On the Performance of Vector Quantizers Empirically Designed From Dependent Sources

Assaf J. Zeevi*
Information Systems Lab
Stanford University
Stanford, CA. 94305-9510

Abstract

Suppose we are given n real valued samples $Z_1, Z_2, \ldots, Z_n$ from a stationary source P. We consider the following question. For a compression scheme that uses blocks of length k, what is the minimal distortion (for encoding the true source P) induced by a vector quantizer of fixed rate R, designed from the training sequence. For a certain class of dependent sources, we derive conditions ensuring that the empirically designed quantizer performs as well (on the average) as the optimal quantizer, for almost every training sequence emitted by the source. In particular, we observe that for a code rate R, the optimal way to choose the dimension of the quantizer is $k_n = \lfloor (1-\delta) R^{-1} \log n \rfloor$. The problem of empirical design of vector quantizer of fixed dimension k based on a vector valued training sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is also considered. For a class of dependent sources, it is shown that the mean squared error (MSE) of the empirically designed quantizer w.r.t the true source distribution converges to the minimum possible MSE at a rate of $O(\sqrt{\log n / n})$, for almost every training sequence emitted by the source. In addition, the expected value of the distortion redundancy – the difference between the MSE's of the quantizers – converges to zero for a sequence of increasing block lengths k, if we have at our disposal corresponding training sequences whose length grows as $n = 2^{(R+\delta)k}$. Some of the derivations extend recent results in empirical quantizer design using an i.i.d. training sequence, obtained by Linder *et al.* [7] and Merhav and Ziv [8]. Proof techniques rely on recent results in the theory of empirical processes, indexed by VC function classes.

*This work was supported in part by Grants JSEP #DAAH04-94-G-0058, NSF #NCR-9628193 and ARPA #J-FBI-94-218-2.

1 Introduction

One of the central problems in lossy data compression concerns designing vector quantizers from training data. The following question is of interest. Suppose we are given a real valued training sequence, drawn according to the source probability measure, from which we are to design a quantization scheme. The empirical design of a k-dimensional vector quantizer may be viewed as a process of *learning* from examples, and the effectiveness of this learning phase may be measured by the performance of the vector quantizer on future data. The discrepancy between the mean squared error when using the empirical quantizer to compress the source P, and that of the optimal quantizer of dimension k and rate R (aka the *operational distortion–rate function* of order k) is of interest. In what follows we will refer to this as the *distortion redundancy*.

The main question here is the following. Can one choose the dimension of the quantizer, based on the the training set, so as to ensure an arbitrarily small distortion redundancy (in a manner that will be made precise). We note in passing that the empirical design via minimizing the empirical distortion is typically implemented using the so-called Lloyd algorithm, or LBG [6].

Pollard [11] proved that for all stationary memoryless sources, with vector valued r.v.'s in $\mathbb{R}^k$ having a finite second order moment, the empirical mean squared error (MSE) will asymptotically converge to the optimal one. These asymptotics were investigated in [3], by experimental verification. A slightly different question was addressed both in [7] and [8]. For a fixed dimension k and rate R, what is the minimal amount of side information needed to design a vector quantizer so that the distortion redundancy is made arbitrarily small. It turns out that the number of training vectors should be $n = 2^{kR}$ in an exponential sense. The 'direct' part was given by [7, Corollary 1], and subsequently strengthened by Merhav and Ziv who established the 'converse' result in [8, Theorem 3], and also state the 'direct' part in a more general setting. Another related result was obtained by Bartlett *et al.* [1], using a minimax framework to derive a lower bound on the distortion redundancy, for fixed dimension vector quantizers. Linder *et al.* [7] also proved that $k_n = (1-\delta)R^{-1}\log n$ is the optimal choice of dimension for empirically designing a quantizer, based on n scalar samples drawn from a memoryless source. Here and throughout $\log \equiv \log_2$. For this choice, they establish that the distortion redundancy is almost surely of order $n^{-\tau}$ with $\tau \in (0, \delta/4)$. The random variables are assumed to take values in a bounded set.

All of the above papers make the fairly stringent assumption of i.i.d. training samples, which in fact only holds for a class of memoryless sources. As pointed out in [8] "It is an open problem to prove the 'direct part' for dependent training vectors". We note that the independence assumption is usually not true in practice, and in fact it is much more common to encounter data exhibiting strong dependence or correlation.

