

## Lagrangian Empirical Design of Variable-Rate Vector Quantizers: Consistency and Convergence Rates

Tamás Linder, *Senior Member, IEEE*

**Abstract**—The Lagrangian formulation of variable-rate vector quantization is known to yield useful necessary conditions for quantizer optimality and generalized Lloyd algorithms for quantizer design. In this correspondence, the Lagrangian formulation is demonstrated to provide a convenient framework for analyzing the empirical design of variable-rate vector quantizers. In particular, the consistency of empirical design based on minimizing the Lagrangian performance over a stationary and ergodic training sequence is shown for sources with finite second moment. The finite sample performance is also studied for independent training data and sources with bounded support.

**Index Terms**—Consistency, convergence rates, empirical design, Lagrangian performance, variable-rate quantization.

### I. INTRODUCTION

In empirical quantizer design we are presented with a finite sample of vector-valued data generated by an unknown source, and the goal is to construct a vector quantizer of given dimension and rate yielding best performance on future data generated by the same source. In this sense, empirical design is a process of learning from examples where the effectiveness of learning is measured by applying the designed quantizer to future data.

In fixed-rate quantization, where the rate of the quantizer is determined by the number of its codevectors, empirical design has been studied in detail. Pollard [1] showed that the design based on minimizing the empirical quantization error over a stationary and ergodic training sequence is consistent, i.e., the distortion of the designed quantizer converges to the optimum distortion as the length of the sequence increases without bound. In practice, one has only a finite amount of training data, so it is of interest to quantify how quantizer performance improves as the size of the training data increases. The finite sample performance (rate of convergence) for empirically optimal quantizers trained on independent data was studied in [2], and the convergence rate derived there was shown to be optimal in the minimax sense in [3]. Further results on empirical design of fixed-rate quantizers and related statistical clustering problems are given in [4]–[8].

Our goal is to study the empirical design of variable-rate quantizers where the quantizer rate is measured by the average length of a variable-length lossless code used to encode the quantizer output. To avoid some of the considerable difficulties inherent in the traditional formulation of optimality for variable-rate quantization (see, e.g., [9]), we use the Lagrangian formulation of Chou *et al.* [10]. The Lagrangian formulation matches the generalized Lloyd algorithm [10] that is used in practice to design variable-rate quantizers, and it has recently been used to rigorously prove and generalize Zador's classic result for the asymptotic behavior of optimal entropy-constrained quantizers [11]. More importantly, the Lagrangian formulation yields an optimality condition

Manuscript received February 12, 2002; revised May 24, 2002. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The material of this correspondence was presented in part at the Advanced School on the Principles of Nonparametric Learning, CISM, Udine, Italy, July 2001 and at the IEEE International Symposium on Information Theory, Lausanne, Switzerland, June/July 2002.

The author is with the Department of Mathematics and Statistics and the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: linder@mast.queensu.ca).

Communicated by R. Zamir, Associate Editor for Source Coding.  
Digital Object Identifier 10.1109/TIT.2002.804114.

(the generalized nearest neighbor condition) which proves essential in our development. The Lagrangian formulation allows us to adapt techniques from the theory of empirical processes and statistical learning that have previously been used to analyze empirical fixed-rate quantizer design.

This correspondence is organized as follows. In Section II, the basic definitions of variable-rate quantization are introduced in both the traditional and the Lagrangian settings. In Section III, we prove the consistency of Lagrangian design for stationary and ergodic training data for sources with finite second moment. In Section IV, the finite sample performance is studied for independent training data drawn from a source with bounded support. After establishing a key property of optimal quantizers for sources with bounded support, we show that the (Lagrangian) distortion of the empirically optimal variable-rate vector quantizer converges to that of the truly optimal variable-rate vector quantizer at a rate  $O(n^{-1/2})$ , where  $n$  is the size of the training data. This result matches the convergence rate results developed for fixed-rate empirical design [2], [3].

In variable-rate quantization, one can alternatively define the rate of a quantizer by the entropy of the quantizer output [10], [12]. Although our rate definition uses the average code length, the difference between the two approaches is not significant, and the entropy-constrained counterparts of our results are easily seen to hold. Section V discusses the connection with entropy-constrained quantization.

