straightforward scaling, see, e.g., [5].)

It is of interest how fast $J(\mu, q_n^*)$ converges to 0. To our knowledge the best accessible upper bound concerning its dependence on n, N, and d, can be obtained by combining results of [3] with recent developments of [4], implying:

$$0 \leq J(\mu, q_n^*) \leq 2C_1\sqrt{\frac{Nd}{n}}$$

for all $n \geq 1$ and $\mu \in \mathcal{P}(1)$, where $C_1 = 96$. The result in [4] also implies an upper bound on the size of the bias of $D(\mu_n, q_n^*)$:

$$0 \leq D^*(\mu) - \mathbf{E}D(\mu_n, q_n^*) \leq C_1\sqrt{\frac{Nd}{n}}$$

for all $n \geq 1$ and $\mu \in \mathcal{P}(1)$. (Sometimes this bound is referred to as a *lower* bound on $\mathbf{E}D(\mu_n, q_n^*)$.)

Minimax Lower Bounds. A natural question is whether there exists a method, perhaps different from empirical distortion minimization, which provides an empirically designed quantizer with substantially smaller test distortion. In case of $N = 1$, it is easy to check that $J(\mu, q_n^*) = \mathbf{Var}X/n$. Similarly, for the training distortion $D^*(\mu) - \mathbf{E}D(\mu_n, q_n^*) = \mathbf{Var}X/n$. Thus the convergence rate is $O(1/n)$ in both cases, substantially faster than the $O(1/\sqrt{n})$ rate above. However, for $N \geq 3$ lower bounds in [9] and [4] show that the $O(1/\sqrt{n})$ convergence rate above cannot be improved in the following sense (Φ denotes the standard normal distribution function):

Proposition 1 (Bartlett, Linder, and Lugosi [9]). *If $N \geq 3$, then for any empirically designed N-level quantizer q_n trained on $n \geq n_0 = 16N/(3\Phi(-2)^2) \approx 10305N$ samples, we have*

$$\sup_{\mu \in \mathcal{P}(1)} J(\mu, q_n) \geq CN^{-2/d}\sqrt{\frac{N}{n}}$$

where $C = \Phi(-2)^4 2^{-12}/\sqrt{6} \approx 2.67 \cdot 10^{-11}$.

The construction and the proof in [9] is extremely complicated, which also leads to losses in the constants. In what follows, we give a much simpler construction and a shorter proof for Proposition 1 with about 10^7-times better constant C and better n_0.

A somewhat similar lower bound on the bias of the training distortion of empirically optimal quantizers is the following:

Proposition 2 (Linder [4]). *If $N \geq 3$, then there exists a distribution $\mu \in \mathcal{P}(1)$ such that for all training set size $n \geq n_0 = 2N/3$*

$$D^*(\mu) - \mathbf{E}D(\mu_n, q_n^*) \geq cN^{-2/d}\sqrt{\frac{N}{n}}$$

where $c = 2^{-8}/3$.

This is often referred to as an *upper* bound on $\mathbf{E}D(\mu_n, q_n^*)$. Note that this Proposition holds only for one empirically designed quantizer, the empirically optimal one, but on the other hand, it gives also an *individual* rate in the sense, that it provides a fix distribution for all n.

3 Main Results

Improved Minimax Lower Bounds. First we state the improved versions of Proposition 1:

Theorem 1. *Proposition 1 holds also with* $n_0 = 8N$ *and* $C = \frac{211}{360^2\sqrt{3}}\Phi(-0.92) \approx 1.68 \cdot 10^{-4}$.

