Empirical Quantizer Design in the Presence of Source Noise or Channel Noise

Tamás Linder, *Member, IEEE*, Gábor Lugosi, and Kenneth Zeger, *Senior Member, IEEE*

***Abstract*— The problem of vector quantizer empirical design for noisy channels or for noisy sources is studied. It is shown that the average squared distortion of a vector quantizer designed optimally from observing clean independent and identically distributed (i.i.d.) training vectors converges in expectation, as the training set size grows, to the minimum possible mean-squared error obtainable for quantizing the clean source and transmitting across a discrete memoryless noisy channel. Similarly, it is shown that if the source is corrupted by additive noise, then the average squared distortion of a vector quantizer designed optimally from observing i.i.d. noisy training vectors converges in expectation, as the training set size grows, to the minimum possible mean-squared error obtainable for quantizing the noisy source and transmitting across a noiseless channel. Rates of convergence are also provided.**

***Index Terms*—Empirical vector quantizer design, lossy source coding, training sets, convergence rates, channel noise.**

I. Introduction

The design of quantizers has been studied over the last four decades from various perspectives. On the practical side, the Lloyd–Max [1], [2] algorithm provides an efficient iterative method of designing locally optimal quantizers from known source statistics or from training samples. The generalized Lloyd algorithm [3], [4] similarly is useful for designing vector quantizers. A theoretical problem motivated by practice is the question of consistency: if the observed training set size is large enough, can one expect a performance nearly as good as in the case of known source statistics? The consistency of design based on global minimization of the empirical distortion was established with various levels of generality by Pollard [5], Abaya and Wise [6], and Sabin [7]. The finite sample performance was also analyzed by Pollard [8], Linder, Lugosi, and Zeger [9], and Chou [10]. The consistency of the generalized Lloyd algorithm was also established by Sabin [7] and Sabin and Gray [11]. An interesting interpretation of the quantizer design problem was given by Merhav and Ziv [12], who obtained lower bounds on the amount of side information a quantizer design algorithm needs to perform nearly optimally for all sources.

Less is known about the more general situation when the quantized source is to be transmitted through a noisy channel (joint source and channel coding), or when the source is corrupted by noise prior to quantization (quantization of a noisy source). In the noisy channel case, theoretical research has mostly concentrated on the questions of optimal rate-distortion performance in the limit of large block length either for separate [13], or joint [14] source and channel coding, as well as for high-resolution source-channel coding [15], [16]. Practical algorithms have also been proposed to iteratively design (locally) optimal source and channel coding schemes [17], [18].

For the noisy source quantization problem, the optimal rate-distortion performance was analyzed by Dobrushin and Tsybakov [19] and Berger [20]. The structure of the optimal noisy source quantizer for squared distortion was studied by Fine [21], Sakrison [22], and Wolf and Ziv [23]. The framework of these works also included transmission through a noisy channel. Properties of optimal noisy source quantizers as well as a treatment of Gaussian sources corrupted by additive independent Gaussian noise were given by Ayanoglu [24]. A Lloyd–Max-type iterative design algorithm was given by Ephraim and Gray [25] for the design of vector quantizers for noisy sources. A design approach based on deterministic annealing was reported by Rao *et al.* [26]. No consistency results have yet been proved for empirical design of noisy channel or noisy source vector quantizers.

In empirical design of standard vector quantizers one can observe a finite number of independent samples of the source vector. The procedure chooses the quantizer which minimizes the average distortion over this data. One is interested in the expected distortion of the designed quantizer when it is used on a source which is independent of the training data. An empirical design procedure is called *consistent* if the expected distortion of the empirical quantizer approaches the distortion of the quantizer which is optimal for the source, as the size of the training data increases. If consistency is established, one can investigate the rate of convergence of the algorithm, i.e., how fast the expected distortion of the empirically optimal quantizer approaches the optimal distortion. Tight convergence rates have practical significance, since consistency alone gives no indication of the relationship between the resulting distortion and the size of the training data.

In this paper, we investigate the consistency of vector quantizers obtained by global empirical error minimization

Manuscript received November 10, 1995; revised June 19, 1996. This research was supported in part by the National Science Foundation, the Hungarian National Foundation for Scientific Research, and the Foundation for Hungarian Higher Education and Research. The material in this paper was presented in part at the Data Compression Conference, Salt Lake City, UT, March 1996.

T. Linder and G. Lugosi are with the Department of Mathematics and Computer Science, Technical University of Budapest, Budapest, Hungary.

K. Zeger was with the Department of Electrical and Computer Engineering, Coordinated Science Laboratory, University of Illinois, Urbana-Champaign, Urban IL 61801 USA. He is now with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093 USA.

Publisher Item Identifier S 0018-9448(97)00782-7.

for noisy channels and noisy sources. In both cases, the notion of empirical (sample) distortion is not as simple as in standard vector quantizer design. For noisy channels, the channel transition probabilities are assumed to be known, and the empirical distortion is defined as the expected value of the distortion between a source symbol and its random reproduction, where the expectation is taken with respect to the channel. For sources corrupted by noise, the density of the noise is assumed to be known and the estimation-quantization structure (see, e.g., [23]) of the optimal quantizer is used. Here the sample distortion has no unique counterpart. Although a modified distortion measure can be introduced [25] which converts the problem into a standard quantization problem, this modified measure cannot directly be used since it is a function of the unknown source statistics. The main difficulty lies in the fact that, in general, the encoding regions of a noisy source vector quantizer need not be either convex or connected. Thus the set of quantizers to be considered in the minimization procedure is more complex than in the clean source or noisy channel case.

In this paper, Section II gives the necessary definitions for noisy channel and noisy source quantization problems. In Section III, consistency of the empirical design for noisy channel quantization is established. In particular, Theorem 1 proves that the expected squared error distortion of the quantizer minimizing the appropriately defined empirical distortion over m training vectors is within $O(\sqrt{\log m/m})$ of the distortion of the quantizer which is optimal for the given source and channel. This is the same rate as that obtained in [9] for the standard vector quantizer problem. In Section IV, empirical design for sources corrupted by additive noise is considered. A method is presented which combines nonparametric estimation with empirical error minimization. Theorem 2 proves that if the conditional mean of the clean source given the noisy source can be consistently estimated, then the method is consistent. Based on this result, Corollary 1 establishes the consistency of empirical design for additive, independent noise. We conjecture that the noisy source design problem is likely to be more difficult than the noisy channel quantizer design problem, when only noisy source samples are available. In Theorem 3 it is shown that consistency and convergence rates can be obtained under much more general conditions on the noise, if training samples from the clean source are also available.

II. Preliminaries

A. Vector Quantizers for Noisy Channels

An *N-level noisy-channel vector quantizer* is defined via two mappings. The *encoder* Q_C maps $\mathbb{R}^k$ into the finite set $\{1, \cdots, N\}$, and the *decoder* Q_D maps $\{1, \cdots, N\}$ onto the set of codewords $\{y_1, y_2, \cdots, y_N\} \subset \mathbb{R}^k$ by the rule $Q_D(j) = y_j$, for $j = 1, \cdots, N$. The rate of the quantizer is $(1/k) \log N$ bits per source symbol. The quantizer takes an $\mathbb{R}^k$-valued random vector X as its input, and produces the index $I = Q_C(X)$. The index I is then transmitted through a noisy channel, and the decoder receives the index $J \in \{1, \cdots, N\}$, a random variable whose conditional distribution given I is

$$\boldsymbol{P}(J = j|I = i) = p(j|i), \qquad 1 \le i, j \le N$$

where the $p(j|i)$ are the channel transition probabilities. The channel is assumed to be discrete with N input and N output symbols, with known transition probabilities, and the channel is assumed to work independently of the source X. The output of the quantizer is

$$Y = Q_D(J) = y_J$$

and the joint distribution of (X, Y) is determined by the source distribution and the conditional distribution

$$\boldsymbol{P}(Y = y_j|X = x) = p(j|Q_C(x)).$$

We will use the notation $Y = Q(X)$ as for an ordinary vector quantizer, but now Q is not a deterministic mapping. The performance of Q will be measured by the mean-squared distortion $\boldsymbol{E}[\|Q(X) - X\|^2]$, where $\|x\|$ denotes the Euclidean norm of the vector x. The quantizer distortion can be written as

$$\begin{aligned}
&\boldsymbol{E}[\|Q(X) - X\|^2] \\
&\quad = \boldsymbol{E}[\boldsymbol{E}[\|Q(X) - X\|^2 \,|\, I, J]] \\
&\quad = \sum_{i=1}^{N} \boldsymbol{P}(I = i) \\
&\qquad \cdot \left(\sum_{j=1}^{N} \boldsymbol{E}[\|Q(X) - X\|^2 \big| I = i, J = j] p(j|i) \right) \\
&\quad = \sum_{i=1}^{N} \int_{R_i} \left(\sum_{j=1}^{N} \|y_j - x\|^2 p(j|i) \right) P_X(dx) \qquad (1)
\end{aligned}$$

where the encoding regions $R_i = \{x\colon Q_C(x) = i\}$, for $i = 1, \cdots, N$ completely determine the encoder Q_C. It is obvious from (1) that given the decoder Q_D, the encoder regions

$$\begin{aligned}
R_i = \Bigg\{ x\colon & \sum_{j=1}^{N} \|y_j - x\|^2 p(j|i) \\
& \le \sum_{j=1}^{N} \|y_j - x\|^2 p(j|l), \quad l = 1, \cdots, N \Bigg\}
\end{aligned}$$

determine an encoder (with ties broken arbitrarily) which minimizes the distortion over all encoders. The above encoding rule is sometimes called the *weighted nearest neighbor condition* (see, e.g., [14], [17], [27], [28]). Note that some of the R_i may be empty in an optimal noisy channel vector quantizer (in contrast to the noiseless channel case).