### II. PRELIMINARIES

A variable-rate vector quantizer  $q$  is described by an encoder  $\alpha: \mathbb{R}^d \rightarrow \mathcal{I}$ , where  $\mathcal{I}$  is a countable index set, a decoder  $\beta: \mathcal{I} \rightarrow \mathbb{R}^d$ , and an index coder  $\psi: \mathcal{I} \rightarrow \{0, 1\}^*$ , where  $\{0, 1\}^*$  denotes the collection of all finite-length binary strings.  $\psi$  is assumed to be a prefix-free binary code. If  $\mathcal{I}$  is finite with  $N$  elements, without loss of generality we always take  $\mathcal{I} = \{1, \dots, N\}$ ; otherwise,  $\mathcal{I}$  is taken to be the set of all positive integers. Thus,  $q$  maps any point  $x \in \mathbb{R}^d$  into one of the codevectors in the codebook  $\{\beta(i); i \in \mathcal{I}\}$ , via the rule  $q(x) = \beta(\alpha(x))$ . Letting  $S_i = \{x: \alpha(x) = i\}$  denote the quantization cells and  $y_i = \beta(i)$  the codepoints of  $q$  for  $i \in \mathcal{I}$ , we have

$$q(x) = y_i, \quad \text{if and only if } x \in S_i.$$

We assume that the  $d$ -dimensional random vector  $X$  with distribution  $\mu$  has finite second moment  $\mathbb{E}\|X\|^2 < \infty$ , and define the mean-squared distortion of  $q$  in the usual way

$$D(\mu, q) \triangleq \mathbb{E}\|X - q(X)\|^2$$

where  $\|\cdot\|$  denotes the Euclidean norm on  $\mathbb{R}^d$ .

The length function  $\ell: \mathcal{I} \rightarrow \{0, 1, 2, \dots\}$  associates with each index  $i$  the length of the corresponding codeword  $\psi(i)$ , i.e.,  $\ell(i) = \text{length}(\psi(i))$ . The rate of the variable-rate quantizer  $q$  is defined as the expected codeword length

$$r(\mu, q) \triangleq \mathbb{E}\ell(\alpha(X)) = \sum_{i \in \mathcal{I}} \ell(i) \mathbb{P}\{q(X) = y_i\}.$$

A length function  $\ell$  is called *admissible* if it satisfies Kraft's inequality  $\sum_{i \in \mathcal{I}} 2^{-\ell(i)} \leq 1$ . Since a prefix-free code  $\psi$  exists for a given length function  $\ell$  if and only if  $\ell$  is admissible [13], for our purposes it is enough to specify a quantizer  $q$  by its encoder  $\alpha$ , decoder  $\beta$ , and admissible length function  $\ell$ . In this case, we write  $q \equiv (\alpha, \beta, \ell)$ .

In variable-rate quantization, ideally one is interested in the minimum distortion achievable for a given rate  $R \geq 0$

$$\delta_R^*(\mu) \triangleq \inf_{q: r(\mu, q) \leq R} D(\mu, q).$$

However, the traditional formulation of quantizer optimality for variable-rate quantizers is hard to handle. For example, the cells of an optimal quantizer achieving  $\delta_R^*(\mu)$  can have a complex structure [9], as opposed to the fixed-rate case [14]. Therefore, we use the Lagrangian formulation of optimal distortion-rate tradeoff for variable-rate quantization that was first introduced by Chou *et al.* [10]. For a variable-rate quantizer  $q \equiv (\alpha, \beta, \ell)$ , and for  $\lambda > 0$ , define the *Lagrangian distortion* by

$$\Delta_\lambda(\mu, q) \triangleq D(\mu, q) + \lambda r(\mu, q) = \mathbb{E} \{ \|X - q(X)\|^2 + \lambda \ell(\alpha(X)) \}$$

and the optimal Lagrangian performance by

$$\Delta_\lambda^*(\mu) \triangleq \inf_q \Delta_\lambda(\mu, q)$$

where the infimum is taken over all variable-rate quantizers.