The idea behind our proof and its differences from the one in [9] are best illustrated by the special case $d = 1$, $N = 6$. Assume that μ is concentrated on four pairs of points: $(-1, -1+\Delta)$, $(-1/3, -1/3+\Delta)$, $(1/3-\Delta, 1/3)$, and $(1-\Delta, 1)$, such that $\mu(-1) = \mu(-1+\Delta) = (1+u_1\delta)/8$, $\mu(1) = \mu(1-\Delta) = (1-u_1\delta)/8$, $\mu(-1/3) = \mu(-1/3+\Delta) = (1+u_2\delta)/8$, and $\mu(1/3) = \mu(1/3-\Delta) = (1-u_2\delta)/8$, where $u_1, u_2 \in \{-1, +1\}$. This gives four distributions. Then if Δ is sufficiently small, the code points of the 6-level optimal quantizer consist of $\{-1+\Delta/2, 1-\Delta, 1\}$ if $u_1 = -1$ or $\{-1, -1+\Delta, 1-\Delta/2\}$ if $u_1 = 1$, plus $\{-1/3+\Delta/2, 1/3-\Delta, 1/3\}$ if $u_2 = -1$ or $\{-1/3, -1/3+\Delta, 1/3-\Delta/2\}$ if $u_2 = -1$. Therefore, an empirical quantizer should "learn" from the data which of the four distributions generates that. This leads to a hypothesis testing problem for u_1 and u_2 whose error may be approximated through inequalities for the binomial distribution. Proper choice of Δ and δ provides the desired $\Omega(1/\sqrt{n})$ lower bound for the expected distortion redundancy. The idea is the same for general case.

For the four distributions above, the weights are $(1-\delta)/8$ and $(1+\delta)/8$, and exactly two of the pairs have mass $(1-\delta)/4$. In the original proof in [9], for the $N = 6$ case, all the $\binom{4}{2} = 6$ distributions with this property were involved. In the general case, this meant an exponentially (in N) larger distribution class and more involved parameterization than in our proof, led to expressions which were really difficult to handle, and resulted in significant loss in the constants. Our simplified proof is somewhat suggested also by ideas in [4, Proof of Theorem 1]. Although our result gives an improvement only in the constant factor of the lower bound and thus it leaves the gap compared to the upper bound, we conjecture that it is sharp in respect of the factor $N^{-2/d}$ and that the upper bound might be improved in this respect.

Note that, as a special case of Theorem 1 for $N = 3$, we can get the following constants by careful analysis (see also [5]):

Corollary 1. *For any empirically designed* 3*-level quantizer* q_n *trained on* $n \geq n_0 = 9$ *samples,*

$$\sup_{\mu \in \mathcal{P}(1)} J(\mu, q_n) \geq \frac{C}{\sqrt{n}}$$

where $C = 3\Phi(-7/\sqrt{54})/80 \approx 6.39 \cdot 10^{-3}$.

Two-Level Quantizers. The results so far leave open the $N = 2$ case. In the following, we prove that in this case we already get the same rates as in the general case of $N \geq 3$. Thus, as soon as the criterion function (the expected distortion of nearest neighbor quantizers as a function of the code vectors) is not convex, the rate is slower than for $N = 1$ when the criterion function is convex. This requires some modification of the constructions in the proofs of Propositions 1 and 2. For the test distortion, we have:

Theorem 2. *Corollary 1 holds also for* $N = 2$ *with* $n_0 = 25$ *and* $C = \frac{1557}{30625}\Phi(-\frac{7}{\sqrt{54}}) \approx 8.66 \cdot 10^{-3}$.

The idea behind the proof and its differences from the $N = 3$ case is the following: In the proof for $N = 3$, μ was concentrated on four (two pairs) of points. Now assume that μ is concentrated on three points: -1, 0, and 1, such that $\mu(0) = 0.4$, $\mu(-1) = 0.3 + u\delta/2$ and $\mu(1) = 0.3 - u\delta/2$, where u is either -1 or $+1$. Then the code points of the 2-level optimal quantizer consist of $\{-\frac{3-5\delta}{7-5\delta}, 1\}$ if $u = -1$, and $\{-1, \frac{3-5\delta}{7-5\delta}\}$ if $u = 1$. Though these depend on δ, they remain in a small interval distinguishing sharply the two cases. Thus, though the family of sensible quantizers are richer now, we can proceed similarly to the $N \geq 3$ case. For a detailed proof, see the forthcoming full version.

For the training distortion, we have:

Theorem 3. *Proposition 2 holds also for* $N = 2$ *with* $n_0 = 1$ *and* $c = 1/24$.

Continuous Distributions. Certain considerations suggest that for classes of more regularized distributions a faster uniform rate of convergence can be achieved for $J(\mu, q_n)$ for some empirical quantizer q_n. For example, [10, Theorem 2, Corollary 1] and their proofs imply $O(\log n/n)$ uniform rate of $J(\mu, q_n^*)$ for a class of scalar distributions with continuous densities supported on S such that

(*) there is a fixed $\epsilon > 0$ that for all these densities the second derivative matrices of the criterion function have only eigenvalues greater than ϵ at each optimal codebook.