Assuming that $\boldsymbol{E}[\|X\|^2] < \infty$, there always exists an N-level quantizer minimizing the distortion over all N-level quantizers. This is easily seen by adapting an argument for deterministic quantizers by Pollard [5]. Let us denote the distortion of such an optimal quantizer Q_N^* by

$$D_N^* = \boldsymbol{E}[\|Q_N^*(X) - X\|^2] = \min_{Q_C, Q_D} \boldsymbol{E}[\|Q(X) - X\|^2]$$

where the minimum is taken over all (N-level) encoders and decoders operating on the fixed channel and source X. Thus D_N^* depends on N, the source statistics, and on the channel transition probabilities, which we will assume to be fixed and *known* throughout this paper.

B. Vector Quantizers for Noisy Sources

Assume that Y is the noisy version of the source X. Y can be viewed as the output of a channel whose input is X. The noisy source Y is to be quantized by an N-level quantizer Q such that the mean-squared distortion

$$\boldsymbol{E}[\|X - Q(Y)\|^2]$$

is as small as possible. In this problem, an N-level quantizer Q is characterized by its codevectors $\{y_1, \cdots, y_N\} \subset \mathbb{R}^k$ and the measurable sets $R_i = \{x \in \mathbb{R}^k\colon Q(x) = y_i\}$, $i = 1 \cdots, N$, called *encoding regions*. As was noted in several papers dealing with this problem (see, e.g., [19], [21]–[23]), the structure of the optimal N-level quantizer can be obtained via a useful decomposition. Let $M\colon \mathbb{R}^k \to \mathbb{R}^k$ denote a version of the conditional expectation $\boldsymbol{E}[X \mid Y = y]$. Then

$$\begin{aligned}\boldsymbol{E}[\|X - Q(Y)\|^2] &= \boldsymbol{E}[\|X - M(Y)\|^2] + \boldsymbol{E}[\|M(Y) - Q(Y)\|^2] \\ &\quad + 2\boldsymbol{E}[(X - M(Y))^t(M(Y) - Q(Y))] \\ &= \boldsymbol{E}[\|X - M(Y)\|^2] + \boldsymbol{E}[\|M(Y) - Q(Y)\|^2]\end{aligned} \tag{2}$$

where the cross term disappears after taking iterated expectations, first conditioned on Y. Thus to minimize $\boldsymbol{E}[\|X - Q(Y)\|^2]$, the quantizer has to minimize $\boldsymbol{E}[\|M(Y) - Q(Y)\|^2]$. If the codevectors $\{y_1, \cdots, y_N\}$ are given, then the encoding regions minimizing the distortion must satisfy

$$\|M(y) - y_i\| \le \|M(y) - y_j\|, \quad \text{for } j = 1, \cdots, N \quad \text{if } y \in R_i. \tag{3}$$

This means that for any Q

$$\boldsymbol{E}[\|M(Y) - Q(Y)\|^2] \ge \boldsymbol{E}[\|M(Y) - \hat{Q}(M(Y))\|^2]$$

where $\hat{Q}$ is an ordinary *nearest neighbor* quantizer which has the same codevectors as Q. Thus by (2) we have

$$\begin{aligned}D_N^* &\stackrel{\text{def}}{=} \inf_Q \boldsymbol{E}[\|X - Q(Y)\|^2] \\ &= \boldsymbol{E}[\|X - M(Y)\|^2] + \inf_{\hat{Q}} \boldsymbol{E}[\|M(Y) - \hat{Q}(M(Y))\|^2]\end{aligned}$$

where the second infimum is taken over all N-level nearest neighbor quantizers $\hat{Q}$. Since $\boldsymbol{E}[\|M(Y)\|^2] \le \boldsymbol{E}[\|X\|^2]$, it follows from, e.g., Pollard [5] that an optimal quantizer $\hat{Q}^*$ exists. Therefore, the quantizer Q^* minimizing $\boldsymbol{E}[\|X - Q(Y)\|^2]$ is obtained by first transforming Y by M and then quantizing $M(Y)$ by a nearest neighbor quantizer $\hat{Q}^*$, that is,

$$Q^*(Y) = \hat{Q}^*(M(Y)).$$

Furthermore

$$D_N^* = \boldsymbol{E}[\|X - M(Y)\|^2] + \boldsymbol{E}[\|M(Y) - \hat{Q}^*(M(Y))\|^2]. \tag{4}$$

III. Empirical Design for Noisy Channels

In most applications one does not know the actual source statistics, but instead can observe a sequence of independent and identically distributed (i.i.d.) copies $Z_m = (X_1, X_2, \cdots, X_m)$ of X. These m "training samples" induce the empirical distribution P_m which assigns probability to every measurable $G \subset \mathbb{R}^k$ according to the rule

$$P_m(G) = \frac{1}{m} \sum_{l=1}^{m} I_{\{X_l \in G\}}$$

where I is the indicator function of the event of its argument. When the source statistics are not known, one cannot directly search for an optimal quantizer Q^*. Instead, one generally attempts to minimize the empirical distortion, which is a functional of P_m rather than of the true source distribution. The empirical distortion $D_{N,m}$ is the expected value (expectation taken over the channel use) of the average distortion of the quantizer when Z_m is quantized

$$D_{N,m} = \sum_{i=1}^{N} \int_{R_i} \left(\sum_{j=1}^{N} \|y_j - x\|^2 p(j|i) \right) P_m(dx). \tag{5}$$

The empirical distortion can be rewritten in the simple form

$$D_{N,m} = \frac{1}{m} \sum_{l=1}^{m} d_Q(X_l)$$

where $d_Q\colon \mathbb{R}^k \to \mathbb{R}^+$ is a function which depends on the quantizer Q as

$$d_Q(x) = \sum_{i=1}^{N} I_{\{x \in R_i\}} \left(\sum_{j=1}^{N} \|y_j - x\|^2 p(j|i) \right). \tag{6}$$

Note that the empirical distortion is a random variable, a function of the training data Z_m. We remark here that by using the function d_Q, the expected distortion of Q in (1) can be rewritten as

$$\boldsymbol{E}[\|Q(X) - X\|^2] = \boldsymbol{E}[d_Q(X)].$$

Assume we design a quantizer based on the training data by *minimizing the empirical distortion* over all possible quantizers. This minimization can be carried out in principle, since given Z_m and the channel transition probabilities, we can calculate $D_{N,m}$ for any quantizer using weighted nearest neighbor encoding.

Let $Q_N^*(\cdot|Z_m)$ be the quantizer minimizing $D_{N,m}$,

$$Q_N^*(\cdot|Z_m) = \arg\min_Q \frac{1}{m} \sum_{l=1}^{m} d_Q(X_l)$$

and let

$$D_{N,m}^* = \boldsymbol{E}[\|Q_N^*(X|Z_m) - X\|^2]$$

where X is independent of Z_m. Then $D_{N,m}^*$ is the average distortion of the empirically optimal quantizer when it is used on data independent of the training set. A fundamental question is how close this distortion gets to the optimal D_N^*

as the size of the training data increases, and therefore as the source statistics are more and more revealed by the empirical distribution.

One goal in this paper is to investigate how fast the difference between the expected distortion of the empirically optimal quantizer and the optimal distortion

$$\boldsymbol{E}[\|Q_N^*(X|Z_m) - X\|^2] - D_N^*$$

decreases as the training set size m increases. An upper bound on this difference, converging to zero as $m \to \infty$, is given which indicates how large the training set size should be so that the designed quantizer has a distortion near the optimum.

In what follows we assume that the source is bounded almost surely (a.s.), so that $\boldsymbol{P}(\|X\|^2 \le B) = 1$ for some $B > 0$. With this assumption we have the following theorem.

Theorem 1: Assume that a source $X \in \mathbb{R}^k$ is bounded as $\boldsymbol{P}(\|X\|^2 \le B) = 1$ for some $B > 0$, and let $Z_m = (X_1, \cdots, X_m)$, where the X_i are i.i.d. copies of X. Suppose an N-level noisy channel vector quantizer $Q_N^*(\cdot|Z_m)$ is designed by using empirical distortion minimization over the training set Z_m. Then the average distortion of this quantizer is bounded above as

$$\boldsymbol{E}[\|Q_N^*(X|Z_m) - X\|^2] \le D_N^* + c\sqrt{\frac{\log m}{m}} + O(m^{-1/2})$$

where D_N^* is the distortion of the N-level quantizer that is optimal for the source and the channel, and $c = 8B\sqrt{kN+1}$.