Unlike the traditional formulation, the Lagrangian formulation yields Lloyd optimality conditions for variable-rate vector quantizers [10]. Namely, in the Lagrangian formulation, a necessary condition for optimality is that each of the three components of the quantizer be optimal for the other two. In particular, for a given decoder  $\beta$  and length function  $\ell$ , for  $\mu$  almost all  $x$  an optimal encoder must satisfy

$$\alpha(x) = \arg \min_{i \in \mathcal{I}} (\|x - \beta(i)\|^2 + \lambda \ell(i)) \quad (1)$$

(ties are broken arbitrarily). Since  $\ell$  satisfies Kraft's inequality, the above minimum is easily seen to be achieved by some index  $i$  even if  $\mathcal{I}$  is not finite. An  $\alpha$  satisfying (1) for all  $x \in \mathbb{R}^d$  is called a *generalized nearest neighbor encoder*.

We also need the following auxiliary result from [15] showing the existence of quantizers that minimize the Lagrangian distortion.<sup>1</sup>

*Theorem 1 [15]:* For any  $\mu$  with finite second moment and  $\lambda > 0$  there is a variable-rate quantizer  $q_\lambda^*$  with a generalized nearest neighbor encoder such that

$$\Delta_\lambda(\mu, q_\lambda^*) = \Delta_\lambda^*(\mu).$$

It is well known that for a given rate  $R$ ,  $\delta_R^*(\mu)$  is achievable by a quantizer  $q$  minimizing  $\Delta_\lambda(\mu, q)$  for some  $\lambda > 0$  if and only if  $\delta_R^*(\mu)$  (as a function of  $R$ ) coincides with its convex hull  $\hat{\delta}_R^*(\mu)$  at this rate. Thus, Lagrangian optimization can be used to find the convex hull of  $\delta_R^*(\mu)$ . For values of  $R$  such that  $\hat{\delta}_R^*(\mu)$  is strictly less than  $\delta_R^*(\mu)$ , optimal variable-rate quantizers cannot be obtained by the Lagrangian method. However, this is not a serious limitation in practical applications since any rate and distortion pair  $(R, \hat{\delta}_R^*(\mu))$  on the convex hull can be achieved by "time sharing" between two quantizers that achieve the convex hull, i.e., two quantizers that can be obtained by Lagrangian minimization.

### III. CONSISTENCY OF LAGRANGIAN EMPIRICAL DESIGN

We are interested in the performance of quantizers designed using a finite training sequence. Let  $X_1^n = X_1, \dots, X_n$  be a stationary and ergodic sequence of  $\mathbb{R}^d$ -valued random vectors with marginal distribution  $\mu$ . We assume that  $X_1^n$  and  $X$  are independent. Let  $\mu_n$  denote the empirical distribution of  $X_1^n$ , i.e.,  $\mu_n$  places weight  $1/n$  at each point  $X_k, k = 1, \dots, n$ . Fix  $\lambda > 0$  and define an *empirically optimal* variable-rate quantizer  $q_n^*$  to be one that minimizes the empirical Lagrangian distortion

$$\Delta_\lambda(\mu_n, q_n^*) = \inf_q \Delta_\lambda(\mu_n, q) = \Delta_\lambda^*(\mu_n)$$

<sup>1</sup>More precisely, the result in [15] is for entropy-constrained quantizers, but a slight modification of the proof yields Theorem 1.

i.e.,

$$q_n^* \triangleq \arg \min_{q \equiv (\alpha, \beta, \ell)} \frac{1}{n} \sum_{k=1}^n \|X_k - \beta(\alpha(X_k))\|^2 + \lambda \ell(\alpha(X_k)).$$

We will always assume (as we may by Theorem 1) that the encoder  $\alpha_n^*$  of  $q_n^*$  is a generalized nearest neighbor encoder.

The performance of an empirically optimal variable-rate quantizer  $q_n^* = (\alpha_n^*, \beta_n^*, \ell_n^*)$  is measured by its Lagrangian test distortion, given by

$$\Delta_\lambda(\mu, q_n^*) \triangleq \mathbb{E} [\|X - \beta_n^*(\alpha_n^*(X))\|^2 + \lambda \ell_n^*(\alpha_n^*(X))] X_1^n].$$

The quantity  $\Delta_\lambda(\mu, q_n^*)$  is the Lagrangian distortion of  $q_n^*$  in quantizing independent (future) test data.  $\Delta_\lambda(\mu, q_n^*)$  is a random variable since  $q_n^*$  depends on the training sequence  $X_1^n$ , and it is easy to see that  $\mathbb{E} \Delta_\lambda(\mu, q_n^*) \geq \Delta_\lambda^*(\mu)$ .

