On the Amount of Statistical Side Information Required for Lossy Data Compression *

Neri Merhav and Jacob Ziv

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL

January 12, 1998

Abstract

Consider a vector quantizer that is equipped with N side information bits of an arbitrary representation of the statistics of the input source. We investigate the minimum value of N such that rate-distortion performance of this quantizer would be essentially the same as the optimum quantizer for the given source.

Index Terms: vector quantization, rate-distortion, universal lossy source coding, empirical vector quantizer design, convergence rates.

*This research is partially supported by the U.S.-ISRAEL BNSF and by the Wolfson Research Awards administered by the Israel Academy of Sciences and Humanities.

1 Introduction

Let us consider an l-dimensional vector quantizer with $M = 2^{Rl}$ codewords, where R is the coding rate in bits per source symbol. Suppose that the quantizer has access to N side information bits of some arbitrary representation of the statistics of the source. We focus on the following question: What is the minimum value of N such that for every source, the rate-distortion performance could be essentially as good as that of the optimal rate R, l-dimensional quantizer?

In fact, one half of this question, the sufficiency part, has been answered recently by Linder, Lugosi, and Zeger [4]. Their results imply that if N is exponentially larger than M (i.e., $N \geq 2^{(R+\delta)l}$ for some $\delta > 0$), and the N side information bits represent a sequence of (finely quantized) i.i.d. l-dimensional training vectors, [1] then the distortion is essentially as small as $D_l(R)$, the minimum achievable distortion among all rate R, l-dimensional vector quantizers. We shall henceforth refer to this result as the *direct* theorem.

In this paper, we attempt to answer the second half of the above question, i.e., the necessity part. Specifically, we first show (in Section 2) that if N is exponentially slightly smaller than M, that is $N \leq 2^{(R-\delta)l}$, then no matter what representation of the source is used, there exists at least one probability density function (PDF) of l-vectors for which the distortion must be *significantly* larger than $D_l(R)$. For instance, one can find a PDF for which the distortion cannot be below $2D_l(R)$. We shall henceforth refer to this result as the *converse* theorem.

The direct and converse theorems together tell us, therefore, that $N = 2^{Rl}$ (in the exponential sense) is the minimum amount of side information, and there is an interesting "threshold effect" of a jump in the distortion when the exponent of N crosses the value R.

Both the direct theorem of Linder *et al.* and our converse theorem in Section 2 focus on PDFs of l-dimensional vectors, and it is not immediately apparent that these two results are applicable to stationary ergodic sources. In particular, the direct theorem requires independent training vectors, which are never quite available from any finite length sample unless the stationary ergodic source is memoryless. For memoryless sources, however, the necessary amount of training vectors need not grow with l because it is entirely dictated by

[1] In [4] the training vectors were not quantized. Nevertheless, in order to represent the training vectors by a finite number of bits as in our setting, we think of them as being quantized. This quantization, however, should be sufficiently fine so that the resulting additional distortion will be relatively small.

the one dimensional marginal PDF, which can be estimated efficiently from a relatively short data record. As for the converse theorem, we construct in Section 2, a set of counterexample PDFs, none of which is seemingly an lth order marginal of a stationary ergodic process.

In Section 3, we define a class of stationary ergodic sources w.r.t. which both the direct and the converse theorems still hold though in a slightly different formulation. However, for the direct part we still need to assume that the training vectors from the source are drawn independently, as in [4]. The construction of the class of sources in this section parallels that of Hershkovits and Ziv [3] and it will be detailed here for the sake of completeness.

Finally, we would like to mention some previous related work about the problem of characterizing the minimum amount of statistical side information. Wyner and Ziv [6] have investigated a classification problem in that setting: A classifier accepts an exact characterization of an lth order marginal of a stationary ergodic probability measure Q and N bits of partial information about the l-dimensional statistics of a possibly different stationary ergodic measure P. This classifier is required to decide whether $P = Q$ or else P and Q differ significantly in the sense that $D(P||Q)$ exceeds a prescribed threshold. How large should N be so that the right decision will be made for every P? More recently, Hershkovits and Ziv [3] have investigated the problem of lossless source coding from the same aspect: A lossless source encoder, operating on l-vectors, is informed of N information bits about the statistics of a stationary ergodic source P. How large should N be so that the entropy of the source would be achievable? For both questions the critical exponent value of N turns out to be intimately related to the l-th order entropy H_l. The intuition is that the necessary important information is carried by a set of typical l-sequences of P, and there are about 2^{lH_l} such sequences.

The present work is a natural extension of [3] from lossless to lossy source coding. The main message is that in rate-distortion coding, one no longer needs to know the set of *all* typical sequences, but actually, only the set of all Voronoi region centers of the optimum quantizer. Another interesting aspect of our results is that, similarly as in [3] and [6], their validity is not restricted merely to the asymptotic limit $l \to \infty$. Although we assume that l is large enough that certain quantities are negligible, we still do not require that l is so large that the best l-dimensional vector quantization performance is close to the asymptotic rate-distortion limit. This difference is meaningful for sources where $D_l(R)$ converges very slowly to the distortion-rate function $D(R)$, e.g., sources with very long memory.

2 PDFs of l-Dimensional Vectors

A rate R, l-dimensional vector quantizer Q is a measurable map from the l-dimensional Euclidean space $\mathbb{R}^l$ to a finite set of code words $\{y_1, ..., y_M\} \subset \mathbb{R}^l$, where $M = 2^{Rl}$. This map is defined by the nearest neighbor rule, i.e.,

$$Q(x) = y_i \qquad \text{if } \ ||x - y_i|| \leq ||x - y_j|| \ \text{ for all } j, \tag{1}$$

where $||\cdot||$ denotes the Euclidean norm, and ties are broken arbitrarily. Let X denote an l-dimensional random vector governed by a probability measure P. Given a quantizer Q, define the *distortion* (mean square error distortion) as

$$\Delta(Q) = \frac{1}{l} E||X - Q(X)||^2, \tag{2}$$

where E denotes expectation w.r.t. P. Let

$$D_l(R) = \min_Q \Delta(Q), \tag{3}$$

where the minimum is over all rate R, l-dimensional vector quantizers. The existence of a minimizing quantizer is proved in [5] under the assumption $E||X||^2 < \infty$. The function $D_l(R)$, the minimum attainable distortion at rate R and vector dimension l, can be thought of as the distortion-rate function of P w.r.t. dimension l. This distortion is achievable, of course, only if we design the vector quantizer on the basis of perfect knowledge of P.

But what happens if instead of full information about the lth order statistics P we are given only N information bits of an *arbitrary* representation of this information? This representation may take on many forms, e.g., a set of quantized values of the PDF over some grid, or some approximation of the characteristic function of P, or a set of training vectors, and so on. The question that we investigate here is the following: What is the minimum value of N, as a function of the rate R and the dimension l, such that there exists a quantizer, depending on the N side information bits, that essentially achieves the minimum distortion $D_l(R)$ for every P?

More precisely, the problem is defined as follows. Let $\mathcal{P}_l$ be a certain class of PDFs on the l-dimensional Euclidean space $\mathbb{R}^l$. An *N-bit representation* for sources in $\mathcal{P}_l$ is a deterministic mapping $F : \mathcal{P}_l \to \{0,1\}^N$. For every $b \in \{0,1\}^N$, let Q_b denote a rate R, l-dimensional vector quantizer associated with b. For a given $\epsilon > 0$ and a positive integer l, let $N_l(R, \epsilon)$ be the smallest positive integer N for which there exists an N-bit representation

F for $\mathcal{P}_l$ and a set of 2^N rate R, l-dimensional vector quantizers $\{Q_b, b \in \{0,1\}^N\}$, such that for every $P \in \mathcal{P}_l$

$$\Delta(Q_{F(P)}) \leq D_l(R) + \epsilon. \tag{4}$$

We would like to characterize the behavior of $N_l(R, \epsilon)$ as a function of R and l for small $\epsilon > 0$. In particular, we focus on the asymptotic behavior of $N_l(R, \epsilon)$ when $l \to \infty$ and R is held fixed. It will be assumed that as l grows, the sequence of classes $\{\mathcal{P}_l\}$ contains sources for which $D_l(R)$ is bounded away from zero for all l, and therefore ϵ can be chosen very small compared to $D_l(R)$.