Proof: The proof of the theorem is based on a technique often used in the statistical and learning theory literature (see, e.g., [29]). First we note that the condition $\|X\|^2 \le B$ a.s. implies that both Q_N^* (the globally optimal quantizer) and $Q_N^*(\cdot|Z_m)$ (the empirically optimal quantizer) must have codevectors lying inside the sphere of radius $\sqrt{B}$ centered at the origin, since projecting any codevector outside this sphere back to the surface of the sphere clearly reduces the distortion. Let Q be a quantizer for the noisy channel and introduce the notation

$$\Delta_m(Q) = \frac{1}{m}\sum_{l=1}^{m} d_Q(X_l)$$

where d_Q is defined in (6). Let $\mathcal{F}$ be the class of all functions d_Q, where Q ranges through all N-level noisy channel quantizers Q whose codepoints lie inside the sphere $S(\sqrt{B}) = \{x \in \mathbb{R}^k\colon \|x\| \le \sqrt{B}\}$. These quantizers can be assumed to use the weighted nearest neighbor encoding rule since both $Q_N^*(\cdot)$ and $Q_N^*(\cdot|Z_m)$ use such encoders. For a fixed arbitrary $\epsilon > 0$, let $\mathcal{F}_\epsilon$ be an ϵ-covering of $\mathcal{F}$, i.e., let $\mathcal{F}_\epsilon$ be a set of functions $d_{\hat{Q}}$ such that for each $d_Q \in \mathcal{F}$, there exists an N-level noisy channel quantizer $\hat{Q}$ with $d_{\hat{Q}} \in \mathcal{F}_\epsilon$ satisfying

$$\sup_{\|x\| \le \sqrt{B}} |d_Q(x) - d_{\hat{Q}}(x)| \le \epsilon.$$

Let Q_N^* be an arbitrary fixed optimal quantizer (i.e., Q_N^* has N codevectors and distortion D_N^*), and let Q_N^{**} denote a quantizer such that $d_{Q_N^{**}} \in \mathcal{F}_\epsilon$ satisfies

$$\sup_{\|x\| \le \sqrt{B}} |d_{Q_N^*}(x) - d_{Q_N^{**}}(x)| \le \epsilon. \tag{7}$$

Then

$$\begin{aligned}
&\boldsymbol{E}[\|Q_N^*(X|Z_m) - X\|^2 \,|\, Z_m] - \boldsymbol{E}[\|Q_N^*(X) - X\|^2] \\
&\quad = \boldsymbol{E}[\|Q_N^*(X|Z_m) - X\|^2 \,|\, Z_m] - \Delta_m(Q_N^*(\cdot|Z_m)) \\
&\qquad + \Delta_m(Q_N^*(\cdot|Z_m)) - \boldsymbol{E}[\|Q_N^*(X) - X\|^2] \\
&\quad \le 2\epsilon + \boldsymbol{E}[\|Q_N^{**}(X|Z_m) - X\|^2 \,|\, Z_m] - \Delta_m(Q_N^{**}(\cdot|Z_m)) \\
&\qquad + \Delta_m(Q_N^*) - \boldsymbol{E}[\|Q_N^*(X) - X\|^2] \qquad \text{a.s.}
\end{aligned}$$

where in the inequality we used (7) and the fact that $\Delta_m(Q)$ is minimized by the empirically optimal quantizer. Thus we have that

$$\begin{aligned}
&\boldsymbol{E}[\|Q_N^*(X|Z_m) - X\|^2 \,|\, Z_m] - \boldsymbol{E}[\|Q_N^*(X) - X\|^2] \\
&\quad \le 2\epsilon + \sup_{d_Q \in \mathcal{F}_\epsilon \cup \{d_{Q_N^*}\}} 2|\Delta_m(Q) - \boldsymbol{E}[\Delta_m(Q)]| \quad \text{a.s.}
\end{aligned} \tag{8}$$

The right-hand side of the above inequality is a random variable whose expectation gives an upper bound on $\boldsymbol{E}[\|Q_N^*(X|Z_m) - X\|^2] - D_N^*$. To upper-bound this expectation we will use Hoeffding's [30] probability inequality which says that if $\xi_1, \cdots, \xi_m$ are i.i.d. real-valued random variables such that $\boldsymbol{P}\{\xi_1 \in [a, a+A]\} = 1$ for some a and $A > 0$, then

$$\boldsymbol{P}\left\{\left|\frac{1}{m}\sum_{l=1}^{m} \xi_l - \boldsymbol{E}\xi_1\right| > \epsilon\right\} \le 2e^{-2m\epsilon^2/A^2}. \tag{9}$$

Bounding the Cardinality of a Minimal ϵ-Covering: In order to use the facts above, we derive an upper bound on the cardinality of a minimal ϵ-covering of the class $\mathcal{F} = \{d_Q\colon Q \in \mathcal{Q}_N\}$, where $\mathcal{Q}_N$ is the set of all N-level noisy channel vector quantizers with weighted nearest neighbor encoders and whose codevectors have norm at most $\sqrt{B}$. Since the X_i all lie in the sphere $S(\sqrt{B})$, the set of functions $\mathcal{F}$ has a constant envelope of $4B$. Let us assume now that we are given the quantizers $Q, Q' \in \mathcal{Q}_N$ having codevectors $\{y_1, \cdots, y_N\}$ and $\{y_1', \cdots, y_N'\}$, respectively, such that for some $\rho > 0$, we have $\|y_j - y_j'\| \le \rho$ for all $j = 1, \cdots, N$. For a given $x \in S(\sqrt{B})$, assume without loss of generality that $d_Q(x) \le d_{Q'}(x)$. Setting

$$i = \arg\min_{1 \le l \le N} \sum_{j=1}^{N} \|y_j - x\|^2 p(j|l)$$

and

$$i' = \arg\min_{1 \le l \le N} \sum_{j=1}^{N} \|y_j' - x\|^2 p(j|l)$$

we have by the weighted nearest neighbor property that

$$\begin{aligned}
|d_Q(x) - d_{Q'}(x)| &= \sum_{j=1}^{N} \|y_j' - x\|^2 p(j|i') - \sum_{j=1}^{N} \|y_j - x\|^2 p(j|i) \\
&\le \sum_{j=1}^{N} \|y_j' - x\|^2 p(j|i) - \sum_{j=1}^{N} \|y_j - x\|^2 p(j|i) \\
&\le \sum_{j=1}^{N} p(j|i) 4\sqrt{B} \|y_j' - y_j\| \\
&\le 4\rho\sqrt{B}.
\end{aligned} \tag{10}$$

If we consider a rectangular grid of width δ in $S(\sqrt{B})$, then for any $y \in S(\sqrt{B})$ there is a point y' on this grid such that $\|y - y'\| \leq \delta\sqrt{k}$. Thus letting $\mathcal{Q}'_N$ be the set of all noisy channel quantizers which have all their codepoints on this grid and which use weighted nearest neighbor encoding, we obtain from (10) that for any $Q \in \mathcal{Q}_N$ there exists a $Q' \in \mathcal{Q}'_N$ such that

$$\sup_{x \in S(\sqrt{B})} |d_Q(x) - d_{Q'}(x)| \leq 4\delta\sqrt{kB}.$$

This implies that $\mathcal{F}_\epsilon = \{d_{Q'}\colon Q' \in \mathcal{Q}'_N\}$ is an ϵ-covering of $\mathcal{F}$ for $\epsilon = 4\delta\sqrt{kB}$. Letting V denote the volume of $S(\sqrt{B})$ we thus obtain

$$\begin{aligned} |\mathcal{F}_\epsilon| &\leq \left(\frac{V}{\delta^k}\right)^N \\ &= V^N (4\sqrt{Bk})^{kN} \epsilon^{-kN}. \end{aligned}$$

With this we obtain from (8), the union bound, and Hoeffding's inequality that for any t such that $0 < \epsilon < t/2$

$$\begin{aligned} &\boldsymbol{P}\{\boldsymbol{E}[\|Q^*_N(X|Z_m) - X\|^2 \,|\, Z_m] - \boldsymbol{E}[\|Q^*_N(X) - X\|^2] > t\} \\ &\leq \boldsymbol{P}\left\{ \sup_{Q \in \mathcal{Q}'_N \cup \{Q^*_N\}} \left| \frac{1}{m} \sum_{l=1}^{m} d_Q(X_l) - \boldsymbol{E}[d_Q(X)] \right| > \frac{t}{2} - \epsilon \right\} \\ &\leq (|\mathcal{F}_\epsilon| + 1) \\ &\cdot \sup_{Q \in \mathcal{Q}'_N \cup \{Q^*_N\}} \boldsymbol{P}\left\{ \left| \frac{1}{m} \sum_{l=1}^{m} d_Q(X_l) - \boldsymbol{E}[d_Q(X)] \right| > \frac{t}{2} - \epsilon \right\} \\ &\leq 2\Big(V^N (4\sqrt{Bk})^{kN} \epsilon^{-kN} + 1\Big) e^{-m(t/2-\epsilon)^2/8B^2}. \end{aligned} \tag{11}$$

This inequality holds for all $\epsilon < t/2$. Choose $\epsilon = t/4$. The difference inside the probability on the left-hand side is a.s. upper-bounded by $4B$. Using the simple bound $\boldsymbol{E}(Z) \leq t + 4B\boldsymbol{P}(Z > t)$, valid for any $t > 0$ and random variable Z such that $\boldsymbol{P}(Z > 4B) = 0$, we obtain

$$\begin{aligned} &\boldsymbol{E}[\|Q^*_N(X|Z_m) - X\|^2] - \boldsymbol{E}[\|Q^*_N(X) - X\|^2] \\ &\qquad \leq t + 8B\Big(V^N (16\sqrt{Bk})^{kN} t^{-kN} + 1\Big) e^{-mt^2/128B^2}. \end{aligned}$$