The following theorem shows that the design-based empirical Lagrangian distortion minimization is consistent in the sense that as  $n \rightarrow \infty$ ,  $\Delta_\lambda(\mu, q_n^*)$  converges to the optimal Lagrangian distortion  $\Delta_\lambda^*(\mu)$  for almost every realization of the stationary and ergodic training sequence.

*Theorem 2 (CONSISTENCY OF LAGRANGIAN EMPIRICAL DESIGN):* For any  $\lambda > 0$ , the sequence of empirically optimal variable-rate quantizers  $q_n^*, n = 1, 2, \dots$ , minimizing the empirical Lagrangian distortion satisfies

$$\lim_{n \rightarrow \infty} \Delta_\lambda(\mu, q_n^*) = \Delta_\lambda^*(\mu) \quad \text{a.s.}$$

Pollard's consistency theorem [1] is based on bounding quantizer performance using the so-called  $L_2$  Wasserstein distance  $\rho(\mu, \nu)$  between two probability distributions  $\mu$  and  $\nu$  on  $\mathbb{R}^d$  with finite second moment. This distance is defined by

$$\rho(\mu, \nu) \triangleq \inf_{X \sim \mu, Y \sim \nu} (\mathbb{E} \|X - Y\|^2)^{1/2}$$

where the infimum is taken over all joint distributions of two random vectors  $X$  and  $Y$  such that  $X$  has distribution  $\mu$ , and  $Y$  has distribution  $\nu$  (denoted by  $X \sim \mu$  and  $Y \sim \nu$ , respectively). It is known that the infimum defining  $\rho(\mu, \nu)$  is a minimum, and that  $\rho(\mu, \nu)$  is a metric on the space of probability distributions on  $\mathbb{R}^d$  with finite second moment (see, e.g., [16]).

Our proof uses Pollard's technique, but we need to find a modified version of  $\rho(\mu, \nu)$  suitable for variable-rate quantization.

Let  $\mathcal{D}$  denote the set of all discrete distributions on  $\mathbb{R}^d$  with finite second moment and finite entropy. That is,  $\nu \in \mathcal{D}$  if and only if  $\nu$  is concentrated on a finite or countably infinite set  $\{x_i; i \in \mathcal{I}_\nu\} \subset \mathbb{R}^d$ , and satisfies

$$\sum_{i \in \mathcal{I}_\nu} \|x_i\|^2 \nu(x_i) < \infty, \quad - \sum_{i \in \mathcal{I}_\nu} \nu(x_i) \log_2 \nu(x_i) < \infty.$$

For any  $\nu \in \mathcal{D}$ , let  $L_\nu$  denote the minimum expected code length over all admissible code-length functions  $\ell: \mathcal{I}_\nu \rightarrow \{0, 1, \dots\}$

$$L_\nu = \min_\ell \sum_{i \in \mathcal{I}_\nu} \ell(i) \nu(x_i). \quad (2)$$

Note that a minimizing admissible  $\ell$  always exists [17], and that  $L_\nu < \infty$  by the Shannon–Fano bound [13].

For  $\lambda > 0$ ,  $\mu$  with finite second moment, and  $\nu \in \mathcal{D}$  define

$$\rho_\lambda(\mu, \nu) \triangleq (\rho(\mu, \nu)^2 + \lambda L_\nu)^{1/2}.$$

To interpret  $\rho_\lambda(\mu, \nu)$ , suppose  $X \sim \mu$  and  $Y \sim \nu$  achieve  $\rho(\mu, \nu)$ , and let  $\ell_\nu$  be an admissible code length achieving  $L_\nu$ . Then  $Y$  can be viewed as the output of a variable-rate "random quantizer" that, to each  $x$ , assigns the reproduction vector  $x_i$  and a binary codeword of length

$\ell_\nu(i)$  with probability  $\mathbb{P}\{Y = x_i | X = x\}$ . The quantity  $\rho_\lambda(\mu, \nu)^2$  is the Lagrangian distortion of this random quantizer.