For example, it is easy to see that if $\mathcal{P}_l$ is the class of all PDFs for which each coordinate of X is absolutely bounded with probability one by a constant $B > 0$, then

$$N_l(R, \epsilon) \leq 2^{Rl} \cdot l \log\left(\frac{B}{\sqrt{B^2 + \epsilon} - B}\right), \tag{5}$$

where $\log(\cdot)$ is defined to the base 2 throughout this paper. The right-hand side of this inequality is achieved if F maps P into a binary N-sequence formed by concatenating quantized versions of the code words $\{y_i\}$ of the optimum vector quantizer w.r.t. P. Each coordinate of each code word vector is quantized by a uniform scalar quantizer of $\log(B/\alpha)$ bits and step-size 2α, where $\alpha = \sqrt{B^2 + \epsilon} - B$. Since the quantization error in each coordinate never exceeds α, the overall extra distortion beyond $D_l(R)$ does not exceed $2B|\alpha| + \alpha^2 = \epsilon$, and so eq. (4) is satisfied. Thus roughly speaking, as l grows without bound, $N_l(R, \epsilon)$ is at most of the exponential order of 2^{Rl} in this example.

This simple example, however, has one drawback. The proposed representation F that is given by quantizing the code words of the optimum vector quantizer, is not available in reality if the source is unknown. In practical situations, the statistical side information is normally given in the form of random training data drawn from the same source, and the vector quantizer is designed empirically from the training data. Of course, if the training data is given in limited precision, then the total amount of side information bits N is finite. Thus, using the above terminology, the N-bit representation is given by a *random* rather than a deterministic mapping F in this case. Nevertheless, if one can claim the existence of a good random mapping, this is stronger than the parallel claim about a deterministic mapping. The latter follows from the former by invoking a simple 'random coding' argument: if for every $P \in \mathcal{P}_l$, $N = N_0$ bits of (quantized) random training data are sufficient to keep the *expected* distortion less than $D_l(R) + \epsilon$ (where the expectation involves the ensemble of training sets as well), then there must be a *deterministic* binary N_0-sequence $F(P)$ for which the distortion is as small, and so, $N_l(R,\epsilon) \leq N_0$. For these reasons, achievability results stated in terms of random training data are more desirable, though they provide merely an existence proof with no constructive strategy.

Linder, Lugosi, and Zeger [4, eq. (15)] have established a result in this spirit, though without quantization of the training data. This result, with slight modifications in the formalism, is summarized in the following theorem.

Theorem 1 *[4, eq. (15)] Let R and B be given positive constants, and let P be any member of the class $\mathcal{P}_l = \mathcal{P}_l(B)$ of all sources that satisfy $Pr\{||X||^2 \leq Bl\} = 1$. Let $Z = \{Z_1, Z_2, ..., Z_m\}$ be i.i.d. random vectors in $\mathrm{I\!R}^l$ drawn from P, independently of X. Let $Q^*(\cdot|Z)$ minimize $m^{-1}\sum_{i=1}^m ||Z_i - Q(Z_i)||^2$ over all rate R, l-dimensional vector quantizers, and let*

$$D_l(R|Z) = \frac{1}{l} E\{||X - Q^*(X|Z)||^2|Z\}, \tag{6}$$

where the expectation is taken w.r.t. the ensemble of X. Then,

$$ED_l(R|Z) \leq D_l(R) + 16\sqrt{2}Bl\sqrt{(l+1)2^{Rl}+1}\sqrt{\frac{\log m}{m}} + o\left(\sqrt{\frac{\log m}{m}}\right), \tag{7}$$

where the expectation is taken w.r.t. the ensemble of Z.

In words, if m is large the performance of the empirically-optimum quantizer is essentially as good as that of the optimum quantizer on the average. Now, if we fix $\delta > 0$ and let $m = 2^{(R+\delta)l}$, the excess distortion beyond $D_l(R)$ in eq. (7) vanishes as $l \to \infty$. If each training vector Z_i is quantized into kl bits, then the quantized training set is represented by $N = klm = kl2^{(R+\delta)l}$ bits. If, in addition, k is sufficiently large (though fixed), then the additional distortion due to quantization of the training data can be made negligibly small. This follows from the following consideration.

Let us cover the sphere of radius $\sqrt{Bl}$ with non-overlapping l-dimensional cubes of size 2ϵ, centered at points whose coordinates are integer multiples of 2ϵ. Now suppose that every training vector Z_i, whose norm is less than $\sqrt{Bl}$, is quantized to the center of its cube. By doing this, we cause a quantization error whose absolute value never exceeds ϵ in each coordinate, and the number of bits is approximately the logarithm of the ratio between the volume of the sphere and the volume of the cube, i.e., $k \approx 0.5\log[\pi eB/(2\epsilon^2)]$. Now, by using the quantized training data, we obtain a sequence of empirically-designed quantizers that tends to the optimum for the probability distribution $\tilde{P}$ of these quantized training vectors. But since the quantization error of the training data is uniformly small and the support of P and $\tilde{P}$ is bounded, then as $\epsilon \to 0$, the expected distortion of every Q w.r.t. $\tilde{P}$ tends to $\Delta(Q)$ uniformly. Therefore, an optimum quantizer for $\tilde{P}$ is nearly optimum for P.

Thus, we are again led to the conclusion that the exponential growth rate of $N_l(R, \epsilon)$ does not exceed R. This time, however, this conclusion was reached more generally from the viewpoint of random training set representations. This result is now stated formally in the following theorem.

Theorem 2 *Let $R > 0$ be given and let $\mathcal{P}_l$ be defined as in Theorem 1. Then, for every $\epsilon > 0$,*

$$\limsup_{l \to \infty} \frac{1}{l} \log N_l(R, \epsilon) \leq R. \tag{8}$$

We now state a converse to Theorem 2 that tells us that not only the converse inequality holds true as well, but moreover, if N is of exponential order strictly less than 2^{Rl}, the distortion must be *significantly* larger than $D_l(R)$ at least for one source.

Theorem 3 *Let $R > 0$ be given and let $\mathcal{P}_l$ be defined as in Theorem 1. Then,*

$$\lim_{\epsilon \to 0} \liminf_{l \to \infty} \frac{1}{l} \log N_l(R, \epsilon) \geq R. \tag{9}$$

Furthermore, for every $\epsilon > 0$ and $\delta > 0$, if $N < 2^{(R-\delta)l}$ and l is sufficiently large, then for any deterministic N-bit representation $F : \mathcal{P}_l \to \{0,1\}^N$ and any set of 2^N rate R, l-dimensional vector quantizers $\{Q_b : \ b \in \{0,1\}^N\}$, there exists a PDF $P \in \mathcal{P}_l$ such that $D_l(R) > \epsilon$, and at the same time

$$\Delta(Q_{F(P)}) > 2D_l(R). \tag{10}$$

At this point, a few comments are in order.

- The reason for requiring $D_l(R) > \epsilon$ is to guarantee that eq. (10) contradicts the achievability inequality (4).
- Note that Theorem 3 is stated for *deterministic* representations and hence is stronger than a strict converse to Theorem 1.