In this paper we focus on the problem of empirical quantizer design for a class of

stationary dependent sources, satisfying certain mixing conditions. The term *source* relates to the underlying stochastic process, or the associated probability measure, depending on the context. Given a real valued training sequence $Z_1^n = (Z_1, Z_2, \ldots, Z_n)$, we derive an upper bound on the distortion redundancy for a fixed block size k. Subsequently, we argue that the block size must increase at a rate which is logarithmic with the length of the training sequence, to ensure that the distortion redundancy tends to zero for almost every training sequence. Alternatively, for an increasing sequence of block lengths k, the length of the training sequence should be $n = 2^{(R+\delta)k}$ so that the distortion redundancy converge to zero in expectation. Consequently, we are able to obtain finite sample bounds on the minimal length of the training sequence, for which the distortion redundancy is arbitrarily small with high probability. The rate of the quantizer for which the distortion redundancy tends to zero is restricted by the dependence structure in the process. The intuitive explanation is: If we take a large block size, this will make the dependence effects smaller, and incidently also improve the mean squared error, at the expense of decreasing the number of training vectors available for the design algorithm. Thus, the dimension of the quantizer, the size of the training sequence, and the dependence structure of the process all play a role in determining the optimal tradeoff. Our results quantify the effects of the dependent structure of the process, on the distortion redundancy, and consequently on the length of the required training sequence. Another result we obtain may be viewed as an extension of [7, Corollary 1], and can be used to extend the 'direct' theorem in [8, Theorem 2]. We show that for a class of vector valued dependent sources, the distortion redundancy converges to zero almost surely at a rate of $O(\sqrt{\log n/n})$. In addition, for a sequence of increasing dimensions the expected distortion redundancy can be made arbitrarily small for a training sequence that grows as $n = 2^{k(R+\delta)}$. The essential ingredients in the derivations are the results obtained in [7] for the i.i.d. case, and some recent results from the theory of empirical processes indexed by VC classes [13].

This paper is organized as follows. In section 2 we give some preliminary definitions. Section 3 presents the main results and Section 4 gives some concluding remarks. The proofs are omitted, and can be found in the full paper.

2 Definitions and Notation

Let $\mathbb{R}$ denote the real line, and let $\mathbb{R}^k$ denote the k-dimensional Euclidean space. A k-dimensional N-level *vector quantizer* is a measurable mapping $Q : \mathbb{R}^k \to \mathcal{C}_N$, where $\mathcal{C}_N = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ is a finite collection of vectors in $\mathbb{R}^k$, referred to as the *codebook* of Q. Thus, Q induces a partition of $\mathbb{R}^k$ with cells $S_i = \{\mathbf{x} \ : \ Q(\mathbf{x}) = \mathbf{y}_i\}$, and the classification of input vectors into codewords is done according to the *nearest neighbor* partition rule, i.e.,

$$Q(\mathbf{x}) = \mathbf{y}_i, \quad \text{if } \|\mathbf{x} - \mathbf{y}_i\| < \|\mathbf{x} - \mathbf{y}_j\| \ \ \forall j \quad .$$

The norm $\|\cdot\|$ denotes the Euclidean norm, and ties are broken arbitrarily. The *rate* of this quantizer is $R = (\log N)/k$ bits per symbol. Here and throughout we use bold

face notation to distinguish vectors from scalars.

Let $\{Z_i\}_{i\geq 1}$ be a discrete time real valued stationary stochastic process, with probability measure P. Here and throughout we will denote r.v.'s using uppercase type and corresponding realizations using lowercase. In what follows we will be interested in the design of a fixed rate R lossy encoder, by means of quantizing source blocks of length k using a k-dimensional, $N = 2^{kR}$-level, vector quantizer Q. There are many reasons to focus on vector quantization as opposed to scalar (see [5] for a discussion and list of related references). It is sufficient to note that even for simple i.i.d. processes a substantial improvement in performance may be achieved using a vector quantizer.