In this interpretation, the next lemma states that deterministic quantizers always outperform random quantizers in the Lagrangian sense.

*Lemma 1:*

$$\Delta_\lambda^*(\mu) = \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu, \nu)^2.$$

*Proof:* Suppose  $q \equiv (\alpha, \beta, \ell)$  is a variable-rate quantizer such that  $\Delta_\lambda(\mu, q) < \infty$ . Let  $\nu_q$  denote the distribution of the discrete random variable  $q(X)$  and note that  $\nu_q \in \mathcal{D}$ . Since  $X \sim \mu$  and  $q(X) \sim \nu_q$

$$\begin{aligned} \Delta_\lambda(\mu, q) &= \mathbb{E} \|X - q(X)\|^2 + \lambda \mathbb{E} \ell(\alpha(X)) \\ &\geq \rho(\mu, \nu_q)^2 + \lambda L_{\nu_q} = \rho_\lambda(\mu, \nu_q)^2 \end{aligned}$$

and hence

$$\Delta_\lambda^*(\mu) = \inf_q \Delta_\lambda(\mu, q) \geq \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu, \nu)^2.$$

To show the reverse inequality, assume  $X \sim \mu$  and  $Y \sim \nu$  achieve  $\rho(\mu, \nu)$ , where  $\nu \in \mathcal{D}$  is concentrated on a countable set of points  $\{y_i; i \in \mathcal{I}_\nu\}$ . Define the variable-rate quantizer  $q$  with index set  $\mathcal{I}_\nu$  to have decoder  $\beta(i) = y_i, i \in \mathcal{I}_\nu$ , code length  $\ell_\nu$  such that

$$L_\nu = \sum_{i \in \mathcal{I}_\nu} \ell_\nu(i) \nu(y_i)$$

and encoder  $\alpha$  that is optimized for  $\beta$  and  $\ell_\nu$ , i.e.,

$$\alpha(x) = \arg \min_{i \in \mathcal{I}_\nu} (\|x - y_i\|^2 + \lambda \ell_\nu(i)).$$

Then

$$\Delta_\lambda(\mu, q) = \mathbb{E} \min_{i \in \mathcal{I}_\nu} (\|X - y_i\|^2 + \lambda \ell_\nu(i)).$$

Since  $X$  and  $Y$  achieve  $\rho(\mu, \nu)$ , and  $Y$  takes values in  $\{y_i; i \in \mathcal{I}_\nu\}$

$$\begin{aligned} \rho_\lambda(\mu, \nu)^2 &= \mathbb{E} \{\|X - Y\|^2 + \lambda L_\nu\} \\ &= \int_{\mathbb{R}^d} \sum_{i \in \mathcal{I}_\nu} (\|x - y_i\|^2 + \lambda \ell_\nu(i)) \\ &\quad \cdot \mathbb{P}\{Y = y_i | X = x\} \mu(dx) \\ &\geq \int_{\mathbb{R}^d} \min_{i \in \mathcal{I}_\nu} (\|x - y_i\|^2 + \lambda \ell_\nu(i)) \mu(dx) \\ &= \Delta_\lambda(\mu, q) \end{aligned}$$

and we obtain

$$\inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu, \nu)^2 \geq \inf_q \Delta_\lambda(\mu, q). \quad \square$$

As a consequence of the previous lemma, we obtain the following stability result for the optimal Lagrangian performance.

*Lemma 2:* For any  $\mu$  and  $\mu'$  with finite second moment

$$\left| \Delta_\lambda^*(\mu)^{1/2} - \Delta_\lambda^*(\mu')^{1/2} \right| \leq \rho(\mu, \mu').$$

*Proof:* Assume  $\Delta_\lambda^*(\mu) \geq \Delta_\lambda^*(\mu')$  without loss of generality. Fix  $\epsilon > 0$  and let  $\nu' \in \mathcal{D}$  be such that

$$\rho_\lambda(\mu', \nu') \leq \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu', \nu) + \epsilon.$$