- Clearly, the combination of Theorem 2 and the first part of Theorem 3 establishes the fact that for $\mathcal{P}_l$ defined as in Theorem 1,

$$\lim_{\epsilon \to 0} \lim_{l \to \infty} \frac{1}{l} \log N_l(R, \epsilon) = R. \tag{11}$$

- The factor of two on the right-hand side of eq. (10) is immaterial. In fact, it can be replaced by any arbitrarily large (but finite) number.

- The significance of Theorem 3 is primarily in sharpening the result of Linder *et al.* by claiming that from the viewpoint of vector quantization, there is essentially no more efficient way to represent a source than that of using independent training vectors.

The remaining part of this section is devoted to the proof of Theorem 3.

Proof. Since the first part of the theorem follows from the second part, it will be sufficient to prove the second part. The main idea of the proof of the second part is in applying a "sphere covering" argument similar to that in [6] and [3]. We shall construct a counterexample set of sufficiently many PDFs that are "far apart" from one another in the sense that only a small fraction of them can be quantized by a *single* quantizer with distortion less than $2D_l(R)$. Thus, if N is not large enough, then any set of 2^N different vector quantizers cannot possibly "cover" the whole family of PDFs, and therefore there must be at least one PDF for which the distortion exceeds $2D_l(R)$.

Let R, δ, and ϵ be given positive reals defined as in Theorem 3. Select two positive reals D_0 and A, so that $D_0 > 2\epsilon$ and $A > 36D_0 4^R$. (The reason for these choices will become apparent shortly.) Construct a set of K l-dimensional vectors $\{u_1, u_2, ..., u_K\}$ in the following manner: The first vector u_1 is chosen arbitrarily from the sphere $S_0(\sqrt{Al})$, where $S_u(r)$ denotes the l-dimensional sphere of radius r centered at u. The second vector u_2 is chosen arbitrarily from $S_0(\sqrt{Al}) - S_{u_1}(6\sqrt{lD_0})$, the third vector u_3 is selected from $S_0(\sqrt{Al}) - [S_{u_1}(6\sqrt{lD_0}) \bigcup S_{u_2}(6\sqrt{lD_0})]$, and so on. This procedure terminates when $S_0(\sqrt{Al})$ is exhausted. The total number K of vectors generated by this procedure is lower bounded by the volume of a sphere of radius $\sqrt{Al}$ divided by the volume of a sphere of radius $6\sqrt{lD_0}$, i.e.,

$$K \geq \frac{\mathrm{Vol}\{S_0(\sqrt{Al})\}}{\mathrm{Vol}\{S_{u_1}(6\sqrt{lD_0})\}} = \exp_2\left[\frac{l}{2}\log\left(\frac{A}{36D_0}\right)\right] \triangleq 2^{Gl}. \tag{12}$$

The above choice of A guarantees that $G > R$ and hence $K >> M = 2^{Rl}$ for large l.

Consider now a finite class $\mathcal{W}$ of PDFs defined as follows. Each PDF corresponds to a particular subset of M out of the K vectors $u_1, ..., u_K$. To avoid cumbersome notation, let us re-index the M vectors associated with a given source in $\mathcal{W}$ as $u_1, ..., u_M$. For a given subset of vectors $u_1, ..., u_M$, the corresponding PDF is defined as

$$P(x) = \frac{1}{M}\sum_{i=1}^{M} H(x - u_i), \qquad\qquad x \in \mathbb{R}^l, \tag{13}$$

where $H(x)$ is the uniform PDF on the *surface* of $S_0(\sqrt{lD_0})$. In words, x is uniformly distributed over all the (disjoint) surfaces of spheres of radius $\sqrt{lD_0}$ centered at u_i, $i = 1, 2, ..., M$. It is shown in Appendix A, that for each such source,

$$D_0 \geq D_l(R) \geq D_0 - \zeta_l, \tag{14}$$

where $\zeta_l \to 0$ as $l \to \infty$, hence $D_l(R) > 2\epsilon - \zeta_l > \epsilon$ for all large enough l. Since all PDFs in $\mathcal{W}$ emit vectors whose norms never exceed $(\sqrt{A} + \sqrt{D_0})^2 l$, then every PDF of $\mathcal{W}$ is a member of $\mathcal{P}_l$ with $B = (\sqrt{A} + \sqrt{D_0})^2$. Thus, we know that all these sources are in $\mathcal{P}_l$ and they all satisfy $D_l(R) > \epsilon$, as required in the assertion of Theorem 3.

We would now like to upper bound the number of PDFs in $\mathcal{W}$ for which a single quantizer distorts by less than $2D_0$. Suppose that we have a quantizer Q that induces distortion less than $2D_0$ for a given PDF in $\mathcal{W}$. Let

$$\Delta_i = \frac{1}{l}E\{||X - Q(X)||^2 | X \in S_{u_i}(\sqrt{lD_0})\}, \qquad i = 1, ..., M. \tag{15}$$

Then, by our hypothesis

$$2D_0 \geq \frac{1}{M}\sum_{i=1}^{M} \Delta_i, \tag{16}$$

or, equivalently,

$$\frac{1}{2} \geq \frac{\frac{1}{M}\sum_{i=1}^{M}\Delta_i}{4D_0} \geq \frac{1}{M}|\{i:\ \Delta_i \geq 4D_0\}|, \tag{17}$$

where the second inequality follows from Chebychev's inequality. This means that more than $M/2$ spheres associated with the PDF contribute distortion less than $4D_0$. But in order for a certain sphere $S_{u_i}(\sqrt{lD_0})$ to contribute distortion less than $4D_0$, there must be at least one code word y_j within distance $3\sqrt{lD_0}$ from the center of that sphere. By construction of the set $\{u_1, ..., u_K\}$ and the triangle inequality, it is clear that if a certain code word is at distance less than $3\sqrt{lD_0}$ from one sphere center u_i, it must be at a larger distance from any other sphere center. Consequently, in order for a quantizer to induce distortion less than $2D_0$ for as many sources as possible, it is necessary that every code vector y_i be at distance less than $3\sqrt{lD_0}$ from some source center u_i. Therefore, the maximum number L of PDFs in $\mathcal{W}$ that can be covered by one quantizer in the sense of providing distortion less than $2D_0$ is bounded by

$$\begin{aligned} L &\leq \sum_{i=M/2}^{M} \binom{M}{i} \binom{K-M}{M-i} \\ &\leq \sum_{i=M/2}^{M} \binom{M}{i} \binom{K}{M-i} \\ &\leq M \binom{M}{M/2} \binom{K}{M/2}, \end{aligned} \tag{18}$$

where we have assumed, without essential loss of generality, that M is even. The first inequality follows from the fact that for every quantizer, the count of PDFs corresponding to distortion less than $2D_0$ must contain sources for which at least $M/2$ PDF centers $\{u_i\}$ are chosen in the vicinity of code vectors while the other centers may be chosen freely from the remaining $K-M$ sphere centers $\{u_i\}$. The first binomial coefficient on the right-most side of eq. (18) is upper bounded by 2^M. As for the second binomial coefficient, we will now use the following chain of inequalities for arbitrary nonnegative integers m and n, where $m \leq n$:

$$\begin{aligned} \log \binom{n}{m} &\leq nh(\frac{m}{n}) \\ &= m \log \frac{n}{m} + (n-m) \log \frac{1}{1-\frac{m}{n}} \\ &\leq m \log \frac{n}{m} + (n-m)(\frac{1}{1-\frac{m}{n}} - 1) \log e \\ &= m(\log \frac{n}{m} + \log e), \end{aligned} \tag{19}$$

where $h(\cdot)$ is the binary entropy function. The first inequality can be found in [1] and the second inequality follows from $\ln u \leq u - 1$. In a similar manner [2, p. 285, eq. (12.5.2)] (see also [1]),

$$\begin{aligned} \log \binom{n}{m} &\geq nh(\frac{m}{n}) - \log(n+1) \\ &\geq m \log \frac{n}{m} - \log(n+1). \end{aligned} \tag{20}$$