(Note that, under similar smoothness assumptions, Chou [11] pointed out that $D(\mu, q_n^*) - D^*(\mu)$ decreases as $O(1/n)$ in probability based on an asymptotic normality result of Pollard [12].) On the other hand, approximating the discrete distributions used in the proof of Proposition 1 by higher and higher density function peaks would lead to an unbounded sequence of densities. Considering this, one might think that the ugly technical condition (*) can be omitted in proving faster uniform convergence rates, that is, for example, for the class $\mathcal{D}_K$ of distributions with continuous densities supported on S and bounded by $K > 0$, there is a sequence $\{a_n\} = o(1/\sqrt{n})$ depending only on N, d, and K such that for all $n \geq 1$ and $\mu \in \mathcal{D}_K$,

$$J(\mu, q_n^*) \leq a_n \quad .$$

Here we show that, surprisingly, this conjecture is false, moreover it fails for any empirical quantizer q_n, proving that Proposition 1 remains true restricting ourselves for the case $\mu \in \mathcal{D}_K$. For simplicity, this is given only for the special case $d = 1$ and $N = 3$. The basic idea is the same for the general case.

Theorem 4. *If $K \geq 12/5$ then Corollary 1 holds with $C = \frac{3}{80}(1-\frac{6}{5K})^2 \Phi(-\frac{7}{\sqrt{54}})$ $\approx 6.39 \cdot 10^{-3}\left(1-\frac{6}{5K}\right)^2$ also if the supremum is taken over only $\mathcal{D}_K$.*

For the limit $K \to \infty$, we get

Corollary 2. *Corollary 1 holds also if the supremum is taken over only for $\mu \in \mathcal{P}(1)$ with continuous densities.*

Theorem 4 might be surprising, also because replacing the atoms by intervals of a density, the cell interval boundaries of the quantizers can split them anywhere, which seems to be an essential difference compared to the discrete case. The idea behind the proof is the following: Introduce $h = 3/(10K)$. Note that $h \leq 1/8$ for $K \geq 12/5$. In contrast with the discrete case, where μ was concentrated on the set $\{-1, -1/2, 1/2, 1\}$, here assume instead that μ is concentrated on the set

$$A = [-1, -1+2h] \cup [-1/2-2h, -1/2] \cup [1/2, 1/2+2h] \cup [1-2h, 1] \qquad (1)$$

such that $\mu([-1,-1+2h]) = \mu([-1/2-2h,-1/2]) = (1+u\delta)/4$ and $\mu([1/2, 1/2+2h]) = \mu([1-2h, 1]) = (1-u\delta)/4$, where u is either -1 or $+1$, and the distribution inside all of these intervals have a symmetric continuous density vanishing at the end points. Then the code points of the 3-level optimal quantizer consist of $\{-3/4, 1/2+h, 1-h\}$ if $u = -1$, and $\{-1+h, -1/2-h, 3/4\}$ if $u = 1$. Using small (but n-independent) h, we can assure that none of the cell interval boundaries of the sensible quantizers actually splits any interval of the support A, thus we can proceed similarly to the discrete case. See the forthcoming full version for a detailed proof which uses the following lemma:

Lemma 1. *If $h \leq 1/8$ and $\delta \leq 1/4$ then for any 3-level scalar quantizer q there exists a 3-level nearest neighbor scalar quantizer q' whose two cell interval boundaries do not intersect the support A in (1) and, in particular, one of these boundaries is in $(-1/2, 1/2)$, such that for $u = \pm 1$, $D(\mu_u, q') \leq D(\mu_u, q)$.*