We will focus on a class of stationary mixing processes, for which we are assured that events sufficiently 'far apart' are 'almost' independent. More precisely, let

$$\mathcal{F}_1^j = \sigma(Z_1, Z_2, \ldots, Z_j)$$

and

$$\mathcal{F}_{j+\ell}^{\infty} = \sigma(Z_{j+\ell}, Z_{j+\ell+1}, \ldots)$$

be the sigma-algebras generated by the respective r.v.'s. In the sequel, the following definition of mixing will be utilized

Definition 1 *For any sequence $\{Z_i\}_{i\geq 1}$, the β-mixing (or completely regular) coefficient $\beta_z(\ell)$ is defined as follows:*

$$\beta_z(\ell) = \sup_{j\geq 1} \mathbb{E}\left[\sup_{B\in\mathcal{F}_{j+\ell}^{\infty}} |P(B|\mathcal{F}_1^j) - P(B)|\right] \quad .$$

A process for which $\beta_z(\ell) \to 0$ for $\ell \to \infty$ is called *β-mixing*. Further properties and relation to other mixing conditions can be found in [2], and [4]. We note in passing that there exist other definitions of mixing, most of which are more restrictive than β-mixing (see the above references for more details).

Remark 1 We focus on this measure of dependence since the machinery underlying our main results is the uniform convergence of sample means to their expectations, over certain so called VC-classes. For this framework, the β-mixing condition is a natural choice (see the discussion in [13] following Lemma 4.1, and [10] for more details). In particular, in this work we follow closely the results in [13], which have been derived for the β-mixing class.

In order to obtain finite sample results in the uniform convergence framework, we must further restrict the class of sources. In what follows we focus on the following subclass of sources

$$\mathcal{P}(B, \tilde{\beta}, b) = \left\{P : \ P(|Z| \leq B) = 1, \ \beta(\ell) \leq \tilde{\beta}\exp\{-b\ell\} \ \forall \ell \geq 1\right\}$$

where $B < \infty$. Thus we are assuming that the scalar random variables emitted by the source are uniformly bounded, and that the 'memory' of the process vanishes exponentially fast (in the sense of the β-mixing definition), for events which are sufficiently far apart. For examples of such processes the reader is referred to [13] and the list of references therein. In particular, classes of Harris recurrent stationary Markov processes are exponentially β-mixing under appropriate drift conditions. Examples include stationary ARMA processes with innovations that have distribution absolutely continuous w.r.t. Lebesgue measure (see [9] for further details).

Remark 2 The boundedness assumption is standard in deriving exponential inequalities (c.f. [7] and [8] in this context). It is mainly put forth to avoid technical conditions. In particular, growth rate of moments, existence of the moment generating function in a certain region, and similar assumptions may be substituted in for boundedness of the random variables. However, to obtain distribution free uniform convergence rates, the boundedness assumption cannot be dispensed with so easily. In particular, the proof technique used herein will not hold in the absence of this assumption. The assumption that the memory of the process decays exponentially fast is also essential. A polynomial rate will not suffice in order to obtain bounds and rates of convergence. Thus, $\mathcal{P}$ is in fact the most general class of processes given the present proof techniques.

Recall Z_1^n denotes a training sequence emitted by a source P. We divide the n-sequence into blocks of length k, one after the other. Let $m_n = \lfloor n/2k \rfloor$, so that there are $2m_n$ blocks of length k, and a remainder block of length $n - 2m_n k$. Let us assume, with no loss of generality, that $n/2k$ is an integer, and thus the remainder block is empty. The proof of Theorem 1 shows that the remainder block is uniformly bounded and therefore does not affect the analysis. Denote the vector valued block sequence by $(\mathbf{X}_i)_{1 \leq i \leq 2m_n}$. The choice of m is made to adhere with the proofs of the main theorem, where the above definition is convenient.

We now return to the problem of vector quantizer design. Given a quantizer $Q : \mathbb{R}^k \to \mathbb{R}^k$, define its mean square *average distortion*

$$\Delta(Q) = \mathbb{E}\|\mathbf{X} - Q(\mathbf{X})\|^2$$

where the expectation is taken w.r.t. the probability measure that P induces on $\mathcal{B}([-B, B]^k)$, and the *empirical distortion* as

$$\Delta_n(Q) = \frac{1}{2m_n} \sum_{i=1}^{2m_n} \|\mathbf{X}_i - Q(\mathbf{X}_i)\|^2 \quad .$$

Let Q^* denote the quantizer with minimal average distortion, and let Q_n^* be the quantizer with minimal empirical distortion. That is,

$$Q^* = \arg \min_{Q \in \mathcal{Q}(R;k)} \Delta(Q)$$

and

$$Q_n^* = \arg \min_{Q_n \in \mathcal{Q}(R;k)} \Delta_n(Q)$$

where $\mathcal{Q}(R;k)$ stands for the class of k-dimensional, $N = 2^{kR}$-level quantizers, for some rate R fixed and given. Denote

$$D_k(R) = \frac{1}{k}\Delta(Q_*^n)$$

the k'th order operational distortion–rate function. Since the encoding is done according to the nearest neighbor rule, finding the optimal quantizer amounts to finding the set of codewords $\mathcal{C}_N$.