By applying (19) to eq. (18), we get

$$L \leq \exp_2 \left[\frac{M}{2} (\log \frac{K}{M} + 3 + \log e) + \log M \right]. \tag{21}$$

By applying eq. (20), we get the following lower bound on the total number of sources in $\mathcal{W}$:

$$|\mathcal{W}| = \left(\begin{array}{c} K \\ M \end{array}\right) \geq \exp_2[M \log \frac{K}{M} - \log(K+1)]. \tag{22}$$

The ratio $|\mathcal{W}|/L$, which expresses the minimum number of quantizers needed to "cover" $\mathcal{W}$, is therefore lower bounded by

$$\frac{1}{L}\left(\begin{array}{c} K \\ M \end{array}\right) \geq \exp_2\{\frac{M}{2}[\log \frac{K}{M} - 3 - \log e] - \log M - \log(K+1)\}. \tag{23}$$

Now, $M = 2^{Rl}$ and $K \geq 2^{Gl}$. Since $[0.5M \log K - \log(K+1)]$ is a monotone nondecreasing function of K for every $M \geq 2$, then substituting 2^{Gl} instead of K would further decrease the right most side of eq. (23), and we obtain

$$\frac{1}{L}\left(\begin{array}{c} K \\ M \end{array}\right) \geq \exp_2\left\{\frac{1}{2}2^{Rl}[l(G-R) - 3 - \log e] - Rl - \log(2^{Gl}+1)\right\}. \tag{24}$$

This quantity in turn, is larger than $2^{2^{(R-\delta)l}}$ for every $\delta > 0$ (chosen above) and sufficiently large l. Thus, if the total number of quantizers 2^N is less than this number, that is, if $N < 2^{(R-\delta)l}$, there must be at least one PDF with distortion larger than $2D_0 \geq 2D_l(R)$. This completes the proof of Theorem 3. □

3 Stationary and Ergodic Processes

So far we focused on lth order marginals of sources without any concern as to whether these marginals can be obtained from any stationary ergodic sources. Indeed, we are not aware of the existence of a stationary ergodic process whose lth order marginal agrees exactly with the PDF of any of the sources constructed in the proof of Theorem 3.

Our main concern in this section is to provide a converse theorem for stationary and ergodic processes. Roughly speaking, the direct theorem for processes will be largely a re-statement of Theorem 1 where now the l-th order marginal P should be thought of as being derived from a process. In other words, we still require independent training vectors similarly as in [4]. Strictly speaking, this assumption cannot be met for a non-memoryless process. However, it can be approached provided that the memory of the process fades away and the time gap between consecutive training vectors is sufficiently large. It is an open problem, however, to prove the direct part for *dependent* training vectors drawn from the underlying process.

Before we turn to the converse theorem for processes, we first extend Theorem 1 so as to apply to a broader class of sources than in Section 2. The reason for this extension is that the counterexample processes that will be constructed in the proof of the forthcoming converse theorem will have lth order marginal PDFs with unbounded support, and hence will not belong to $\mathcal{P}_l$ of Section 2. Specifically, we define a class of stationary processes $\mathcal{M}$ as follows.

For a given stationary process μ, let $\sigma^2(\mu) = E|X_1|^2$, where X_1 is the first coordinate of X. For a given $\epsilon > 0$, and a given l, let $B(\mu, \epsilon, l)$ denote the infimum value of B such that

$$\frac{1}{l}E\left(||X||^2 \cdot 1\{||X||^2 > Bl\}\right) \leq \epsilon, \tag{25}$$

where X is a random vector in $\mathbb{R}^l$ drawn from μ, and $1\{\cdot\}$ is the indicator function of an event. For two given positive reals σ^2 and B_0, and a given function $L_0(\cdot)$, let $\mathcal{M} = \mathcal{M}(\sigma^2, B_0, L_0)$ be the set of all stationary processes $\{\mu\}$ that uniformly satisfy the following conditions:

1. $\sigma^2(\mu) \leq \sigma^2$.

2. For every $\epsilon > 0$, $l \geq L_0(\epsilon)$ implies $B(\mu, \epsilon, l) \leq B_0$.

Generally speaking, $\mathcal{M}$ is a class of stationary processes for which there is a uniform bound on the second order moment, and the 'tails' decay uniformly rapidly in a certain sense. For example, Gaussian processes and finite mixtures of Gaussian processes with variance less than or equal to σ^2 are all in $\mathcal{M}$ for a certain choice of B_0. The following is an extension of Theorem 1 for processes in $\mathcal{M}$.

Theorem 4 *Let $R > 0$ be given and let $Z = (Z_1, ..., Z_m)$, $m = 2^{(R+\delta)l}$, be i.i.d. random l-vectors drawn from μ. For a given $B > 0$, let $Q(\cdot|Z)$ minimize $\sum_{i \in I} ||Z_i - Q(Z_i)||^2$ over all rate R, l-dimensional quantizers, where I is the set of all $1 \leq i \leq m$ for which $||Z_i||^2 \leq Bl$. Then, for every $\epsilon > 0$ there exists a sufficiently large $B > 0$ such that for all sufficiently large l,*

$$\frac{1}{l}E\{E[||X - Q(X|Z)||^2|Z]\} \leq D_l(R) + \epsilon \tag{26}$$

for every $\mu \in \mathcal{M}$.

The proof of Theorem 4 appears in Appendix B.

Note that Theorem 4 makes a claim about an empirically-designed quantizer that is trained selectively only on training vectors whose norms fall within a certain bound. It does not reflect a belief that this is the best training strategy, but it makes the proof easier.

Let us now turn to the converse part for stationary processes. For some positive constants σ^2, B_0, and C_1 and C_2 (to be determined later), let us define $\mathcal{M} = \mathcal{M}(\sigma^2, B_0, L_0)$, where $L_0(\epsilon) = C_1 \log(C_2/\epsilon)$.

Theorem 5 *Let $R > 0$ be given and let $\mathcal{M}$ be defined as above. Then, for every $\epsilon > 0$ and $\delta > 0$, if $N < 2^{(R-\delta)l}$ and l is sufficiently large, then for any deterministic N-bit representation $F : \mathcal{P}_l \to \{0,1\}^N$ and any set of 2^N rate R, l-dimensional vector quantizers $\{Q_b, b \in \{0,1\}^N\}$, there exists a stationary and ergodic process $\mu \in \mathcal{M}$ whose lth order marginal PDF P satisfies $D_l(R) > 2\epsilon$, and at the same time*

$$\Delta(Q_{F(P)}) > 2D_l(R) - \epsilon. \tag{27}$$

The remaining part of this section is devoted to the proof of Theorem 5.

Proof. The idea of the proof is to construct a set of stationary and ergodic processes in $\mathcal{M}$ whose lth order marginals are nearly the same as those constructed in the proof of Theorem 3. The construction will be based on the same ideas. We first describe the set of counterexample processes, and then prove that:

(a) The processes in this set are all members of $\mathcal{M}$, and

(b) The assertion of Theorem 5 holds.

For a given R, δ, and ϵ, let us select $D_0 > 2\epsilon$ and $A > 36D_0 4^R$ as in the proof of Theorem 3. Consider again the l-dimensional sphere $S_0(\sqrt{Al})$. As in Theorem 1, we first generate a collection of vectors $\{u_i\}$ that will serve as centers of spheres of radius $\sqrt{lD_0}$. But now, for reasons that will become apparent later, we would like to guarantee that the set of all u-vectors *as well as all their cyclic shifts* are at distance at least $6\sqrt{lD_0}$ from each other.