Finally, if we choose $t = c\sqrt{\log m/m}$ with constant $c = 8B\sqrt{(kN+1)}$, then the second term on the right-hand side of the above inequality is on the order of $m^{-1/2}$, and the proof of the theorem is complete. □

IV. Empirical Design for Noisy Source

In the noisy source quantizer design problem we are given the samples $Z_m = (Y_1, \cdots, Y_m)$ drawn independently from the distribution of Y. We also assume that the conditional distribution of the noisy source Y given X is known (i.e., the channel between X and Y is known), and that $\boldsymbol{P}(\|X\|^2 \leq B) = 1$ for some *known* constant B. In this situation the method of empirical distortion minimization cannot be applied directly, since we only have the indirect (noisy) observations $Y_1, \cdots, Y_n$ about X. However, the decomposition (4) suggests the following method for noisy source quantizer design:

i) Split the data Z_m into two parts, $Z_m^{(1)} = (Y_1, \cdots, Y_{m/2})$ and $Z_m^{(2)} = (Y_{m/2+1}, \cdots, Y_m)$ (assume m is even) and estimate $M(y) = \boldsymbol{E}[X|Y = y]$ from the first half of the samples $Z_m^{(1)}$ and the *known* conditional distribution $P_{Y|X}$. The estimate $M_m(\cdot) = M_m(\cdot, Z_m^{(1)})$ is required to be L_2 consistent:

$$\boldsymbol{E}[\|M_m(Y) - M(Y)\|^2] = a_m \to 0 \qquad \text{as} \quad m \to \infty. \tag{12}$$

Since the upper bound B on $\|X\|^2$ is known we also require that

$$\sup_{y \in \mathbb{R}^k} \|M_m(y)\|^2 \leq B. \tag{13}$$

ii) Using the second half of the training data define a new set of $m/2$ training vectors

$$M_m(Y_{m/2+1}), \cdots, M_m(Y_m)$$

and consider a nearest neighbor quantizer $\hat{Q}^*_m$ minimizing the empirical distortion

$$\hat{Q}^*_m = \arg\min_Q \frac{1}{m/2} \sum_{i=m/2+1}^{m} \|M_m(Y_i) - Q(M_m(Y_i))\|^2. \tag{14}$$

Here the minimization is over all N-level nearest neighbor quantizers. The quantizer for the noisy source designed from the noisy samples is then obtained from $\hat{Q}^*_m$ and M_m as

$$Q^*_m = \hat{Q}^*_m \circ M_m.$$

The following theorem gives an estimate for the difference between the distortion of Q^*_m and the minimum achievable distortion D^*_N.

Theorem 2: Assume that a source $X \in \mathbb{R}^k$ is bounded as $\boldsymbol{P}(\|X\|^2 \leq B) = 1$ for some $B > 0$ and let $(Y_1, \cdots, Y_m)$ be i.i.d. samples of the noisy source Y. Suppose, furthermore, that the conditional distribution of Y given X, and the constant B are known, and that the estimator $M_m(y)$ of $M(y) = \boldsymbol{E}[X \,|\, Y = y]$ has L_2 error

$$\boldsymbol{E}[\|M_m(Y) - M(Y)\|^2] = a_m$$

and is bounded as

$$\sup_{y \in \mathbb{R}^k} \|M_m(y)\|^2 \leq B.$$

Then the N-level Q^* quantizer designed in steps i) and ii) above satisfies

$$\begin{aligned} \boldsymbol{E}[\|X - Q^*_m(Y)\|^2] &\leq D^*_N + c\sqrt{\frac{\log m}{m}} + O(m^{-1/2}) \\ &\quad + 8\sqrt{Ba_m} + a_m \end{aligned}$$

where D^*_N is the distortion of the optimal N-level quantizer for the noisy source problem, and $c = 8B\sqrt{2(kN+1)}$.

Additive Independent Noise: Before proving the theorem we show how it can be applied to the special (but very important) case when $Y = X + \nu$, where the noise ν is independent of X. Theorem 2 implies that if there exists an L_2 consistent estimate of M, then the quantizer design procedure using this estimate will be consistent, i.e.,

$$\lim_{m\to\infty} \left(\boldsymbol{E}[\|X - Q_m^*(Y)\|^2] - D_N^*\right) = 0.$$

Such consistent estimators M_m exist, for example, when X has a density f, ν has a *bounded* density h, and the characteristic function of ν is nonzero almost everywhere. To see this, we use the following lemma (proved in the Appendix).

Lemma 1: Let $Y = X + \nu$ be a random vector in $\mathbb{R}^k$, where X and ν are independent absolutely continuous random variables. Assume that the density h of ν is known, and its characteristic function $\zeta(t) = \boldsymbol{E}[e^{it\nu}]$ is nonzero for almost all $t \in \mathbb{R}^k$. Assume that n i.i.d. copies $Y_1, \cdots, Y_n$ of Y are observed. Then for every density f of X there exists an estimator $f_n(x)$ of f such that

$$\lim_{n\to\infty} \int |f(x) - f_n(x)|\, dx = 0 \qquad \text{a.s.}$$

Take $n = m/2$ in Lemma 1. The estimator f_m in the lemma integrates to one, but it may take negative values. Also, even though f has a bounded support, f_m can have an unbounded support since it is obtained by deconvolving a kernel density estimate of the density g of Y. Let $A = \{x\colon\ f_m(x) \ge 0\}$. Then

$$\hat{f}_m(x) = \frac{f_m(x) I_{\{A\cap S(\sqrt{B})\}}}{\displaystyle\int_{A\cap S(\sqrt{B})} f_m(y)\, dy}$$

is a probability density with support contained in $S(\sqrt{B})$. Moreover, by [31, pp. 12–13] we have

$$\int |\hat{f}_m(x) - f(x)|\, dx \le \int |f_m(x) - f(x)|\, dx$$

so that $\hat{f}_m$ is also strongly L_1 consistent, and we can actually use the notation f_m instead of $\hat{f}_m$. Since

$$g(y) = \int h(y-x) f(x)\, dx$$

we have

$$M(y) = (g(y))^{-1} \left(\int x h(y-x) f(x)\, dx \right)$$

for all y such that $g(y) > 0$. We define our estimate M_m as

$$M_m(y) = \frac{\displaystyle\int x f_m(x) h(y-x)\, dx}{\displaystyle\int f_m(x) h(y-x)\, dx}.$$

It is immediate that $\|M_m(y)\|^2 \le B$ since

$$\frac{f_m(x) h(y-x)}{\displaystyle\int f_m(x') h(y-x')\, dx'}$$

is a probability density in x, with support contained in the convex set $S(\sqrt{B})$, and thus $M_m(y) \in S(\sqrt{B})$. We also have

$$\left| \int f_m(x) h(y-x)\, dx - \int f(x) h(y-x)\, dx \right| \le \|h\|_\infty \int |f_m(x) - f(x)|\, dx \to 0 \qquad \text{a.s.} \quad (15)$$

where $\|h\|_\infty = \sup_{x\in\mathbb{R}} h(x) < \infty$. Thus

$$\begin{aligned} \lim_{m\to\infty} g_m(y) &= \lim_{m\to\infty} \int f_m(x) h(y-x)\, dx \\ &= \int f(x) h(y-x)\, dx \\ &= g(y) \quad \text{a.s.} \end{aligned} \quad (16)$$

that is, g_m is pointwise strongly consistent. Letting $x = (x_1, \cdots, x_k)$ we have by (15) and (16) that for all $i = 1, \cdots, k$ and every y,

$$\lim_{m\to\infty} \frac{\displaystyle\int x_i f_m(x) h(y-x)\, dx}{\displaystyle\int f_m(x) h(y-x)\, dx} = \frac{\displaystyle\int x_i f(x) h(y-x)\, dx}{\displaystyle\int f(x) h(y-x)\, dx} \quad \text{a.s.}$$

since both f and f_m vanish outside $S(\sqrt{B})$. It follows that

$$\lim_{m\to\infty} \|M_m(y) - M(y)\|^2 = 0 \qquad \text{a.s.}$$

for almost every y. Fubini's theorem and the dominated convergence theorem then imply that

$$\lim_{m\to\infty} \boldsymbol{E}[\|M(Y) - M_m(Y)\|^2] = 0.$$

The consistency of the design procedure now follows from Theorem 2. Thus we have proved the following.