4 Proofs

Proof of Theorem 1. For the sake of simplicity, we assume that N is divisible by 3. (This assumption is clearly insignificant, see [4, proof of Theorem 1] for an argument.) Certainly,

$$\sup_{\mu \in \mathcal{P}(1)} J(\mu, q_n) \geq \sup_{\mu \in \mathcal{D}} J(\mu, q_n) \qquad (2)$$

where $\mathcal{D} \subseteq \mathcal{P}(1)$ is any restricted class. Define $\mathcal{D}$ as follows: each member of $\mathcal{D}$ is concentrated on the set of $4k = 4N/3$ fixed points $\{z_i, z_i' \in S : i = 1, \ldots, 2k\}$, where $z_i' = z_i + (\Delta, 0, 0, \ldots, 0)$ and $0 < \Delta \le 1/4$. The positions of $z_1, \ldots, z_{2k}$ satisfy the property that the distance between any two of them is greater than 3Δ. The members of $\mathcal{D}$ are parameterized by the sign vectors $u = (u_1, \ldots, u_k) \in \{-1, +1\}^k$. For $1 \le i \le k$, denote z_{k+i} and z'_{k+i} also by y_i and y_i', respectively, and define $u_{k+i} = -u_i$. For $0 < \delta \le 1/2$ and $i = 1, \ldots, k$, set

$$\mu_u(\{z_i\}) \stackrel{\text{def}}{=} \mu_u(\{z_i'\}) \stackrel{\text{def}}{=} \frac{1 + u_i\delta}{4k}$$
$$\mu_u(\{y_i\}) \stackrel{\text{def}}{=} \mu_u(\{y_i'\}) \stackrel{\text{def}}{=} \frac{1 - u_i\delta}{4k} \ .$$

(The total mass adds up to one.) Let $\mathcal{D}$ contain all such distributions. Thus $|\mathcal{D}| = 2^k$ and $N = 3k$.

Let $\mathcal{Q}^*$ denote the collection of N-level nearest neighbor quantizers q such that for each $1 \le i \le k$, q has code points either at z_i, z_i', and $(y_i + y_i')/2$, or at $(z_i + z_i')/2$, y_i, and y_i'. Then for each μ_u the essentially unique optimal N-level quantizer is in $\mathcal{Q}^*$. It is easy to see that for all u,

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

Let $\mathcal{Q}$ denote the collection of N-level nearest neighbor quantizers q such that for some k values of $i \in \{1, \ldots, 2k\}$, q has code points at both z_i and z_i', and for the remaining k values of i, q has a single code point at $(z_i + z_i')/2$. Note that $\mathcal{Q} \supseteq \mathcal{Q}^*$. One can show that for any N-level quantizer q there is a $q' \in \mathcal{Q}$ such that

$$D(\mu_u, q') \le D(\mu_u, q), \qquad \text{for all } u \ . \tag{3}$$

See [9, Appendix, Proof of Step 3] for a proof.

Let $\mathcal{Q}^{(n)}$ denote the family of empirically designed quantizers q_n such that for every fixed $x_1, \ldots, x_n$, we have $q_n(\cdot, x_1, \ldots, x_n) \in \mathcal{Q}$. Then (3) implies that it is enough to take the infimum over $\mathcal{Q}^{(n)}$, that is,

$$\inf_{q_n} \sup_{\mu \in \mathcal{P}(1)} J(\mu, q_n) \ge \inf_{q_n} \max_{u \in \{-1,1\}^k} J(\mu_u, q_n) = \inf_{q_n \in \mathcal{Q}^{(n)}} \max_{u \in \{-1,1\}^k} J(\mu_u, q_n) \ . \tag{4}$$

Any $q_n \in \mathcal{Q}^{(n)}$ can be represented by the 3-tuple $(A_n, A_n', \{u_{n,i}\}_{i \le k})$, where A_n,A_n', and $\{u_{n,i}\}$ are determined by q_n (and thus depend on the data X_1^n) the following way:

- $A_n = \{i \le k : (z_i + z_i')/2$ (exclusive) or $(y_i + y_i')/2$ is code point of $q_n\}$,
- $A_n' = \{i \le k :$ all of z_i, z_i', y_i, y_i' are code points of $q_n\}$,
- $u_{n,i} = \begin{cases} +1 \text{ if } z_i \text{ and } z_i' \text{ are code points of } q_n \\ -1 \text{ otherwise} \end{cases} \qquad (i = 1, \ldots, k).$