In the sequel we will be interested in the *distortion redundancy*, which we define as follows

$$\mathcal{E}(Q_n^*, Q^*) = \underbrace{\frac{1}{k}\Delta(Q_n^*)}_{\text{(i)}} - \underbrace{D_k(R)}_{\text{(ii)}} \tag{1}$$

where

$$\Delta(Q_n^*) = \mathbb{E}\left[\|\mathbf{X} - Q_n^*(\mathbf{X})\|^2 \mid \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{2m_n}\right] \quad .$$

The difference between the two mean squared errors may be interpreted as follows.

- The first term, (i) in eq. (1), may be understood as the error induced by the empirically designed vector quantizer, when applied to a 'very large' future data set. That is, once we have designed the quantizer using the training sequence, it is held fixed and put to use on future data. Since the quantizer is being applied to a 'very large' data set, we may take the performance measure to be the average distortion rather than the empirical distortion.
- The second term, (ii) in eq. (1), is the average distortion, induced by the *optimal* vector quantizer, i.e., one that is 'custom' designed using the source's statistics. This distortion is achievable only if we have perfect knowledge of the source k marginal.

In essence, the difference between the two MSE's represents the loss incurred by using the empirically optimal quantizer, instead of the optimal one.

3 Main Results

Our objective is to determine the optimal choice of quantizer dimension given the training sequence, and derive a nonasymptotic lower bound on the length of the training sequence which will assure a distortion redundancy which is arbitrarily small. Let the rate of the quantizer R be fixed and given. The first theorem gives a large deviation bound on the generalization error.

Theorem 1 *Suppose that a rate R, k dimensional quantizer Q_n^* is designed to minimize $\Delta_n(Q)$, the empirical distortion, over a scalar valued training sequence Z_1^n emitted by a source $P \in \mathcal{P}(B, \tilde{\beta}, b)$ (and blocked into k blocks). Then for any $\epsilon > 0$, and*

$n \geq \max\{c_1/\epsilon^2, k/\epsilon\}$ *the distortion redundancy is bounded as follows*

$$P\left\{z_1^n : \mathcal{E}(Q_n^*, Q^*) > \epsilon\right\} \leq 8(n/k)^{N(k+1)} e^{-n\epsilon^2/ck} + \tilde{\beta}(n/k)e^{-bk}$$

for c, and c_1 absolute constants, whose values can be explicitly determined.

Note that the bound is comprised of two terms. The first is the classical exponential bound obtained using the Vapnik-Chervonenkis inequality [12]. This result was derived in [7, Theorem 1] for memoryless sources. The second is a bound on the error induced by the memory structure of the process. Note also that the dependence on the block size k is explicit in the upper bound. In fact, the block size plays a crucial role in balancing the two error terms. This relation is pursued in Corollary 2. Clearly, a quantizer of large dimension can give rise to smaller distortion, at the expense of having less training data available. Thus, if k is increased, the quantizer tends to underfit the source. However, this increase allows us to control the error term that results from the dependent structure of the blocks.

By controlling the growth of block size (i.e., the quantizer dimension k), we may obtain consistency results. Moreover, using the bound derived in Theorem 1, we can determine the rate at which the distortion redundancy converges to zero. This result is stated in the following corollary.

Corollary 1 *Fix $\delta \in (0,1)$ and take the block size $k_n = \lfloor \frac{1-\delta}{R} \log n \rfloor$. Then, for any source $P \in \mathcal{P}$ code rate $R \in (0, R_{\max})$, and $\tau \in (0, \frac{\delta}{2})$ we have*

$$\frac{1}{k_n}\Delta(Q_n^*) - D_{k_n}(R) = O(n^{-\tau}) \quad a.s. - P$$

where $R_{\max} \stackrel{\triangle}{=} (1-\delta) b \log e/2$.