Before we describe how this is done, we introduce some new notation. For a given $u \in \mathbb{R}^l$, let Tu denote the one-step right cyclic shift of u. We shall think of the operator T as the $l \times l$ permutation matrix that performs this operation. Thus, T^i causes i cyclic shifts to the right while T^{-i} causes i cyclic shifts to the left.

Let $R < G < 0.5 \log(A/36D_0)$ and let $K = 2^{Gl}$. Next, perform the following steps:

1. Select an arbitrary vector $u_1 \in S_0(\sqrt{Al})$ such that $||T^i u_1 - T^j u_1||^2 \geq 36lD_0$ for every $i, j = 0, 1, ..., l-1$, $j \neq i$.

2. For $m = 2, 3, ...K$, define the remainder set

$$W_m = S_0(\sqrt{Al}) - \bigcup_{j=1}^{m-1} \bigcup_{i=0}^{l-1} S_{T^i u_j}(6\sqrt{lD_0}) \tag{28}$$

   and select $u_m \in W_m$ such that the following conditions are met:

   (c1) $T^i u_m \in W_m$ for all $i = 0, 1, ..., l-1$.

   (c2) $||T^i u_m - T^j u_m||^2 \geq 36lD_0$ for every $i, j = 0, 1, ..., l-1$, $j \neq i$.

After completing this procedure, one has generated a set of Kl vectors which can be partitioned into K disjoint subsets, each of which includes l cyclically shifted versions of a certain *representative* vector u_i. All vectors, including the cyclic shifts are at distance at least $6\sqrt{lD_0}$ from each other.

The reader might wonder whether it is always possible to find at each step a vector u_m that satisfies Conditions (c1) and (c2). It is shown in Appendix C, that not only is such a choice possible, but moreover, most points (in the Lebesgue measure sense) in W_m satisfy these conditions when l is large. The intuition is that the union of the small spheres is very small compared to the big sphere, and that a randomly chosen vector in a sphere looks typically almost like an i.i.d. vector and hence is essentially orthogonal to its cyclic shifts.

We shall now construct stationary and ergodic processes from subsets of the representative vectors $\{u_i\}$ in the following manner. Let $M = 2^{Rl}/l$, and consider the collection of all subsets of M out of K representative vectors. Every process in the class we construct corresponds to a certain combination of M representative vectors, hereafter re-indexed $u_1, ..., u_M$. Let U_i denote the (kl)-dimensional vector formed by k concatenated repetitions of u_i, $i = 1, 2, ..., M$, that is, $U_i = (u_i, u_i, ..., u_i)$. The source will be defined by a finite-state machine (FSM) that generates a sequence of random variables $..., X_{-1}, X_0, X_1, ...$ in the following way (see also [3]). Suppose that at time n the machine is in a state labeled 0. Then, the following steps are performed.

1. Select at random with uniform distribution an integer $I \in \{1, 2, , ..., M\}$ and a binary random variable $\beta \in \{0, 1\}$, independent of I, with probability $\Pr\{\beta = 1\} = \alpha$, where $0 < \alpha < 1$ is a small number.

2. If $\beta = 0$, set $(X_n, ..., X_{n+kl-1}) = U_I + V$, where V is a (kl)-dimensional random vector, independent of I, with i.i.d. zero-mean, variance-D_0 Gaussian components. During the time interval $(n+1, ..., n+kl-1)$ the FSM goes through states labeled $(I,1,0), (I,2,0), ..., (I, kl-1, 0)$ and then returns to state 0. Set the time counter n at $n + kl$ and go to 1.

3. If $\beta = 1$, select a random integer $J \in \{l+1, l+2, ..., kl-1\}$ with uniform distribution. Set $(X_n, ..., X_{n+J-1})$ as the J first components of U_I, contaminated by an additive noise vector $V \in \mathbb{R}^J$ (independent of I and J) with i.i.d. zero-mean, variance-D_0 Gaussian components. During the time interval $(n+1, ..., n+J-1)$ the FSM goes through states labeled $(I,1,J), (I,2,J), ..., (I, J-1, J)$, and then returns to state 0. Set the time counter n at $n + J$ and go to 1.

We shall henceforth refer to the process of concatenated U-vectors and the Gaussian noise process as the *U-process*, and the *V-process*, respectively. Throughout the sequel we shall re-index the corresponding sequences of random variables as $\{U_n\}$ and $\{V_n\}$, respectively, according to the time indexes of $\{X_n\}$. The probability measures of l-vectors associated with the U-process and the V-process will be denoted P_U and P_V, respectively. Throughout the sequel X_i^j, $i \leq j$, will denote the segment $(X_i, ..., X_j)$. Similar notations will be used for segments of the U-process and the V-process.

Since $\{U_n\}$ is a two-sided, irreducible, aperiodic finite-state process it is stationary and ergodic. Since $\{V_n\}$ is an independent i.i.d. process, then $\{U_n\}$ and $\{V_n\}$ are are jointly stationary and ergodic and hence so is $X_n = U_n + V_n$. Moreover, due to the underlying finite-state process, $\{X_n\}$ has an exponentially fast vanishing memory (see also [3]).

We next show that each process $\{X_n\}$ in the set we constructed is also in $\mathcal{M}$. First, observe that for every process in this set,

$$E|X_1|^2 = \frac{1}{l}E||X||^2 = \frac{1}{l}E||U_0^{l-1}||^2 + D_0. \tag{29}$$

Since every realization of U_0^{l-1} is either a cyclic shift of some u_i or it includes the boundary between two different u-vectors, then its norm is bounded by $2Al$, and so, the first condition for membership in $\mathcal{M}$ holds with $\sigma^2 = 2A + D_0$. As for the second condition, it is easy to check that for a Gaussian l-vector Y, with mean vector u and independent components with variance D_0,

$$Ee^{s||Y||^2} = \exp\left[\frac{s||u||^2}{1-2sD_0} - l\log(1-2sD_0)\right], \qquad s < \frac{1}{2D_0}. \tag{30}$$

Now,

$$\begin{aligned} E\left(||Y||^2 \cdot 1\{||Y||^2 > Bl\}\right) &\leq E\left[||Y||^2 \cdot e^{s(||Y||^2 - Bl)}\right] \\ &= e^{-sBl}\frac{\partial}{\partial s}Ee^{s||Y||^2} \\ &= e^{-sBl} \cdot \frac{||u||^2 + 2D_0 l(1-2sD_0)}{(1-2sD_0)^2} \cdot \\ &\quad \exp\left[\frac{s||u||^2}{1-2sD_0} - l\log(1-2sD_0)\right]. \end{aligned} \tag{31}$$

Again, since X_0^{l-1} is distributed according to a finite mixture of Gaussian PDFs with mean vectors depending on $\{u_i\}$ and their concatenations, and with covariance matrix $D_0 I$ (I being the $l \times l$ identity matrix), and since the norm of any underlying mean vector never exceeds $2Al$, it is readily seen that

$$\begin{aligned} &E\left(||X_0^{l-1}||^2 \cdot 1\{||X_0^{l-1}||^2 > Bl\}\right) \leq \\ &2le^{-sBl} \cdot \frac{A + D_0(1-2sD_0)}{(1-2sD_0)^2} \cdot \exp\left\{l\left[\frac{2sA}{1-2sD_0} - \log(1-2sD_0)\right]\right\}. \end{aligned} \tag{32}$$

Thus, by simple algebraic manipulation, it is easy to verify that for a given allowable choice of A, D_0, and s, the last expression normalized by l, can be made less than ϵ, provided that B is at least as large than some constant B_0 that satisfies

$$B_0 > \frac{2sA}{1-2sD_0} - \log(1-2sD_0), \tag{33}$$

and then

$$L_0(\epsilon) = \frac{\log\{2[A + D_0(1-2sD_0)]/[\epsilon(1-2sD_0)^2]\}}{B_0 - 2sA/(1-2sD_0) + \log(1-2sD_0)}. \tag{34}$$

This defines the constants C_1 and C_2 that were mentioned before the statement of Theorem 5.