Corollary 1: Assume the conditions of Theorem 2 and suppose $Y = X + \nu$, where ν is independent of X and has a bounded density whose characteristic function is almost everywhere nonzero. Then there exists a bounded estimator M_m of M such that

$$\lim_{m\to\infty} \boldsymbol{E}[\|M(Y) - M_m(Y)\|^2] = 0$$

and the noisy source design procedure is consistent, i.e.,

$$\lim_{m\to\infty} \boldsymbol{E}[\|X - Q_m^*(Y)\|^2] = D_N^*.$$

Proof of Theorem 2: Using the same decomposition as in (2), the distortion of $Q_m^* = \hat{Q}_m^* \circ M_m$ can be written

$$\begin{aligned} \boldsymbol{E}[\|X - Q_m^*(Y)\|^2] =& \boldsymbol{E}[\|X - M(Y)\|^2] \\ &+ \boldsymbol{E}[\|M(Y) - \hat{Q}_m^*(M(Y))\|^2]. \end{aligned} \quad (17)$$

Then, by the Cauchy–Schwarz inequality, one obtains

$$\begin{aligned} &\boldsymbol{E}[\|M(Y) - \hat{Q}_m^*(M(Y))\|^2] \\ &\quad = \boldsymbol{E}[\|M_m(Y) - M(Y)\|^2] \\ &\qquad + \boldsymbol{E}[\|M_m(Y) - \hat{Q}_m^*(M(Y))\|^2] \\ &\qquad + 2\boldsymbol{E}[(M(Y) - M_m(Y))^t (M_m(Y) - \hat{Q}_m^*(M(Y)))] \\ &\quad \le \boldsymbol{E}[\|M_m(Y) - Q_m^*(M(Y))\|^2] + a_m + 4\sqrt{B a_m} \end{aligned} \quad (18)$$

where $a_m = \boldsymbol{E}[\|M_m(Y) - M(Y)\|^2]$. Recall now that M_m depends only on the samples $Z_m^{(1)} = (Y_1, \cdots, Y_{m/2})$ but is independent of $Z_m^{(2)} = (Y_{m/2+1}, \cdots, Y_m)$. With this in mind, we introduce an auxiliary N-level quantizer Q_m (used only in the analysis) which minimizes the conditional distortion

$$\boldsymbol{E}[\|M_m(Y) - Q_m(M_m(Y))\|^2 \mid Z_m^{(1)}].$$

Note that $Q_m(\cdot) = Q_m(\cdot, Z_m^{(1)})$ depends on $Z_m^{(1)}$. By definition, $\hat{Q}_m^*$ is an N-level quantizer minimizing the empirical distortion over the samples $M_m(Y_{m/2+1}), \cdots, M_m(Y_m)$ for a given $Z_m^{(1)}$. This fact and the independence of Y, $Z_m^{(1)}$, and $Z_m^{(2)}$ imply that for $Z_m^{(1)} = z$ the conditional probability

$$\begin{aligned}\boldsymbol{P}\{&\boldsymbol{E}[\|\hat{Q}_m^*(M_m(Y)) - M_m(Y)\|^2 \mid Z_m^{(2)}] \\ &- \boldsymbol{E}[\|Q_m(M_m(Y)) - M_m(Y)\|^2] > \epsilon \mid Z_m^{(1)} = z\}\end{aligned}$$

can be upper-bounded using the same technique as in the proof of Theorem 1. In fact, if the channel is made noiseless by substituting the transition probabilities $p(j|i) = \delta_{i,j}$ in Theorem 1, then the quantizers there become ordinary nearest neighbor quantizers. Since $\|M_m(Y)\|^2 \le B$, for a fixed z, the inequality (11) implies, after replacing m by $m/2$, that for a.e. z,

$$\begin{aligned}\boldsymbol{P}\{&\boldsymbol{E}[\|\hat{Q}_m^*(M_m(Y)) - M_m(Y)\|^2 \mid Z_m^{(2)}] \\ &- \boldsymbol{E}[\|Q_m(M_m(Y)) - M_m(Y)\|^2] > t \mid Z_m^{(1)} = z\} \\ &\le 2\Big(V^N(16\sqrt{Bk})^{kN} t^{-kN} + 1\Big) e^{-mt^2/256B^2}. \quad (19)\end{aligned}$$

Since the upper bound is independent of z, it follows that

$$\begin{aligned}\boldsymbol{P}\{&\boldsymbol{E}[\|\hat{Q}_m^*(M_m(Y)) - M_m(Y)\|^2 \mid Z_m^{(2)}] \\ &- \boldsymbol{E}[\|Q_m(M_m(Y)) - M_m(Y)\|^2] > t\} \\ &\le 2(V^N(16\sqrt{Bk})^{kN} t^{-kN} + 1) e^{-mt^2/256B^2}\end{aligned}$$

and one obtains in the same way as in Theorem 1 that

$$\begin{aligned}\boldsymbol{E}[&\|\hat{Q}_m^*(M_m(Y)) - M_m(Y)\|^2] \\ &\le \boldsymbol{E}[\|Q_m(M_m(Y)) - M_m(Y)\|^2] \\ &\quad + c\sqrt{\frac{\log m}{m}} + O(m^{-1/2}) \quad (20)\end{aligned}$$

where $c = 8B\sqrt{2(kN+1)}$.

Now recall that $\hat{Q}^*$ is an optimal nearest neighbor quantizer for $M(Y)$ and that Q_m is an optimal nearest neighbor quantizer for the conditional distribution of $M_m(Y)$ given $Z_m^{(1)}$. Thus

$$\begin{aligned}\boldsymbol{E}[&\|Q_m(M_m(Y)) - M_m(Y)\|^2 \mid Z_m^{(1)}] \\ &\le \boldsymbol{E}[\|\hat{Q}^*(M_m(Y)) - M_m(Y)\|^2 \mid Z_m^{(1)}] \\ &\le \boldsymbol{E}[\|\hat{Q}^*(M(Y)) - M_m(Y)\|^2 \mid Z_m^{(1)}]\end{aligned}$$

where the first inequality holds because Q_m is optimal for the distribution of $M_m(Y)$ given $Z_m^{(1)}$, and the second inequality follows because $\hat{Q}^*$ is a nearest neighbor quantizer. Therefore,

$$\begin{aligned}\boldsymbol{E}[&\|Q_m(M_m(Y)) - M_m(Y)\|^2 \mid Z_m^{(1)}] \\ &- \boldsymbol{E}[\|\hat{Q}^*(M(Y)) - M(Y)\|^2] \\ &\le \boldsymbol{E}[\|\hat{Q}^*(M(Y)) - M_m(Y)\|^2 \\ &\quad - \|\hat{Q}^*(M(Y)) - M(Y)\|^2 \mid Z_m^{(1)}] \\ &\le 4\sqrt{B}\boldsymbol{E}[\|M_m(Y) - M(Y)\| \mid Z_m^{(1)}].\end{aligned}$$

In the last inequality the uniform boundedness of M and M_n, and the triangle inequality were used. It follows that

$$\begin{aligned}\boldsymbol{E}[&\|Q_m(M_m(Y)) - M_m(Y)\|^2] \\ &\le \boldsymbol{E}[\|\hat{Q}^*(M(Y)) - M(Y)\|^2] + 4\sqrt{Ba_m}.\end{aligned}$$

Combining this with (18) and (20) gives

$$\begin{aligned}\boldsymbol{E}[&\|M(Y) - \hat{Q}_m^*(M(Y))\|^2] \\ &\le \boldsymbol{E}[\|\hat{Q}^*(M(Y)) - M(Y)\|^2] + c\sqrt{\frac{\log m}{m}} \\ &\quad + O(m^{-1/2}) + 8\sqrt{Ba_m} + a_m\end{aligned}$$

and since $Q^*(Y) = \hat{Q}^*(M(Y))$, one finally gets from (17) that

$$\begin{aligned}\boldsymbol{E}[\|X - Q_m^*(Y)\|^2] &\le \boldsymbol{E}[\|X - Q^*(Y)\|^2] + c\sqrt{\frac{\log m}{m}} \\ &\quad + O(m^{-1/2}) + 8\sqrt{Ba_m} + a_m\end{aligned}$$

and the proof is complete. □

V. Empirical Design from Clean Source Samples

So far we have assumed that the training data consisted of samples from the noisy source. In practice, it is often the case that there might be samples available from the clean source. In what follows this situation is explored and the consistency of empirical design is proved. Moreover, it will be shown that, as opposed to the case of empirical design from noisy samples, in this case the convergence rate of $O(\sqrt{\log n/n})$ is easily achievable.

Assume that we are given as training data the i.i.d. samples $X_1, \cdots, X_m$ drawn from the distribution of the clean source X, and that the conditional distribution of Y given X is known. For the sake of concreteness suppose that Y has a conditional density $h(y|x)$ given $X = x$. Then $M(y)$ is estimated again using the first half of the samples and $h(y|x)$. The empirical design of Theorem 2 can be used with the modification that now $\hat{Q}^*$ is defined as

$$\begin{aligned}\hat{Q}_m^* = \arg\min_Q \frac{1}{m/2} &\sum_{i=m/2+1}^{m} \\ &\int_{\mathbb{R}^k} \|M_m(y) - Q(M_m(y))\|^2 h(y|X_i)\, dy \quad (21)\end{aligned}$$

where the minimization is over all N-level nearest neighbor vector quantizers Q whose codepoints lie inside $S(\sqrt{B})$. The following result states that the procedure is consistent in general, and if $h(y|x)$ satisfies some additional conditions, then we can obtain the convergence rate $O(\sqrt{\log m/m})$.

Theorem 3: Assume that source X is bounded as

$$\boldsymbol{P}(\|X\|^2 \le B) = 1$$

and let $(X_1, \cdots, X_m)$ be i.i.d. copies of X. Suppose that B and the conditional density $h(y|x)$ of Y, given X, are known. Then the quantizer $Q_m^* = \widehat{Q}_m^* \circ M_m$ is consistent, i.e.,

$$\lim_{m\to\infty} \boldsymbol{E}[\|X - Q_m^*(Y)\|^2] = D_N^*$$

where the estimator is

$$M_m(y) = \frac{\sum_{i=1}^{m/2} X_i h(y|X_i)}{\sum_{i=1}^{m/2} h(y|X_i)}$$

and

$$\hat{Q}_m^* = \arg\min_Q \frac{1}{m/2} \sum_{i=m/2+1}^{m} \int_{\mathbb{R}^k} \|M_m(y) - Q(M_m(y))\|^2 h(y|X_i)\, dy.$$

If, additionally, $h(y|x)$ is uniformly bounded and $\|Y\|$ is almost surely bounded, then

$$\boldsymbol{E}[\|X - Q_m^*(Y)\|^2] \le D_N^* + c\sqrt{\frac{\log m}{m}} + O(m^{-1/2})$$

where $c = 8B\sqrt{2(kN+1)}$.