Then by Lemma 1

$$\begin{aligned} \Delta_\lambda^*(\mu)^{1/2} - \Delta_\lambda^*(\mu')^{1/2} \\ = \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu, \nu) - \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu', \nu) \end{aligned}$$

$$\begin{aligned} &\leq \inf_{\nu \in \mathcal{D}} \rho_\lambda(\mu, \nu) - \rho_\lambda(\mu', \nu') + \epsilon \\ &\leq \rho_\lambda(\mu, \nu') - \rho_\lambda(\mu', \nu') + \epsilon \\ &= (\rho(\mu, \nu')^2 + \lambda L_{\nu'})^{1/2} - (\rho(\mu', \nu')^2 + \lambda L_{\nu'})^{1/2} + \epsilon \\ &\leq |\rho(\mu, \nu') - \rho(\mu', \nu')| + \epsilon \leq \rho(\mu, \mu') + \epsilon \end{aligned}$$

where the third inequality holds because

$$(a + c)^{1/2} - (b + c)^{1/2} \leq a^{1/2} - b^{1/2}, \quad \text{for all } a \geq b \geq 0, c \geq 0$$

by the concavity of the square root, and the last inequality follows from the triangle inequality since  $\rho$  is a metric. Since  $\epsilon > 0$  was arbitrary, we obtain  $\Delta_\lambda^*(\mu)^{1/2} - \Delta_\lambda^*(\mu')^{1/2} \leq \rho(\mu, \mu')$ .  $\square$

*Proof of Theorem 2:* Recall that  $\mu_n$  is the empirical distribution of the first  $n$  samples of a stationary and ergodic sequence with marginal distribution  $\mu$ . It is not hard to show (see, e.g., [1]) that ergodicity implies

$$\lim_{n \rightarrow \infty} \rho(\mu, \mu_n) = 0 \quad \text{a.s.} \quad (3)$$

To make use of this fact, we write

$$\Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu) = \Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu_n) + \Delta_\lambda^*(\mu_n) - \Delta_\lambda^*(\mu). \quad (4)$$

Applying Lemma 2 with  $\mu' = \mu_n$  gives

$$\left| \Delta_\lambda^*(\mu)^{1/2} - \Delta_\lambda^*(\mu_n)^{1/2} \right| \leq \rho(\mu, \mu_n)$$

and hence (3) implies that the second difference on the right-hand side of (4) converges to zero almost surely (a.s.).

To bound the first difference on the right-hand side of (4), recall that by assumption, the encoder of  $q_n^* \equiv (\alpha_n^*, \beta_n^*, \ell_n^*)$  uses the generalized nearest neighbor rule. Thus, for any  $x, y \in \mathbb{R}^d$

$$\|x - q_n^*(x)\|^2 + \lambda \ell_n^*(\alpha_n^*(x)) \leq \|x - q_n^*(y)\|^2 + \lambda \ell_n^*(\alpha_n^*(y)). \quad (5)$$

Now let  $X \sim \mu$  and  $Y \sim \mu_n$  achieve  $\rho(\mu, \mu_n)$ , and, in addition, suppose that the pair  $X, Y$  is independent of the training sequence  $X_1^n$ . Letting  $\mathbb{E}_n$  denote conditional expectation with respect to  $X_1^n$ , (5) implies

$$\begin{aligned} \Delta_\lambda(\mu, q_n^*) &= \mathbb{E}_n \{\|X - q_n^*(X)\|^2 + \lambda \ell_n^*(\alpha_n^*(X))\} \\ &\leq \mathbb{E}_n \{\|X - q_n^*(Y)\|^2 + \lambda \ell_n^*(\alpha_n^*(Y))\} \\ &\leq \mathbb{E}_n \|X - Y\|^2 + \mathbb{E}_n \{\|Y - q_n^*(Y)\|^2 + \lambda \ell_n^*(\alpha_n^*(Y))\} \\ &\quad + 2\mathbb{E}_n \{\|X - Y\| \|Y - q_n^*(Y)\|\} \\ &= \rho(\mu, \mu_n)^2 + \Delta_\lambda(\mu_n, q_n^*) \\ &\quad + 2\mathbb{E}_n \{\|X - Y\| \|Y - q_n^*(Y)\|\} \\ &\leq \rho(\mu, \mu_n)^2 + \Delta_\lambda(\mu_n, q_n^*) + 2(\mathbb{E}_n \|X - Y\|^2)^{1/2} \\ &\quad \cdot (\mathbb{E}_n \|Y - q_n^*(Y)\|^2)^{1/2} \\ &= \rho(\mu, \mu_n)^2 + \Delta_\lambda^*(\mu_n) + 2\rho(\mu, \mu_n)D(\mu_n, q_n^*)^{1/2} \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Since