It remains to demonstrate that at least one of the sources constructed above satisfies the assertion of the theorem if N is not large enough. The underlying idea is that most of the time the source creates long repetitions of u-vectors, and therefore, the quantizer, which is not necessarily synchronized to the source, will "see" noisy versions of cyclic shifts of the u_i's. By the law of large numbers, the l-dimensional noise vectors will usually lie near the surface of $S_0(\sqrt{lD_0})$. Since there are $2^{Rl}/l$ representative center vectors and l cyclic shifts of each one, the total number of sphere centers is 2^{Rl}. Due to the occasional random selection of J in step 3, the phase of the cyclic shift will be random.

More precisely, consider the vector X_0^{l-1} as the current input of the quantizer. Define the events

$$E_1 = \left\{U_0^{l-1}: \text{ State 0 has not been visited in the time interval } \{0,1,...,l-1\}\right\} \tag{35}$$

and

$$E_2 = \left\{V_0^{l-1}: \ |\frac{1}{l}\sum_{i=0}^{l-1} V_i^2 - D_0| \leq \epsilon_0\right\} \tag{36}$$

for some small $\epsilon_0 > 0$. Note that E_1 is defined in terms of the underlying U-process while E_2 corresponds to the noise process. Hence E_1 and E_2 are independent events. Both events have high probability when k and l are large and α is small. Let us denote $\xi = P(E_1^c)+P(E_2^c)$ and select k, l, and α such that ξ would be arbitrarily small. Thus, the joint event $E = E_1 \cap E_2$ has probability at least $1-\xi$.

Given the event E, the lth order marginal is essentially (for small ϵ_0) as in the proof of Theorem 3, and hence according to this theorem, there exists a process in the class for which the distortion must exceed $2(D_0-\epsilon_0)$. Thus, for any N-bit representation F and any set of 2^N quantizers $\{Q_b\}$, there must be a process in the set we have defined for which the overall distortion for the worst process in the class is therefore lower bounded by

$$\begin{aligned}
& E\{||X_0^{l-1} - Q_{F(P)}(X_0^{l-1})||^2\} = E\{||U_0^{l-1} + V_0^{l-1} - Q_{F(P)}(U_0^{l-1} + V_0^{l-1})||^2\} \\
\geq \ & \int_{u_0^{l-1}\in E_1}\int_{v_0^{l-1}\in E_2} ||u_0^{l-1} + v_0^{l-1} - Q_{F(P)}(u_0^{l-1} + v_0^{l-1})||^2 P_U(du_0^{l-1})P_V(dv_0^{l-1}) \\
= \ & P(E_1 \cap E_2)\int ||u_0^{l-1} + v_0^{l-1} - Q_{F(P)}(u_0^{l-1} + v_0^{l-1})||^2 P_U(du_0^{l-1}|E_1)P_V(dv_0^{l-1}|E_2) \\
\geq \ & (1-\xi)\cdot 2(D_0-\epsilon_0).
\end{aligned} \tag{37}$$

Since $D_l(R)$ is essentially D_0 for all sources in this class, the proof is completed by an appropriate choice of ϵ_0.

Appendix A

Proof of eq. (14).

The upper bound to $D_l(R)$ is obvious since D_0 is achievable by the quantizer whose codewords are $y_i = u_i$, $1 \leq i \leq M$. As for the lower bound, we have the following. Let

$I(X;\hat{X})$ denote the (unnormalized) mutual information between the source vector X and the reproduction vector $\hat{X}$. Then,

$$D_l(R) \geq \min \frac{1}{l} E\{||X - \hat{X}||^2\} \tag{A.1}$$

where the minimization is over the set of all channels $P(\hat{X}|X)$ such that $I(X;\hat{X}) \leq Rl$. Now,

$$I(X;\hat{X}) = h(X) - h(X|\hat{X}) \tag{A.2}$$

where $h(X)$ is the differential entropy of X and $h(X|\hat{X})$ is the conditional differential entropy of X given $\hat{X}$. Since X is uniformly distributed over the surfaces of 2^{Rl} disjoint spheres of radius $\sqrt{lD_0}$, then

$$h(X) = Rl + \log \mathrm{Surf}\{S_0(\sqrt{lD_0})\} \tag{A.3}$$

where $\mathrm{Surf}\{S_0(\sqrt{lD_0})\}$ designates the surface area of each such sphere. Thus,

$$\begin{aligned} I(X;\hat{X}) &= Rl + \log \mathrm{Surf}\{S_0(\sqrt{lD_0})\} - h(X|\hat{X}) \\ &= Rl - h(X - \hat{X}|\hat{X}) + \log \left[\frac{2\pi^{l/2}(lD_0)^{(l-1)/2}}{,\,(l/2)}\right] \\ &\geq Rl - h(X - \hat{X}) + \frac{l}{2}\log[2\pi e(D_0 - \zeta_l)], \end{aligned} \tag{A.4}$$

where in the last step, $\zeta_l \to 0$ as $l \to \infty$, following Stirling's formula. Denoting $\tilde{X} \triangleq X - \hat{X}$, we then further lower bound $D_l(R)$ by

$$\min \frac{1}{l} E||\tilde{X}||^2 \tag{A.5}$$

where the minimum is over all random vectors $\tilde{X}$ such that

$$h(\tilde{X}) \geq \frac{l}{2}\log[2\pi e(D_0 - \zeta_l)]. \tag{A.6}$$

Since the right-hand side of the last inequality is the maximum entropy of a random vector whose expected norm does not exceed $l(D_0 - \zeta_l)$, the minimum in eq. (A.5) is obviously $D_0 - \zeta_l$. This completes the proof of eq. (14).

Appendix B

Proof of Theorem 4.

The idea of the proof is to approximate the lth order marginal P of the process μ by a bounded support probability measure like in Section 2, and to show that the contribution of vectors that fall outside the bounded support set is negligibly small.

We first show that there is no essential loss in optimality if the lth order PDF P of any process in $\mathcal{M}$ is treated as having the bounded support $S_0(\sqrt{Bl}) = \{x: \; ||x||^2 \leq Bl\}$, where the constant $B > 0$ is sufficiently large but independent of the specific process in $\mathcal{M}$. For a given lth order marginal P of $\mu \in \mathcal{M}$, and $B > 0$, let

$$P_B(x) = P(x|x \in S_0(\sqrt{Bl})) = \begin{cases} P(x)/P\{S_0(\sqrt{Bl})\} & x \in S_0(\sqrt{Bl}) \\ 0 & x \in S_0^c(\sqrt{Bl}) \end{cases} \tag{B.1}$$

where $S_0^c(\sqrt{Bl})$ denotes the complimentary set. Note that by definition of $\mathcal{M}$, $P\{S_0(\sqrt{Bl})\}$ is arbitrarily close to one for large B, because for $B > 1$,

$$P\{S_0^c(\sqrt{Bl})\} = \int_{S_0^c(\sqrt{Bl})} P(dx) \leq \frac{1}{l} \int_{S_0^c(\sqrt{Bl})} ||x||^2 P(dx) \leq \epsilon. \tag{B.2}$$

Let Q^B denote an optimal rate R, l-dimensional quantizer for P_B. Then, the per-letter distortion of Q^B w.r.t. P can be decomposed as follows.

$$\begin{aligned} \Delta(Q^B) &= \frac{1}{l} \int_{S_0(\sqrt{Bl})} ||x - Q^B(x)||^2 P(dx) + \frac{1}{l} \int_{S_0^c(\sqrt{Bl})} ||x - Q^B(x)||^2 P(dx) \\ &\stackrel{\Delta}{=} I_1 + I_2 \end{aligned} \tag{B.3}$$