Note that $2|A_n'| + |A_n| = k$. For any u and $q_n \in \mathcal{Q}^{(n)}$,

$$D(\mu_u, q_n) - D^*(\mu_u) = \frac{\Delta^2\delta}{4k}\left(|A_n'| + \sum_{i \in A_n} I_{\{u_{n,i} \ne u_i\}}\right)$$

$$
\begin{aligned}
&= \frac{\Delta^2\delta}{4k}\left(\sum_{i\le k, i\notin A_n}\frac{1}{2} + \sum_{i\in A_n} I_{\{u_{n,i}\neq u_i\}}\right) \\
&= \frac{\Delta^2\delta}{4k}\sum_{i=1}^{k}\left(I_{\{i\notin A_n\}}\frac{1}{2} + I_{\{i\in A_n\}}I_{\{u_{n,i}\neq u_i\}}\right) \\
&\ge \frac{\Delta^2\delta}{4k}\sum_{i=1}^{k}\frac{1}{2}I_{\{u_{n,i}\neq u_i\}} = \frac{\Delta^2\delta}{8k}\sum_{i=1}^{k}I_{\{u_{n,i}\neq u_i\}}
\end{aligned}
$$

and thus

$$
\max_{u\in\{-1,1\}^k} J(\mu_u, q_n) \ge \frac{\Delta^2\delta}{8k}\max_{u\in\{-1,1\}^k}\sum_{i=1}^{k}\mathbf{P}\{u_{n,i}\neq u_i\} \ . \tag{5}
$$

We write $X_1^n(u)$ for X_1^n to indicate how the samples depend on u, and use randomization such that u is replaced by a vector $U = (U_1, \ldots, U_k) \in \{-1, +1\}^k$ of i.i.d. zero mean random signs. Then for any $q_n \in \mathcal{Q}^{(n)}$,

$$
\begin{aligned}
\max_{u\in\{-1,+1\}^k}\sum_{i=1}^{k}\mathbf{P}\{u_{n,i}(X_1^n(u))\neq u_i\} &\ge \mathbf{E}\left\{\sum_{i=1}^{k}\mathbf{P}\{u_{n,i}(X_1^n(U))\neq U_i|U\}\right\} \\
&\ge \sum_{i=1}^{k}\mathbf{P}\{u_{n,i}(X_1^n(U))\neq U_i\} \ . \qquad (6)
\end{aligned}
$$

Here $\mathbf{P}\{u_{n,i}(X_1^n(U))\neq U_i\}$ is the error probability of the classification problem of whether U_i is -1 or $+1$ based on the observation $X_1^n(U)$, and thus it is not less than the corresponding optimal error probability, which is given by the following formula (see, e.g., [13]):

$$
\begin{aligned}
\inf_{u_{n,i}}\mathbf{P}\{u_{n,i}(X_1^n(U))\neq U_i\} &\ge \mathbf{E}\{\min(\mathbf{P}\{U_i=1|X_1^n\}, \mathbf{P}\{U_i=-1|X_1^n\})\} \\
&= \mathbf{E}\{\min(\mathbf{P}\{U_1=1|X_1^n\}, \mathbf{P}\{U_1=-1|X_1^n\})\} \quad (7)
\end{aligned}
$$

by symmetry. Let $M = |\{j \le n : X_j \in \{z_1, z_1'\}\}|$ and $\bar{M} = |\{j \le n : X_j \in \{y_1, y_1'\}\}|$. It is easy to check that $(M, \bar{M})$ is a sufficient statistics of X_1^n for U_1, and that if $M \le$ (or $\ge$) $\bar{M}$ then $\mathbf{P}\{U_1 = 1|M, \bar{M}\} \le$ (or $\ge$) $\mathbf{P}\{U_1 = -1|M, \bar{M}\}$, respectively. Thus

$$
\begin{aligned}
&\mathbf{E}\{\min(\mathbf{P}\{U_1=1|X_1^n\}, \mathbf{P}\{U_1=-1|X_1^n\})\} \\
&= \mathbf{E}\{\min(\mathbf{P}\{U_1=1|M,\bar{M}\}, \mathbf{P}\{U_1=-1|M,\bar{M}\})\} \\
&\ge \mathbf{E}\{I_{\{M\le\bar{M}\}}\min(\mathbf{P}\{U_1=1|M,\bar{M}\}, \mathbf{P}\{U_1=-1|M,\bar{M}\})\} \\
&= \mathbf{E}\{I_{\{M\le\bar{M}\}}\mathbf{P}\{U_1=1|M,\bar{M}\}\} \qquad (8) \\
&= \mathbf{P}\{M\le\bar{M}, U_1=1\} \\
&= \frac{1}{2}\mathbf{P}\{M\le\bar{M}|U_1=1\} \ .
\end{aligned}
$$