Proof: To prove the consistency of Q_m^*, we first show that M_m is consistent. Introduce the notation

$$N_m(y) = \frac{1}{m/2} \sum_{i=1}^{m/2} X_i h(y|X_i)$$
$$g_m(y) = \frac{1}{m/2} \sum_{i=1}^{m/2} h(y|X_i)$$
$$N(y) = \boldsymbol{E}[N_m(y)]$$

and

$$g(y) = \boldsymbol{E}[g_m(y)]$$

(note that g is the density of Y). Then by the strong law of large numbers, for every y we have $\lim_m N_m(y) = N(y)$ and $\lim_m g_m(y) = g(y)$ a.s. Thus for all y such that $g(y) > 0$, we obtain

$$\lim_{m\to\infty} M_m(y) = \lim_{m\to\infty} \frac{N_m(y)}{g_m(y)} = \frac{N(y)}{g(y)} = M(y) \quad \text{a.s.}$$

Since $\|X_i\|^2 \le B$ a.s., it follows that $\|M_m(y)\|^2 \le B$ a.s. for all y. Then the dominated convergence theorem implies that $M_m(y)$ is L_2 consistent, i.e.,

$$\lim_{m\to\infty} \boldsymbol{E}[\|M(Y) - M_m(Y)\|^2] = 0. \tag{22}$$

To finish the consistency part of the theorem, we copy the proof of Theorem 2 after redefining the training data as $Z_m^{(1)} = (X_1, \cdots, X_{m/2})$ and $Z_m^{(2)} = (X_{m/2+1}, \cdots, X_m)$. Clearly, one need only check that the $\hat{Q}_m^*$ defined in (21) satisfies (19), i.e.,

$$\begin{aligned} \boldsymbol{P}\{&\boldsymbol{E}[\|\hat{Q}_m^*(M_m(Y)) - M_m(Y)\|^2 \,|\, Z_m^{(2)}] \\ &- \boldsymbol{E}[\|Q_m(M_m(Y)) - M_m(Y)\|^2] > t \,|\, Z_m^{(1)} = z\} \\ &\le 2\left(V^N\left(16\sqrt{Bk}\right)^{kN} t^{-kN} + 1\right) e^{-mt^2/256B^2}. \end{aligned} \tag{23}$$

This is seen by noticing that according to (21), the empirically optimal $\hat{Q}^*$ has to minimize the functional

$$\frac{1}{m/2} \sum_{i=1}^{m/2} d_Q^{(m)}(X_i)$$

where $d_Q^{(m)}(x)$ is defined as

$$d_Q^{(m)}(x) = \int_{\mathbb{R}^k} \|M_m(y) - Q(M_m(y))\|^2 h(y|x)\, dy.$$

Let Q and Q' be N-level nearest neighbor vector quantizers whose codevectors $\{y_1, \cdots, y_N\}$ and $\{y_1', \cdots, y_N'\}$ lie inside $S(\sqrt{B})$ and satisfy $\|y_j - y_j'\| \le \rho$ for all $1 \le j \le N$. Since $\|M_n(y)\| \le \sqrt{B}$, the nearest neighbor property implies that

$$\begin{aligned} |\,\|M_m(y) - Q'(M_m(y))\|^2 - \|M_m(y) - Q(M_m(y))\|^2| \\ \le 4\rho\sqrt{B} \end{aligned}$$

and therefore,

$$\begin{aligned} |d_Q^{(m)}(x) - d_{Q'}^{(m)}(x)| &\le \int_{\mathbb{R}^k} \Big|\,\|M_m(y) - Q'(M_m(y))\|^2 \\ &\qquad - \|M_m(y) - Q(M_m(y))\|^2\Big| h(y|x)\, dy \\ &\le \int_{\mathbb{R}^k} 4\rho\sqrt{B}\, h(y|x)\, dy \\ &= 4\rho\sqrt{B}. \end{aligned}$$

Thus for fixed $Z_m^{(1)}$, the family of functions $d_Q^{(m)}(x)$ parameterized by Q has the same ϵ-covering as $d_Q(x)$ between (10) and (11). It follows from Theorem 1 that $\hat{Q}^*$ of (21) satisfies (23). The rest of the proof is identical to that of Theorem 2, and we obtain that

$$\begin{aligned} \boldsymbol{E}[\|X - Q_m^*(Y)\|^2] \le D_N^* + c\sqrt{\frac{\log m}{m}} + O(m^{-1/2}) \\ + 8\sqrt{Ba_m} + a_m \end{aligned} \tag{24}$$

where $a_m = \boldsymbol{E}[\|M(Y) - M_m(Y)\|^2]$. Since $\lim_m a_m = 0$ by (22), the consistency part of the theorem is proved.

To obtain the convergence rate it suffices to prove that

$$\boldsymbol{E}[\|M_m(Y) - M(Y)\|] \le \frac{C}{\sqrt{m}} \tag{25}$$

for some constant C, since the boundedness of X implies that $\|M_m(y)\| \le \sqrt{B}$ and thus

$$\begin{aligned} a_m &= \boldsymbol{E}[\|M_m(Y) - M(Y)\|^2] \\ &\le 2\sqrt{B}\, \boldsymbol{E}[\|M_m(Y) - M(Y)\|]. \end{aligned}$$

The term $8\sqrt{Ba_m}$ in (24) comes from the upper bound $\sqrt{\boldsymbol{E}[\|M_m(Y) - M(Y)\|^2]}$ on $\boldsymbol{E}[\|M_m(Y) - M(Y)\|]$ in the proof of Theorem 3, and can be replaced by $8\sqrt{B}\, \boldsymbol{E}[\|M_m(Y) - M(Y)\|]$. Substituting this and (25) into (24) gives the stated convergence rate.

Finally, the estimate (25) is proved. For all y such that $g(y) > 0$ we have

$$\begin{aligned}\|M_m(y)-M(y)\| &= \left\|\frac{N_m(y)}{g_m(y)}-\frac{N(y)}{g(y)}\right\| \\ &\le \left\|\frac{N_m(y)}{g_m(y)}-\frac{N_m(y)}{g(y)}\right\|+\left\|\frac{N_m(y)}{g(y)}-\frac{N(y)}{g(y)}\right\| \\ &= A_1(y)+A_2(y).\end{aligned}$$

The expectation of the first term can be upper-bounded as

$$\begin{aligned}\boldsymbol{E}[A_1(y)] &= \boldsymbol{E}\left[\left\|\frac{N_m(y)(g_m(y)-g(y))}{g_m(y)g(y)}\right\|\right] \\ &\le \frac{\sqrt{B}}{g(y)}\boldsymbol{E}[|g_m(y)-g(y)|] \\ &\le \frac{\sqrt{B}}{g(y)}\sqrt{\boldsymbol{E}[|g_m(y)-g(y)|^2]} \\ &= \frac{\sqrt{2B}}{g(y)\sqrt{m}}\sqrt{\mathrm{Var}\,[h(y|X)]} \\ &\le \frac{C_1}{g(y)\sqrt{m}}\end{aligned}$$

for a constant C_1, where the first inequality follows from the fact that the $\|X_i\| \le \sqrt{B}$ a.s., and last inequality holds because $h(y|x)$ is uniformly bounded. For the second term, we similarly have

$$\begin{aligned}\boldsymbol{E}[A_2(y)] &= \frac{1}{g(y)}\boldsymbol{E}[\|N_m(y)-N(y)\|] \\ &\le \frac{1}{g(y)}\sqrt{\boldsymbol{E}[\|N_m(y)-N(y)\|^2]} \\ &= \frac{1}{g(y)}\sqrt{\frac{2}{m}}\sqrt{\boldsymbol{E}[\|Xh(y|X)-N(y)\|^2]} \\ &\le \frac{C_2}{g(y)\sqrt{m}}\end{aligned}$$

for some constant C_2. By the assumption on the distribution of Y, $g(y)=0$ outside some compact set S, so that

$$\boldsymbol{E}[\|M_m(Y)-M(Y)\|] \le \int_S (A_1(y)+A_2(y))g(y)\,dy \le \frac{C}{\sqrt{m}}$$

for a constant C, which proves (25). □

VI. Conclusion

We have investigated the problem of empirical vector quantizer design for noisy channels or noisy sources. The notion of empirical distortion minimization was suitably defined for both cases, and proofs of consistency of the methods were given. For the noisy channel problem it was shown that the average squared distortion of an optimal vector quantizer designed from observing m clean i.i.d. training vectors converges, in expectation, as the training set size grows, to the minimum possible mean-squared error obtainable for quantizing the clean source and transmitting across a discrete memoryless noisy channel. The convergence rate $O(\sqrt{\log m/m})$ was also obtained. The comparison of this rate with that obtained in [9] for empirical design for ordinary vector quantizers shows that noisy channel vector quantizer design is not a harder problem from a statistical viewpoint. Consistency of an empirical design method for sources corrupted by noise was also proved under some regularity conditions. Determining a good convergence rate is an open problem for the case when only noisy training samples are available. The estimation problem involved in the design indicates that, in general, this problem is significantly harder than ordinary vector quantizer design. When training samples from the clean source are available, we can obtain the same convergence rate as for the standard vector quantizer design problem or for the noisy channel problem under mild conditions on the noise distribution.