Now, the first term I_1 is upper bounded by

$$\begin{aligned} I_1 &= \frac{1}{l} \int_{S_0(\sqrt{Bl})} ||x - Q^B(x)||^2 P(dx) \\ &= \frac{1}{l} \min_Q \int_{S_0(\sqrt{Bl})} ||x - Q(x)||^2 P(dx) \\ &\leq \frac{1}{l} \min_Q \int_{\mathbb{R}^l} ||x - Q(x)||^2 P(dx) \\ &= D_l(R), \end{aligned} \tag{B.4}$$

and the second term I_2 is upper bounded by

$$\begin{aligned} I_2 &= \frac{1}{l} \int_{S_0^c(\sqrt{Bl})} ||x - Q^B(x)||^2 P(dx) \\ &\leq \frac{1}{l} \int_{S_0^c(\sqrt{Bl})} (||x|| + \sqrt{Bl})^2 P(dx) \\ &\leq \frac{4}{l} \int_{S_0^c(\sqrt{Bl})} ||x||^2 P(dx). \end{aligned} \tag{B.5}$$

The first inequality follows from the fact Q^B must have all its code words in $S_0(\sqrt{Bl})$, and so $||x-Q^B(x)||^2$ cannot exceed the distance of x from the most distant point in $S_0(\sqrt{Bl})$, which is $(||x||+\sqrt{Bl})^2$. The second inequality follows from the fact that $(||x||+\sqrt{Bl})^2 \leq 4||x||^2$ for every $x \in S_0^c(\sqrt{Bl})$. Combining eqs. (B.3), (B.4), and (B.5), and choosing $B \geq B_0$, we get

$$\Delta(Q^B) \leq D_l(R) + \frac{\epsilon}{3}, \tag{B.6}$$

for all $l \geq L_0(\epsilon/12)$. In words, Q^B is optimum for P within extra distortion of $\epsilon/3$. Thus, it will be sufficient to prove that the average distortion incurred by the empirically-designed quantizer is less than or equal to $\Delta(Q^B) + 2\epsilon/3$.

The average distortion of the empirically-designed quantizer $Q(\cdot|Z)$ can be decomposed as follows.

$$\begin{aligned}
\frac{1}{l}E\{E(||X-Q(X|Z)||^2|Z)\} &= \frac{1}{l}E\left\{\int_{S_0(\sqrt{Bl})}||x-Q(x|Z)||^2P(dx)\right\}+ \\
&\quad \frac{1}{l}E\left\{\int_{S_0^c(\sqrt{Bl})}||x-Q(x|Z)||^2P(dx)\right\} \\
&\leq \frac{1}{l}E\left\{P\{S_0(\sqrt{Bl})\}E_B(||X-Q(X|Z)||^2|Z)\right\}+ \\
&\quad \frac{4}{l}E\left\{\int_{S_0^c(\sqrt{Bl})}||x||^2P(dx)\right\} \\
&\leq \frac{1}{l}E\{E_B(||X-Q(X|Z)||^2|Z)\}+\frac{\epsilon}{3},
\end{aligned} \tag{B.7}$$

where E_B denotes expectation w.r.t. P_B, and the last two steps follow from the same considerations as in eq. (B.5), since the codewords of the empirically-designed quantizer are also in $S_0(\sqrt{Bl})$. Thus, it remains to show that

$$\frac{1}{l}E\{E_B(||X-Q(X|Z)||^2|Z)\} \leq D_l(R)+\frac{\epsilon}{3}. \tag{B.8}$$

For a given m and a training set $Z=(Z_1,...,Z_m)$, let $Y=(Y_1,...,Y_m)$ denote the binary sequence where $Y_i = 1\{Z_i \in S_0(\sqrt{Bl})\}$. Let $y=(y_1,...,y_m)$ denote a specific realization of Y. Then,

$$\frac{1}{l}E\{E_B(||X-Q(X|Z)||^2|Z)\} = \sum_y \Pr\{Y=y\}\frac{1}{l}E\{E_B(||X-Q(X|Z)||^2|Y=y)\}. \tag{B.9}$$

Now, let $m(y)$ denote the number of ones in y. Given that $Y=y$, with $y_{i_1}=y_{i_2}=...=y_{i_{m(y)}}=1$, it is clear that the relevant training vectors $Z_{i_1},Z_{i_2},...,Z_{i_{m(y)}}$ are i.i.d. and each Z_{i_j} is governed by the bounded-support density P_B. Since the expectation over the ensemble of X is taken w.r.t. P_B as well, we are actually back in the situation of Section 2. However, for some y-sequences, the number of training vectors $m(y)$ for learning P_B is small.

Let us partition the set of binary y-sequences into two complementary subsets according to $m(y) \leq 2^{(R+\delta/2)l}$ or $m(y) > 2^{(R+\delta/2)l}$. As for y-sequences in the first subset, the distortion might be large but it is bounded by the maximum possible per-letter distortion within $S_0(\sqrt{Bl})$, which is $4B$. Furthermore, since Y is a Bernoulli process with $\Pr\{Y_i = 1\} = P\{S_0(\sqrt{Bl})\} \geq 1-\epsilon$, then the probability that $m(y) \leq 2^{(R+\delta/2)l}$, or equivalently, $m(y)/m \leq 2^{-\delta l/2}$, decays exponentially with m and hence double-exponentially with l. Thus, for large enough l, we can make the overall contribution of all y-sequences with small $m(y)$, less than $\epsilon/6$, and so, the proof will be complete if we bound the overall extra distortion of the complimentary subset by $\epsilon/6$.

But for every y with $m(y) > 2^{(R+\delta/2)l}$, we can invoke Theorem 1 and obtain

$$\frac{1}{l}E\{E_B(||X - Q(X|Z)||^2|Y = y)\} \leq D_l(R;B) + \frac{\epsilon}{12} \tag{B.10}$$

for all sufficiently large l, where $D_l(R;B)$ is the minimum achievable distortion w.r.t. P_B over all rate R, l-dimensional quantizers. However, if ϵ is sufficiently small and B and l are sufficiently large, then

$$\begin{aligned} D_l(R;B) &= \min_Q E_B||X - Q(X)||^2 \\ &= \frac{\min_Q \int_{S_0(\sqrt{Bl})} ||x - Q(x)||^2 P(dx)}{P\{S_0(\sqrt{Bl})\}} \\ &\leq \frac{\min_Q \int_{\mathbb{R}^l} ||x - Q(x)||^2 P(dx)}{P\{S_0(\sqrt{Bl})\}} \\ &= \frac{D_l(R)}{P\{S_0(\sqrt{Bl})\}} \\ &\leq D_l(R) + \frac{\epsilon}{12}, \end{aligned} \tag{B.11}$$

where the last step follows from the fact that $D_l(R)$ is obviously upper bounded by σ^2. This completes the proof of Theorem 4.

Appendix C

Most Vectors in the Sphere Satisfy Conditions (c1) and (c2).

It will be sufficient to show that even in the last step of the procedure ($m = K$), the relative volume of points in $S_0(\sqrt{Al})$ that satisfy simultaneously conditions (c1) and (c2), tends to unity as $l \to \infty$. In other words, a random selection of u under a uniform PDF within $S_0(\sqrt{Al})$ will be successful with high probability.