From (4)–(8) we conclude that

$$\inf_{q_n} \sup_{\mu \in \mathcal{P}(1)} J(\mu, q_n) \geq \frac{\Delta^2 \delta}{16} \mathbf{P}\{M \leq \bar{M} | U_1 = 1\} \ . \tag{9}$$

Let $B = M + \bar{M}$, which has binomial distribution with parameters $(n, 1/k)$ for any u. Now

$$\begin{aligned} \mathbf{P}\{M \leq \bar{M} | U_1 = 1\} &= \mathbf{P}\{\bar{M} \geq B/2 | U_1 = 1\} \\ &= \mathbf{E}\{\mathbf{P}\{\bar{M} \geq B/2 | U_1 = 1, B\} | U_1 = 1\} \ . \end{aligned} \tag{10}$$

Note that given $U_1 = 1$ and B, $\bar{M}$ has binomial distribution with parameters B and $p = (1-\delta)/2$. We lower bound the above probability using an inequality by Slud [14] which states that for all $Bp \leq l \leq B(1-p)$,

$$\mathbf{P}\{\bar{M} \geq l | U_1 = 1, B\} \geq \Phi\left(-\frac{l - Bp}{\sqrt{Bp(1-p)}} \right) \ . \tag{11}$$

For $1/\delta \leq B \leq 9/(25\delta^2)$, the choice $l = \lceil B/2 \rceil$ satisfies the conditions of Slud's inequality, and for $\delta < 0.1$ we obtain

$$\begin{aligned} \mathbf{P}\{\bar{M} \geq B/2 | U_1 = 1, B\} &\geq \mathbf{P}\{\bar{M} \geq \lceil B/2 \rceil | U_1 = 1, B\} \\ &\geq \Phi\left(-\frac{\lceil B/2 \rceil - B(1-\delta)/2}{\sqrt{B(1-\delta^2)/4}} \right) \\ &\geq \Phi\left(-\frac{1 + B\delta}{\sqrt{B(1-\delta^2)}} \right) \\ &\geq \Phi\left(-\frac{1 + 3\sqrt{B}/5}{\sqrt{B(1 - 9/(25B))}} \right) = \Phi\left(-\frac{5 + 3\sqrt{B}}{\sqrt{25B - 9}} \right) \\ &\geq \Phi\left(-\frac{5 + 3\sqrt{11}}{\sqrt{25 \cdot 11 - 9}} \right) > \Phi(-0.92) \end{aligned} \tag{12}$$

where we used the conditions $\delta \leq 3/(5\sqrt{B})$ and $B \geq 1/\delta > 10$ above. Hence, using that B and U_1 are independent,

$$\begin{aligned} \mathbf{P}\{M \leq \bar{M} | U_1 = 1\} &\geq \mathbf{E}\left\{ I_{\{\frac{1}{\delta} \leq B \leq \frac{9}{25\delta^2}\}} \Phi(-0.92) \middle| U_1 = 1 \right\} \\ &= \Phi(-0.92) \mathbf{P}\left\{ \frac{1}{\delta} \leq B \leq \frac{9}{25\delta^2} \right\} \ . \end{aligned} \tag{13}$$

For $n \geq 24k = 8N$, letting

$$\delta = \frac{12}{25}\sqrt{\frac{k}{n}} \ \ (< 0.1) \tag{14}$$

by Chebyshev's inequality

$$\begin{aligned}
\mathbf{P}\left\{\frac{1}{\delta} \le B \le \frac{9}{25\delta^2}\right\} &= \mathbf{P}\left\{\frac{25}{12}\sqrt{\frac{n}{k}} \le B \le \frac{25}{16}\frac{n}{k}\right\} \\
&\ge \mathbf{P}\left\{\frac{7}{16}\frac{n}{k} \le B \le \frac{25}{16}\frac{n}{k}\right\} = 1 - \mathbf{P}\left\{|B - \mathbf{E}B| > \frac{9}{16}\frac{n}{k}\right\} \\
&\ge 1 - \frac{n/k(1-1/k)}{\frac{9^2}{16^2}\frac{n^2}{k^2}} \ge 1 - \frac{16^2 k}{9^2 n} \qquad (15) \\
&\ge 1 - \left(\frac{2}{3}\right)^5 = \frac{211}{243} \ .
\end{aligned}$$