$$D(\mu_n, q_n^*)^{1/2} \leq \Delta_\lambda^*(\mu_n)^{1/2} \leq \Delta_\lambda^*(\mu)^{1/2} + \rho(\mu, \mu_n)$$

by Lemma 2, we obtain

$$\begin{aligned} \Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu_n) \\ \leq \rho(\mu, \mu_n)^2 + 2\rho(\mu, \mu_n) \left( \Delta_\lambda^*(\mu)^{1/2} + \rho(\mu, \mu_n) \right). \end{aligned}$$

Since  $\lim_{n \rightarrow \infty} \rho(\mu, \mu_n) = 0$  a.s. and  $\Delta_\lambda^*(\mu, q_n^*) \geq \Delta_\lambda^*(\mu_n)$ , the first difference in (4) converges to zero a.s. as  $n \rightarrow \infty$ , completing the proof of the theorem.  $\square$

#### IV. FINITE SAMPLE PERFORMANCE

Our goal is to establish a connection between the performance of the empirically designed variable-rate quantizer and the number of training samples used in the design. We assume that the training sequence  $X_1^n = X_1, \dots, X_n$  consists of independent copies of  $X$ . In addition, we also assume that the source distribution is an element of  $\mathcal{P}(T)$ , the set of probability distributions on  $\mathbb{R}^d$  supported on  $B_T = \{x: \|x\| \leq T\}$ , the closed ball of radius  $T > 0$  centered at the origin.

The finite sample performance (rate of convergence) has been extensively studied for the empirical design problem in fixed-rate quantization [2], [7], [3], [18]. The techniques developed there will turn out to be easily adaptable to the variable-rate case once we have established a result (Lemma 3) concerning optimal variable-rate quantizers for sources with bounded support. A key to the subsequent development, this result shows that for sources with a given bounded support, variable-rate quantizers that are optimal in the Lagrangian sense cannot have too many codevectors or very large codeword lengths. The proof uses an idea of Chou and Betts [19]. The strength of the Lagrangian approach is evident here; no such general result is known for variable-rate quantizers that minimize the distortion for a given rate constraint.

For  $T > 0$  and positive integers  $N$  and  $L$ , let  $\mathcal{Q}_{N,L}(T)$  denote the collection of all variable-rate quantizers  $q \equiv (\alpha, \beta, \ell)$  with index set  $\mathcal{I}$  such that

- i)  $\|\beta(i)\| \leq T$  for all  $i \in \mathcal{I}$ ;
- ii)  $\alpha$  is a generalized nearest neighbor encoder;
- iii)  $\ell(i) \leq L$  for all  $i \in \mathcal{I}$ , and  $\mathcal{I}$  is finite with cardinality  $|\mathcal{I}| \leq N$ .

*Lemma 3:* For any  $\nu \in \mathcal{P}(T)$  and  $\lambda > 0$

$$\min_q \Delta_\lambda(\nu, q) = \min_{q \in \mathcal{Q}_{N,L}(T)} \Delta_\lambda(\nu, q) \quad (6)$$

where  $N = \lfloor 2^{5T^2/\lambda} \rfloor$  and  $L = \lfloor 5T^2/\lambda \rfloor$ . Thus, there exists  $q_n^* \in \mathcal{Q}_{N,L}(T)$ , and for this  $q_n^*$  we have

$$\Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu) \leq 2 \sup_{q \in \mathcal{Q}_{N,L}(T)} |\Delta_\lambda(\mu_n, q) - \Delta_\lambda(\mu, q)|. \quad (7)$$

*Proof:* The second statement is an easy consequence of the first one. Let  $q^*$  denote a variable-rate quantizer achieving the minimum Lagrangian distortion  $\Delta_\lambda^*(\mu)$ . Since  $\Delta_\lambda(\mu_n, q_n^*) \leq \Delta_\lambda(\mu_n, q^*)$ , we have the basic inequality

$$\begin{aligned} \Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu) &\leq \Delta_\lambda(\mu, q_n^*) - \Delta_\lambda(\mu_n, q_n^*) \\ &\quad + \Delta_\lambda(\mu_n, q^*) - \Delta_\lambda(\mu, q^*). \end{aligned}$$

Since  $\mu, \mu_n \in \mathcal{P}(T)$ , by (6) there exist  $q^*, q_n^* \in \mathcal{Q}_{N,L}(T)$ , and we obtain (7).