The method of empirical distortion minimization (searching for a quantizer globally optimal over the training samples) is computationally prohibitive in practice. It is therefore of practical significance to carry out analyses similar to what we presented here for suboptimal, but computationally feasible methods of design. Such an analysis of consistency was given for the generalized Lloyd–Max algorithm in ordinary vector quantizer design by Sabin and Gray [11]. An interesting area of future research would be to provide convergence rates for suboptimal algorithms for ordinary, as well as noisy channel or noisy source vector quantizer design.

Appendix
Proof of Lemma 1

The estimate with the required property is a k-dimensional extension of the estimator proposed by Devroye [33]. The proof is based on [33], where convergence in expectation was proved. First some notation is introduced. $\phi(t)=\boldsymbol{E}[e^{itX}]$ and $\phi(t)\zeta(t)=\boldsymbol{E}[e^{itY}]$ are the characteristic functions of X and Y, respectively, and the empirical characteristic function of the data is denoted by

$$\phi_n(t)=\frac{1}{n}\sum_{j=1}^{n} e^{itY_j}.$$

The estimator uses a *kernel function* K: $\mathbb{R}^k \to \mathbb{R}$ with $\int K(x)\,dx = 1$, such that its Fourier transform $\psi(t) = \int e^{itx}K(x)\,dx$ satisfies $\sup_{t\in\mathbb{R}^k}|\psi(t)| < \infty$ and $\psi(t)=0$ if $t \notin S(c)$ for some constant $c<\infty$, where $S(r)$ denotes the k-dimensional ball of radius r centered at the origin. We also define a *smoothing parameter* $h>0$, a *tail parameter* $T>0$, and a *noise-control parameter* $r\ge 0$. All of these parameters may change with the sample size n. Introduce the set $A_r=\{t\colon |\zeta(t)|<r\}$, and let $\Re\{z\}$ denote the real part of the complex number z. Our estimate is defined as follows:

$$f_n(x)=\begin{cases}0, & \text{if } \|x\|\ge T\\ \dfrac{1}{(2\pi)^k}\Re\left\{\displaystyle\int_{\mathbb{R}^k-A_r} e^{-itx}\right. & \\ \qquad\left.\cdot\,\psi(ht)\zeta^{-1}(t)\phi_n(t)\,dt\right\}, & \text{if } \|x\|<T.\end{cases}$$

We claim that this estimate satisfies the required consistency property if the parameters vary with n as follows:

$$\lim_{n\to\infty} T=\infty \tag{26}$$

$$\lim_{n\to\infty} h=0 \tag{27}$$

$$\lim_{n\to\infty} T^k \lambda(A_r \cap S(c/h)) = 0 \tag{28}$$

$$\lim_{n\to\infty} \frac{T^{2k}\log n}{nh^{2k}r^2} = 0 \tag{29}$$

where λ denotes the Lebesgue measure.

To see why the estimate is consistent, we introduce the notation $K_h(x) = (1/h^k)K(x/h)$ and

$$(f * K_h)(x) = (2\pi)^{-k} \int e^{-itx}\psi(th)\phi(t)\, dt.$$

Next define the auxiliary function

$$q_n(x) = \frac{1}{(2\pi)^k}\Re\left\{\int_{\mathbb{R}^k - A_r} e^{-itx}\psi(ht)\phi(t)dt\right\}$$

and write the decomposition

$$\begin{aligned}\int |f_n - f| &\le \int_{\|x\|<T} |f_n - q_n| + \int_{\|x\|<T} |q_n - f * K_h| \\ &\quad + \int |f * K_h - f| + \int_{\|x\|\ge T} f \\ &\le \sqrt{V_k T^k}\sqrt{\int_{\|x\|<T} (f_n - q_n)^2} \\ &\quad + V_k T^k \sup_x |q_n(x) - (f * K_h)(x)| \\ &\quad + \int |f * K_h - f| + \int_{\|x\|\ge T} f \\ &= I + II + III + IV\end{aligned}$$

where V_k denotes the volume of the unit ball in $\mathbb{R}^k$. It is now shown that each of the four terms tends to zero as $n \to \infty$, almost surely.

Clearly $IV \to 0$ by (26). Since $\int K = 1$ and $h \to 0$ by (27), we have by the well-known "approximation of the identity" property of the family $\{K_h; h > 0\}$ that $III \to 0$ (see, e.g., [34, Theorem 9.6]). Also,

$$\begin{aligned}II &\le \frac{V_k T^k}{(2\pi)^k}\int_{A_r} |\psi(th)| \cdot |\phi(t)|\, dt \\ &\le \frac{V_k C T^k}{(2\pi)^k}\lambda(A_r \cap S(c/h))\end{aligned}$$

which converges to zero by (28). To show that $\boldsymbol{E}[I] \to 0$, we introduce the random variables

$$\begin{aligned}L_{n,j}(x) &= \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k - A_r} e^{it(Y_j - x)}\psi(th)\zeta^{-1}(t)\, dt \\ K_{n,j}(x) &= \Re L_{n,j}(x).\end{aligned}$$

Then

$$\begin{aligned}(\boldsymbol{E}[I])^2 &= V_k T^k\left(\boldsymbol{E}\sqrt{\int_{\|x\|<T} (f_n - q_n)^2}\right)^2 \\ &\le V_k T^k \int_{\|x\|<T} \boldsymbol{E}[(f_n - q_n)^2] \\ &= V_k T^k \int_{\|x\|<T} \boldsymbol{E}\Bigg[\Bigg(\frac{1}{n(2\pi)^k}\sum_{j=1}^n \\ &\qquad \Re\left\{\int_{\mathbb{R}^k - A_r} e^{-itx}\psi(th)\zeta^{-1}(t)e^{itY_j}\, dt\right. \\ &\qquad \left.\left. - \int_{\mathbb{R}^k - A_r} e^{-itx}\psi(th)\zeta^{-1}(t)\boldsymbol{E}[e^{itY_1}]\, dt\right\}\Bigg)^2\Bigg] \\ &= \frac{V_k T^k}{n}\int_{\|x\|<T} \mathrm{Var}\Bigg[\Re\Bigg\{\frac{1}{(2\pi)^k}\int_{\mathbb{R}^k - A_r} \\ &\qquad e^{-itx}\psi(th)\zeta^{-1}(t)e^{itY_j}dt\Bigg\}\Bigg] \\ &\le \frac{V_k T^k}{n}\boldsymbol{E}\left[\int_{\|x\|<T} K_{n,1}^2(x)\right] dx \\ &\le \frac{V_k T^k}{n}\boldsymbol{E}\left[\int_{\|x\|<T} |L_{n,1}(x)|^2\right] dx \\ &\le \frac{V_k T^k}{n(2\pi)^k}\int_{\mathbb{R}^k - A_r} \frac{|\psi(th)|^2}{|\zeta(t)|^2}\, dt \text{ (by Parseval's identity)} \\ &\le \text{constant} \cdot \frac{T^k}{nh^k r^2}\end{aligned}$$

which converges to zero by (29). Summarizing, we have proved that for every density f

$$\lim_{n\to\infty} \boldsymbol{E}\left[\int |f(x) - f_n(x)|\, dx\right] = 0.$$

To prove convergence with probability one, recall a powerful inequality of McDiarmid [35] (see also Devroye [36]). According to this inequality, if $g\colon \mathbb{R}^n \to \mathbb{R}$ is an arbitrary function satisfying the boundedness condition

$$\begin{aligned}\sup_{x_1,\cdots,x_n,x_j'\in\mathbb{R}^k} &|g(x_1, \cdots, x_{j-1}, x_j, x_{j+1}, \cdots, x_n) \\ &- g(x_1, \cdots, x_{j-1}, x_j', x_{j+1}, \cdots, x_n)| \le c_j\end{aligned}$$

then for any independent random variables $Y_1, \cdots, Y_n$

$$\begin{aligned}\boldsymbol{P}\{|g(Y_1, \cdots, Y_n) - \boldsymbol{E}[g(Y_1, \cdots, Y_n)]| > \epsilon\} \\ \le 2e^{-\epsilon^2/\left(\sum_{i=1}^n c_i^2\right)}.\end{aligned}$$

We apply this inequality to the L_1-error

$$g(Y_1, \cdots, Y_n) = \int |f_n - f|.$$

It suffices to obtain a good upper bound on the variability of the L_1-error if we replace Y_j by an arbitrary Y_j'. Denote the modified estimate by $f_n'(x)$. Then (see (30) at the top of the following page). Therefore, McDiarmid's inequality implies that

$$\boldsymbol{P}\left\{\left|\int |f_n - f| - \int |f_n' - f|\right| > \epsilon\right\} \le 2e^{-\text{constant}\cdot\epsilon^2 nh^k r^2/T^k}.$$