As for Condition (c1), let $H_i = \{u : \; T^i u \in S_0(\sqrt{Al}) - W_{K-1}\}$. Clearly, if $u \in \cup_{i=0}^{l-1} H_i$ then Condition (c1) is violated. But, by the union bound, and since all H_i have the same volume, the volume of $\cup_{i=0}^{l} H_i$ cannot exceed $l \cdot \mathrm{Vol}\{H_0\}$. The volume of H_0, in turn is upper bounded as follows.

$$\begin{aligned} \mathrm{Vol}\{H_0\} &= \mathrm{Vol}\{\bigcup_{j=1}^{K-1}\bigcup_{i=0}^{l-1} S_{T^i u_j}(6\sqrt{lD_0})\} \\ &\le Kl \cdot \mathrm{Vol}\{S_{u_1}(6\sqrt{lD_0})\} \\ &= l2^{Gl} \cdot \mathrm{Vol}\{S_{u_1}(6\sqrt{lD_0})\} \\ &\le l \exp_2\{l[G + \frac{1}{2}\log(72\pi e D_0)]\}. \qquad \text{(C.1)} \end{aligned}$$

The last expression, and hence also the volume of $\cup_{i=0}^{l} H_i$, is exponentially negligible compared to $\mathrm{Vol}\{S_0(\sqrt{Al})\}$ since $G < 0.5\log(A/36D_0)$.

As for Condition (c2), obviously, it is sufficient to require only that $||u - T^j u||^2 > 36lD_0$ for $j = 1, 2, ..., l-1$. Consider the ϵ-surface of $S_0(\sqrt{Al})$ defined as $F = S_0(\sqrt{Al}) - S_0(\sqrt{(A-\epsilon)l})$. Let $L_j = \{u : ||u - T^j u||^2 \le 36lD_0\}$. Since F occupies most of the volume of $S_0(\sqrt{Al})$ for large l, it is sufficient to show that the volume of $\cup_{j=1}^{l-1}(L_j \cap F)$ is very small compared to that of $S_0(\sqrt{Al})$. First, observe that $L_j \cap F \subset \{u : u^t T^j u \ge \phi l\} \cap F$, where $\phi = A - 18D_0 - \epsilon$ and u^t denotes the transposition of the column vector u. Therefore,

$$\begin{aligned} \mathrm{Vol}\{\bigcup_{j=1}^{l-1}(L_j \bigcap F)\} &\le \mathrm{Vol}\left\{\bigcup_{j=1}^{l-1}\{u : u^t T^j u \ge \phi l\} \bigcap F\right\} \\ &\le \sum_{j=1}^{l-1} \mathrm{Vol}\left\{\{u : u^t T^j u \ge \phi l\} \bigcap F\right\} \\ &\le (l-1) \cdot \max_{1 \le j \le l-1} \mathrm{Vol}\left\{\{u : u^t T^j u \ge \phi l\} \bigcap F\right\}. \qquad \text{(C.2)} \end{aligned}$$

We next derive an upper bound on the volume of $\{u : u^t T^j u \ge \phi l\} \bigcap F$. We shall prefer to represent the quadratic form $u^t T^j u$ as $u^t E_j u$, where $E_j = (T^j + T^{-j})/2$, because E_j is a symmetric matrix. Now, for every sufficiently small $s > 0$, we have

$$\mathrm{Vol}\left\{\{u : u^t T^j u \ge \phi l\} \cap F\right\} = \int_F du \cdot 1\{u^t E_j u \ge \phi l\}$$

$$
\begin{aligned}
&\leq \int_F du \cdot e^{s(u^t E_j u - \phi l)} \\
&\leq e^{l/2} \int_F du \cdot e^{-u^t u/(2A)} e^{s(u^t E_j u - \phi l)} \\
&\leq e^{l/2} e^{-s\phi l} \int_{\mathbb{R}^l} du \cdot e^{-\frac{1}{2A} u^t (I - 2sAE_j) u} \\
&= (2\pi e A)^{l/2} e^{-s\phi l} |I - 2sAE_j|^{-1/2} \\
&= \exp[l(\frac{1}{2} \ln(2\pi e A) - s\phi \\
&\quad -\frac{1}{2l} \sum_{i=0}^{l-1} \ln \lambda_i (I - 2sAE_j))],
\end{aligned}
\tag{C.3}
$$

where $\lambda_i(I - 2sAE_j)$ denotes the ith eigenvalue of $I - 2sAE_j$. Since $I - 2sAE_j$ is a circulant matrix, the eigenvalues are given by the discrete Fourier transform coefficients of the first row of this matrix, that is,

$$
\lambda_i(I - 2sAE_j) = 1 - 2sA \cos\left(\frac{2\pi ij}{l}\right), \qquad j = 1, 2, ..., l-1. \tag{C.4}
$$

Now the first term in the exponent of the right-most side of eq. (C.3) represents the total volume of F (or $S_0(\sqrt{Al})$). Therefore, the proof will be complete if we show that for some $s > 0$, the expression

$$
J_l^j(s) = s\phi + \frac{1}{2l} \sum_{i=0}^{l-1} \ln \left[1 - 2sA \cos\left(\frac{2\pi ij}{l}\right)\right] \tag{C.5}
$$

is uniformly bounded away from zero for all l and $1 \leq j \leq l-1$. To this end, we next lower bound the second term of $J_l^j(s)$ for $s < 1/(2A)$ and $1 \leq j \leq l-1$.

$$
\begin{aligned}
\frac{1}{2l} \sum_{i=0}^{l-1} \ln \left[1 - 2sA \cos\left(\frac{2\pi ij}{l}\right)\right] &= -\frac{1}{2l} \sum_{i=0}^{l-1} \ln \left\{1 + \sum_{m=1}^{\infty} \left[2sA \cos\left(\frac{2\pi ij}{l}\right)\right]^m\right\} \\
&\geq -\frac{1}{2l} \sum_{i=0}^{l-1} \sum_{m=1}^{\infty} \left[2sA \cos\left(\frac{2\pi ij}{l}\right)\right]^m \\
&= -\frac{1}{2l} \sum_{i=0}^{l-1} \sum_{m=2}^{\infty} \left[2sA \cos\left(\frac{2\pi ij}{l}\right)\right]^m \\
&\geq -\frac{1}{2l} \sum_{i=0}^{l-1} \sum_{m=2}^{\infty} (2sA)^m \\
&= -\frac{2s^2A^2}{1 - 2sA},
\end{aligned}
\tag{C.6}
$$

where the first inequality follows from the fact that $\ln(1+x) \leq x$, and the second inequality follows from the fact that $|\cos\theta| \leq 1$. Thus,

$$
J_l^r(s) \geq s\phi - \frac{2s^2A^2}{1 - 2sA} \triangleq J(s). \tag{C.7}
$$

Since the positive term is linear in s while the negative term is quadratic, it is easy to find a small s for which $J(s)$ is strictly positive. Since we have shown that for every $1 \leq j \leq l-1$, the volume of $L_j \cap F$ is less than $\exp\{l[0.5\ln(2\pi eA) - J(s)]\}$, then the volume of the union of these sets cannot exceed $(l-1)\exp\{l[0.5\ln(2\pi eA) - J(s)]\}$, which is still negligible compared to the volume of the sphere of radius $\sqrt{Al}$.

Finally, since the fraction of points violating Condition (c1) is negligible and the fraction of points violating Condition (c2) is negligible, so is the fraction of points in their union.

Acknowledgement

We would like to thank Tamás Linder and the reviewers for their very useful comments.

References

[1] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, New York, Academic Press, 1981.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.

[3] Y. Hershkovits and J. Ziv, "On fixed-database universal data compression with limited memory," submitted for publication.

[4] 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. IT-40, no. 6, pp. 1728-1740, November 1994.

[5] D. Pollard, "Quantization and the method of k-means," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 199-205, March 1982.

[6] A. D. Wyner and J. Ziv, "Classification with finite memory," *IEEE Trans. Inform. Theory*, Vol. IT-42, no. 2, pp. 337-347, March 1996.