Combining (9), (13)–(15), we obtain

$$\inf_{q_n} \sup_{\mu\in\mathcal{P}(1)} J(\mu, q_n) \ge \frac{211\Delta^2}{8100}\Phi(-0.92)\sqrt{\frac{k}{n}}$$

where $0 < \Delta \le 1/4$ such that there exist $z_1, \ldots, z_{2k} \in S$ with distance at least 3Δ between any two of them and with each $z_i' \in S$. The latter is ensured if each $z_i \in S(0, 3/4) \subseteq S(0, 1-\Delta)$. Thus, we need a lower bound for the cardinality of the maximal 3Δ-packing of $S(0, 3/4)$. It is well known (e.g., [15]) that this is lower bounded by the cardinality of the minimal 3Δ-covering of $S(0, 3/4)$, which is clearly bounded below by the ratio of the volume of $S(0, 3/4)$ and that of $S(0, 3\Delta)$, that is, by $(1/4\Delta)^d$. Hence $2k$ points can be packed in $S(0, 3/4)$ as long as $2k \le (1/4\Delta)^d$, thus the choice

$$\Delta = \frac{1}{4(2k)^{1/d}}$$

satisfies the required property. Resubstitution of this value and $2k = 2N/3$ proves the theorem. □

Proof of Theorem 3. Define $\mu \in \mathcal{P}(1)$ as the uniform distribution concentrated on three fixed points $\{-1, 0, 1\}$. It is easy to see that there are exactly two optimal 2-level quantizers for μ; with codebooks

$$\mathcal{C}_1 = \{-1/2, 1\} \qquad \text{and} \qquad \mathcal{C}_2 = \{-1, 1/2\}$$

respectively. Consequently, $D^*(\mu) = 1/6$.

Let M and $\bar{M}$ be the number of training data points falling on -1 and 1, respectively. Let q_n be the training set dependent 2-point nearest neighbor quantizer whose codebook is $\mathcal{C}_1$ if $M < \bar{M}$ and $\mathcal{C}_2$ if $M \ge \bar{M}$. Then the expected training distortion of q_n is

$$\mathbf{E}D(\mu_n, q_n) = \mathbf{E}\left\{\frac{1}{n}\sum_{j=1}^{n}(X_j - q_n(X_j))^2\right\}$$

$$
\begin{aligned}
&= \frac{1}{4n}\mathbf{E}\{n - \max(M, \bar{M})\} \qquad (16)\\
&= \frac{1}{4} - \frac{1}{4n}\mathbf{E}\{\max(M, \bar{M})\} \ .
\end{aligned}
$$

Since the empirically optimal quantizer q_n^* minimizes the training distortion over all 2-point quantizers, we have

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

Therefore, using (16) and that $\mathbf{E}M = \mathbf{E}\bar{M} = n/3$, we can lower bound the difference $D^*(\mu) - \mathbf{E}D(\mu_n, q_n^*)$ as

$$
\begin{aligned}
D^*(\mu) - \mathbf{E}D(\mu_n, q_n^*) &\geq \frac{1}{6} - \frac{1}{4} + \frac{1}{4n}\mathbf{E}\{\max(M, \bar{M})\}\\
&= -\frac{1}{12} + \frac{1}{4n}\mathbf{E}\left\{\frac{M + \bar{M}}{2} + \frac{|\bar{M} - M|}{2}\right\} \qquad (17)\\
&= -\frac{1}{12} + \frac{1}{12} + \frac{\mathbf{E}\{|\bar{M} - M|\}}{8n} = \frac{\mathbf{E}\{|\sum_{j=1}^n X_j|\}}{8n} \ .
\end{aligned}
$$

We lower bound the the last expectation using the following inequality: for any random variable Z with finite fourth moment,

$$
\mathbf{E}|Z| \geq \frac{(\mathbf{E}\{Z^2\})^{3/2}}{(\mathbf{E}\{Z^4\})^{1/2}} \qquad (18)
$$