An immediate consequence is that the convergence holds also in mean by the bounded convergence theorem (since $P \in \mathcal{P}$). From the bounds obtain in Theorem 1 it is clear that if the β-mixing assumption is weakened to a polynomial mixing rate, there is no choice of k such that the overall error converges in probability. Note also that the rate is restricted by the class of sources $\mathcal{P}$ in such a way, that $b \uparrow \infty$ implies $R_{\max} \uparrow \infty$. In the limit of b sufficiently large the process is 'essentially' i.i.d., and the rate constraint vanishes, as we would expect. The choice of the block size as well as the restriction on rate region are both a result of the proof technique. The block size choice is in fact optimal in this setting, since for $k_n = [(1+\delta)/R] \log n$ the distortion redundancy does not converge to zero.

Another result which follows from the exponential bound in Theorem 1 is a lower bound on the size of the training sequence, needed to ensure that the event $\{\mathcal{E}_n < \epsilon\}$ occur with probability $1-\alpha$, for any arbitrarily small ϵ and α.

Corollary 2 *Let $P \in \mathcal{P}$. Fix $\delta \in (0,1)$, $R \in (0, 2R_{\max})$, and block size $k_n = \lfloor \frac{(1-\delta)}{R} \log n \rfloor$. Then, for any $\epsilon, \alpha > 0$*

$$\mathbb{P}\left\{z_1^n : \frac{1}{k}\Delta(Q_n^*) \leq D_k(R) + \epsilon\right\} \geq 1 - \alpha$$

hold for all n such that

$$\frac{\log n}{n^{\delta/3}} \leq K(\mathcal{P}, R, \delta, \epsilon, \alpha)$$

with K and explicit function of the given parameters.

Thus, for a given confidence level α, and confidence interval of width ϵ around $D_k(R)$, one can use Corollary 2 to read off the sufficient size of the training sequence needed to ensure that the empirical quantizer will have performance in the required confidence region.

The following result is derived for vector valued processes. Following [8], let us define a class of probability measures

$$\mathcal{P}'(B, \tilde{\beta}, b) = \left\{ P : \ P(\|\mathbf{X}\| \leq \sqrt{k}B) = 1, \ \beta(\ell) \leq \tilde{\beta} \exp\{-b\ell\} \ \forall \ell \geq 1 \right\}$$

where $B < \infty$, and $\mathbf{X} : \Omega \to \mathbb{R}^k$ is random variable emitted by the source P. Here $\|\cdot\|$ denotes the usual Euclidean norm.

Then, we have

Theorem 2 *Suppose that a rate $R = \frac{\log N}{k}$, k dimensional quantizer Q_n^* is designed to minimize $\Delta_n(Q)$, the empirical distortion, over a vector valued training sequence $\mathbf{X}_1^n$ emitted by a source $P \in \mathcal{P}'(B, \tilde{\beta}, b)$. Then, we have*

$$\frac{1}{k}\Delta(Q_n^*) - D_k(R) = O\left(\sqrt{\frac{\log n}{n}}\right) \quad a.s. - P$$

The theorem follows from [7, Theorem 1], and Lemma 4.2 in [13]. Note that in this setting we have no rate constraints on the quantizer. An immediate corollary is the convergence in mean, following again from the bounded convergence theorem. Consider now a setup in which the dimension of the vectors k increases. Merhav and Ziv [8] asked the following question: what is the minimal amount of side information bits needed to design a quantizer, such that the expected distortion redundancy is arbitrarily small. Now, by application of [13, Lemma 4.1] to Theorem 1 of [8] we obtain the following. Fix $\delta > 0$ and $n = 2^{k(R+\delta)}$. Then, the expected distortion redundancy vanishes for $k \to \infty$. By proper quantization of the training sequence vectors, as in [8], the amount of side information bits can be shown to be $2^{k(R+\delta)+o(k)}$, where $o(k)/k \to 0$. This immediately extends Theorem 2 of [8] to non i.i.d. training sequence, and implicitly their Theorem 4 which deals with the same questions in the case of stationary processes. We may now assume that the empirical quantizers are trained using a dependent training sequence, rather then make the stringent assumption of an i.i.d. sequence.