The upper bound is summable for every $\epsilon > 0$ if

$$\frac{T^k \log n}{nh^k r^2} = o(1)$$

which is satisfied by (29). Thus by the Borel–Cantelli lemma, $\int |f_n - f| \to 0$ with probability one. To complete the proof

$$\begin{aligned}\left|\int |f_n - f| - \int |f_n' - f|\right| &\le \int |f_n - f_n'| \\ &= \frac{1}{n}\int_{\|x\|<T}\left|\frac{1}{(2\pi)^k}\Re\left\{\int_{\mathbb{R}^k - A_r} e^{-itx}\frac{\psi(th)}{\zeta(t)}\,(e^{itY_j} - e^{itY_j'})\,dt\right\}\right|dx \\ &\le \frac{1}{n}\int_{\|x-Y_j\|<T}\left|\frac{1}{(2\pi)^k}\int_{\mathbb{R}^k - A_r} e^{-itx}\frac{\psi(th)}{\zeta(t)}\,dt\right|dx \\ &\quad + \frac{1}{n}\int_{\|x-Y_j'\|<T}\left|\frac{1}{(2\pi)^k}\int_{\mathbb{R}^k - A_r} e^{-itx}\frac{\psi(th)}{\zeta(t)}\,dt\right|dx \\ &\le \frac{1}{n(2\pi)^k}\sqrt{V_kT^k\int_{\|x-Y_j\|<T}\left|\int_{\mathbb{R}^k - A_r} e^{-itx}\frac{\psi(th)}{\zeta(t)}\,dt\right|^2 dx} \\ &\quad + \frac{1}{n(2\pi)^k}\sqrt{V_kT^k\int_{\|x-Y_j'\|<T}\left|\int_{\mathbb{R}^k - A_r} e^{-itx}\frac{\psi(th)}{\zeta(t)}\,dt\right|^2 dx} \\ &\quad\ \text{(by the Cauchy–Schwarz inequality)} \\ &\le \frac{2}{n(2\pi)^k}\sqrt{V_kT^k\int_{\mathbb{R}^k}\left|\int_{\mathbb{R}^k} e^{-itx}\frac{\psi(th)}{\zeta(t)}I_{\{t\in\mathbb{R}^k - A_r\}}\,dt\right|^2 dx} \\ &= \frac{2}{n(2\pi)^k}\sqrt{V_kT^k\int_{\mathbb{R}^k}\left|\frac{\psi(th)}{\zeta(t)}\right|^2 I_{\{t\in\mathbb{R}^k - A_r\}}\,dt} \\ &\quad\ \text{(by Parseval's identity)} \\ &\le \frac{2}{n(2\pi)^k}\sqrt{\frac{V_kT^k}{r^2}\int_{\mathbb{R}^k}|\psi(th)|^2\,dt} \\ &\le \text{constant}\cdot\frac{T^{k/2}}{nh^{k/2}r}. \end{aligned} \tag{30}$$

of the lemma, it suffices to demonstrate the existence of the parameters of the estimate satisfying the conditions (26)–(29).

For each positive integer l, set $h = 1/l$ and

$$r = \sup\{u\colon \lambda(A_u \cap S(c/h)) < 1/l\}.$$

To see that such an r exists, note that since $\zeta(t) \neq 0$ almost everywhere, the continuity of the Lebesgue measure implies that for any fixed $v > 0$, $\lambda(A_u \cap S(v)) \to 0$ as $u \to 0$. Let $n_l = \lceil (l/(h^{2k}r^2))^{1+\delta}\rceil$ for $\delta > 0$. For all $n \in [n_l, n_{l+1})$, define h and r to be the same as their values for n_l. Then as $l \to \infty$, $n_l \to \infty$, and $h \to 0$, and therefore (27) is satisfied. Also, $\lambda(A_r \cap S(c/h)) \to 0$, and if $\beta = \delta/(1+\delta)$, then $n^{1-\beta}h^{2k}r^2 \to \infty$. Define

$$T^{2k} = \min\left(n^{1-\beta}h^{2k}r^2,\ \frac{1}{\lambda(A_r \cap S(c/h))}\right).$$

Then $T \to \infty$ (so (26) is satisfied) and $T^k\lambda(A_r \cap S(c/h)) \to 0$, so that (28) is satisfied. Finally, $T^k/(n^{1-\beta}h^kr^2) \to 0$, which implies (29). □

ACKNOWLEDGMENT

The authors wish to thank L. Györfi for helpful discussions and J. Fan for pointing out some relevant references in the statistical literature.

REFERENCES

[1] S. P. Lloyd, “Least squared quantization in pcm,” unpublished memorandum, Bell Labs., 1957; reprinted in *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[2] J. Max, “Quantizing for minimum distortion,” *IEEE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.

[3] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.

[4] R. M. Gray, J. C. Kieffer, and Y. Linde, “Locally optimum block quantizer design,” *Inform. Contr.*, vol. 45, pp. 178–198, 1980.

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

[6] E. A. Abaya and G. L. Wise, “Convergence of vector quantizers with applications to optimal quantization,” *SIAM J. Appl. Math.*, vol. 44, pp. 183–189, 1984.

[7] M. J. Sabin, “Global convergence and empirical consistency of the generalized Lloyd algorithm,” Ph.D. dissertation, Stanford Univ., Stanford, CA, 1984.

[8] D. Pollard, “A central limit theorem for k-means clustering,” *Ann. Prob.*, vol. 10, no. 4, pp. 919–926, 1982.

[9] T. Linder, G. Lugosi, and K. Zeger, “Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding,” *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1728–1740, Nov. 1994.

[10] P. A. Chou, “The distortion of vector quantizers trained on n vectors decreases to the optimum as $O_p(1/n)$,” in *Proc. IEEE Int. Symp. on Information Theory* (Trondheim, Norway, 1994).

[11] M. J. Sabin and R. M. Gray, “Global convergence and empirical consistency of the generalized Lloyd algorithm,” *IEEE Trans. Inform. Theory*, vol. IT-32, no. 2, pp. 148–155, Mar. 1986.

[12] M. Merhav and J. Ziv, “On the amount of side information required for lossy data compression,” preprint, 1995.

[13] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” in *IRE Nat. Conv. Rec.*, pt. 4, 1959, pp. 138–143.

[14] J. Dunham and R. M. Gray, "Joint source and noisy channel trellis encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 4, pp. 516–519, July 1981.

[15] K. Zeger and V. Manzella, "Asymptotic bounds on optimal noisy channel quantization via random coding," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1926–1938, Nov. 1994.

[16] S. McLaughlin and D. Neuhoff, "Asymptotic bounds in source-channel coding," in *Proc. IEEE Int. Symp. on Information Theory* (Budapest, Hungary, 1991).

[17] N. Farvardin and V. Vaishampayan, "Optimal quantizer design for noisy channels: An approach to combined source-channel coding," *IEEE Trans. Inform. Theory*, vol. IT-33, no. 6, pp. 827–838, Nov. 1987.

[18] E. Ayanoglu and R. M. Gray, "The design of joint source and channel trellis waveform coders," *IEEE Trans. Inform. Theory*, vol. IT-33, no. 6, pp. 855–865, Nov. 1987.

[19] R. L. Dobrushin and B. S. Tsybakov, "Information transmission with additional noise," *IRE Trans. Inform. Theory*, vol. IT-18, pp. 293–304, 1962.

[20] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[21] T. Fine, "Optimum mean-square quantization of a noisy input," *IEEE Trans. Inform. Theory,* vol. IT-11, pp. 293–294, Apr. 1965.

[22] D. J. Sakrison, "Source encoding in the presence of random disturbance," *IEEE Trans. Inform. Theory,* vol. IT-14, pp. 165–167, Jan. 1968.

[23] J. K. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 406–411, July 1970.

[24] E. Ayanoglu, "On optimal quantization of noisy sources," *IEEE Trans. Inform. Theory*, vol. 36, no. 6, pp. 1450–1452, Nov. 1990.

[25] Y. Ephraim and R. M. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. Inform. Theory*, vol. 34, no. 4, pp. 826–834, July 1988.

[26] A. Rao, D. Miller, K. Rose, and A. Gersho, "Generalized vector quantization: Jointly optimal quantization and estimation," in *Proc. IEEE Int. Symp. on Information Theory* (Whistler, B.C., Canada, Sept. 1995).

[27] H. Kumazawa, M. Kasahara, and T. Namekawa, "A construction of vector quantizers for noisy channels," *Electron. and Eng. in Japan*, vol. 67-B, no. 4, pp. 39–47, 1984.

[28] K. Zeger and A. Gersho, "Vector quantizer design for memoryless noisy channels," in *Proc. IEEE Int. Conf. on Communications*, 1988, pp. 1593–1597.

[29] A. R. Barron, "Complexity regularization with application to artificial neural networks," in *Nonparametric Functional Estimation and Related Topics*, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer, 1991, NATO ASI Ser., pp. 561–576.

[30] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, 1963.

[31] L. Devroye, *A Course in Density Estimation.* Boston, MA: Birkhäuser, 1987.

[32] G. T. Hwang and L. A. Stefanski, "Monotonicity of regression functions in structural measurement error models," *Stat. Prob. Lett.*, vol. 20, pp. 113–116, 1993.

[33] L. Devroye, "Consistent deconvolution in density estimation," *Can. J. Stat.*, vol. 17, no. 2, pp. 235–239, 1989.

[34] R. L. Wheeden and A. Zygmund, *Measure and Integral.* New York: Dekker, 1977.

[35] C. McDiarmid, "On the method of bounded differences," in *Surv. in Combinatorics 1989*. Cambridge, U.K.: Cambridge Univ. Press, 1989, pp. 148–188.

[36] L. Devroye, "Exponential inequalities in nonparametric estimation," in *Nonparametric Functional Estimation and Related Topics*, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer, 1991, NATO ASI Ser., pp. 31–44.