(see [16, p. 194] or [13, Lemma A.4]). Since X_j's are independent, identically distributed, and have zero mean, we have

$$
\mathbf{E}\left\{\left(\sum_{j=1}^n X_j\right)^2\right\} = n\mathbf{E}\{X_1^2\} = 2n/3 \ .
$$

On the other hand, expanding $(\sum_{j=1}^n X_j)^4$ yields

$$
\begin{aligned}
\mathbf{E}\left\{\left(\sum_{j=1}^n X_j\right)^4\right\} &= n\mathbf{E}\{X_1^4\} + 3n(n-1)(\mathbf{E}\{X_1^2\})^2\\
&= \frac{2n}{3} + 3n(n-1)\frac{4}{9}\\
&= \frac{2n(2n-1)}{3}\\
&\leq 4n^2/3 \ .
\end{aligned}
$$

Hence (18) gives

$$
\mathbf{E}\left\{\left|\sum_{j=1}^n X_j\right|\right\} \geq \frac{(2n/3)^{3/2}}{(4n^2/3)^{1/2}} = \frac{\sqrt{2n}}{3} \ .
$$

Combining this with (17), we conclude that for all $n \geq 1$,

$$D^*(\mu) - \mathbf{E}D(\mu_n, q_n^*) \geq \frac{1}{12\sqrt{2}} \frac{1}{\sqrt{n}} \ . \ \square$$

References

1. Pollard, D.: Strong consistency of k-means clustering. Annals of Statistics **9** (1981) 135–140
2. Pollard, D.: Quantization and the method of k-means. IEEE Transactions on Information Theory **IT-28** (1982) 199–205
3. Linder, T., Lugosi, G., Zeger, K.: Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. IEEE Transactions on Information Theory **40** (1994) 1728–1740
4. Linder, T.: On the training distortion of vector quantizers. IEEE Trans. Inform. Theory **IT-46** (2000) 1617–1623
5. Linder, T.: Learning-theoretic methods in vector quantization. In Györfi, L., ed.: Principles of nonparametric learning. Number 434 in CISM Courses and Lectures. Springer-Verlag, Wien, New York (2002) 163–210
6. Linder, T., Lugosi, G., Zeger, K.: Empirical quantizer design in the presence of source noise or channel noise. IEEE Trans. Inform. Theory **IT-43** (1997) 612–623
7. Merhav, N., Ziv, J.: On the amount of side information required for lossy data compression. IEEE Trans. Inform. Theory **IT-43** (1997) 1112–1121
8. Zeevi, A.J.: On the performance of vector quantizers empirically designed from dependent sources. In Storer, J., Cohn, M., eds.: Proceedings of Data Compression Conference, DCC'98, Los Alamitos, California, IEEE Computer Society Press (1998) 73–82
9. Bartlett, P., Linder, T., Lugosi, G.: The minimax distortion redundancy in empirical quantizer design. IEEE Transactions on Information Theory **IT-44** (1998) 1802–1813
10. Antos, A., Györfi, L., György, A.: Improved convergence rates in empirical vector quantizer design. In: Proceedings 2004 IEEE International Symposium on Information Theory, IEEE, IEEE Information Theory Society (2004) 301 (Chicago, IL, June 28–July 2, 2004.) Full paper submitted.
11. Chou, P.A.: The distortion of vector quantizers trained on n vectors decreases to the optimum as $O_p(1/n)$. In: Proceedings 1994 IEEE International Symposium on Information Theory. IEEE, IEEE Information Theory Society (1994) 457 (Trondheim, Norway, June 27–July 1, 1994.).
12. Pollard, D.: A central limit theorem for k-means clustering. Annals of Probability **10** (1982) 919–926
13. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Number 31 in Applications of Mathematics, Stochastic Modelling and Applied Probability. Springer-Verlag, New York (1996)
14. Slud, E.V.: Distribution inequalities for the binomial law. Annals of Probability **5** (1977) 404–412
15. Kolmogorov, A.N., Tikhomirov, V.M.: ϵ-entropy and ϵ-capacity of sets in function spaces. Translations of the American Mathematical Society **17** (1961) 277–364
16. Devroye, L., Györfi, L.: Nonparametric Density Estimation: The L_1 View. John Wiley, New York (1985)