4 Conclusions

The results in this paper extend some existing ones which were derived for an i.i.d. training sequence. Additional results may be obtained by the same techniques. In

particular, a counterpart to Theorem 2 can be derived for convergence in expectation, and this in turn can be used to generalize results in [7] and [8, Theorem 2].

Several interesting questions arise from the analysis and results. For the scalar case we have assumed that the source has marginals which are supported on a bounded set. This is a rather stringent assumption, though it is quite common in the framework we pursue in this paper (see also [7] and [1]). It is still an open problem to see whether one can obtain these results under moment assumptions or tail conditions instead, even in the case of memoryless sources.

It is somewhat disturbing that some of the results for block encoding the scalar valued source are derived under a rate constraint. This constraint, in turn, is determined by the exponential rate of decay of the mixing coefficient associated with the source. Note that for processes with 'short' memory we have $R_{\max}$ large, and as the process is closer to independent, the rate constraint becomes negligible. We note in passing that this restriction follows from the proof technique alone, and therefor should be investigated further.

Results Corollary 1 and 2, are obtained under the condition of growing block size. In the limit, it is known that under weak conditions (e.g., stationarity and ergodicity) the optimal MSE converges (in the block size) to the distortion rate function. That is, $D_k(R) \to D(R)$. An interesting question therefor is to determine at what rate does $k^{-1}\Delta(Q_n^*)$, the per symbol distortion for the empirically designed quantizer, converge to the distortion rate function for the class of weakly dependent sources studied in this paper. For memoryless sources, Linder *et al.* [7, Theorem 2] established that $D_k(R) - D(R) = O(\sqrt{\log k / k})$, and for correlated Gaussian sources Wyner [14] proved that the same convergence rate holds, if the spectral density is Lipschitz continuous. Whether one can establish this rate of convergence for the class of sources $\mathcal{P}$ is subject to further investigation.

Acknowledgment: The author would like to thank Richard Olshen for some discussions, and Robert Gray for his helpful comments on a preliminary version of this paper.

References

[1] Bartlett, P., Linder, T. and Lugosi, G. "A minimax lower bound for empirical quantizer design", unpublished manuscript, 1996.

[2] Bradley, R.C. "Basic properties of strong mixing conditions", in *Dependence in Probability and Statistics*, ed. E. Eberleint and M. Taqqu, Birkhäuser, 1986.

[3] Cosman, P.C., Perlmutter, K.O., Perlmuter, S.M., Olshen, R.A. and Gray, R.M. "Training sequence size and vector quantizer performance", in *Proc. Asilomar Conf. on Signals, Sys. and Comput.*, pp. 434-438, 1991.

[4] Doukhan, P. *Mixing - Properties and Examples*, Springer, 1989.

[5] Gersho, A., and Gray, R.M. *Vector Quantization*, Kluwar, 1991.

[6] Linde, Y., Buzo, A. and Gray, R.M. "An algorithm for vector quantizer design", *IEEE Trans. Commun.*, Vol. 28, pp. 84-95, 1990.

[7] Linder, T., Lugosi, G. and Zeger, K. "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding", *IEEE Trans. on Info. Theory*, Vol. 40, pp. 1728-1740, 1994.

[8] Merhav, N. and Ziv, J. " On the Amount of Side Information for Lossy Data Compression", *IEEE Trans. on Info. Theory*, vol. 43, pp. 1112-1123, 1997.

[9] Mokkadem, A. "Mixing properties of ARMA processes", *Stoch. Proc. and Aplic.*, vol. 29, pp. 309-315, 1988.

[10] Nobel, A. and Dembo, A. "A note on uniform laws of averages for dependent processes", *Statist. and Prob. Lett.*, Vol. 17, pp. 169-172, 1993.

[11] Pollard, D. "Quantization and the method of k-means", *IEEE Trans. on Info. Theory*, Vol. 28, pp. 199-205, 1982.

[12] Vapnik, V.N. and Chervonenkis, A.Y. "On the uniform convergence of relative frequencies of events to their probabilities", *Theory of Prob. and Applic.*, Vol. 16, pp. 264-280, 1971.

[13] Yu, B., "Rates of convergence for empirical processes of stationary mixing sequences", *Ann. of Prob.*, Vol. 22, pp. 94-116, 1994.

[14] Wyner, A.D. "On the transmission of correlated Gaussian over a noisy channel with finite encoding block length", *Infom. Contr.*, vol. 20, pp. 193-215, 1972.