To prove (6), suppose  $q \equiv (\alpha, \beta, \ell)$  with index set  $\mathcal{I}$  achieves  $\Delta_\lambda^*(\nu)$ . It is easy to see that any codevector of  $q$  outside  $B_T$  can be replaced by its projection to the surface of  $B_T$  without increasing the Lagrangian distortion, so we can assume that  $\|\beta(i)\| \leq T$  for all  $i \in \mathcal{I}$ .

Next recall that  $\alpha$  is a generalized nearest neighbor encoder. Also, we can assume that for each  $S_i = \{x: \alpha(x) = i\}$ ,  $i \in \mathcal{I}$ , we have  $S_i \cap B_T \neq \emptyset$ ; otherwise, since  $\nu \in \mathcal{P}(T)$ , we can discard  $i$  from  $\mathcal{I}$  without affecting the performance. Let  $i_0 \in \mathcal{I}$  be an index with minimum codeword length, i.e.,

$$\ell(i_0) = \min_{i \in \mathcal{I}} \ell(i).$$

Since  $\alpha$  is a generalized nearest neighbor encoder, for any  $i \in \mathcal{I}$  and  $x \in S_i$

$$\|x - \beta(i)\|^2 + \lambda \ell(i) \leq \|x - \beta(i_0)\|^2 + \lambda \ell(i_0).$$

Since  $\|\beta(i_0)\| \leq T$ , we have  $\|x - \beta(i_0)\|^2 \leq 4T^2$  for all  $x \in B_T$ , and since  $S_i \cap B_T$  is nonempty, the previous inequality implies that for all  $i \in \mathcal{I}$

$$\ell(i) \leq \frac{4T^2}{\lambda} + \ell(i_0). \quad (8)$$

Now let  $q_1$  denote the quantizer with a single codepoint  $y = 0$  and rate  $r(\nu, q_1) = 0$  (formally, the single binary codeword of  $q_1$  is the empty string of length zero). Then since  $\nu \in \mathcal{P}(T)$

$$\Delta_\lambda(\nu, q_1) = D(\nu, q_1) + \lambda r(\nu, q_1) \leq T^2. \quad (9)$$

On the other hand

$$\Delta_\lambda(\nu, q) \geq \lambda r(\nu, q) \geq \lambda \ell(i_0)$$

which, together with (9) and the fact that  $\Delta_\lambda(\nu, q) \leq \Delta_\lambda(\nu, q_1)$  (since  $q$  minimizes the Lagrangian distortion for  $\nu$ ), implies that

$$\ell(i_0) \leq \frac{T^2}{\lambda}.$$

Hence, by (8), we have for all  $i \in \mathcal{I}$

$$\ell(i) \leq \frac{5T^2}{\lambda}.$$

Now Kraft's inequality implies

$$1 \geq \sum_{i \in \mathcal{I}} 2^{-\ell(i)} \geq |\mathcal{I}| 2^{-5T^2/\lambda}$$

and hence  $|\mathcal{I}| \leq 2^{5T^2/\lambda}$ . Setting  $N = \lfloor 2^{5T^2/\lambda} \rfloor$  and  $L = \lfloor 5T^2/\lambda \rfloor$ , we obtain  $q \in \mathcal{Q}_{N,L}(T)$ , which completes the proof.  $\square$

In view of Lemma 3, one can now use established techniques of empirical processes and statistical learning theory to bound the Lagrangian performance of an empirically optimal variable-rate quantizer for a finite amount of training data. In the following theorem, we assume (as we may by Lemma 3) that  $q_n^* \in \mathcal{Q}_{N,L}(T)$ .

*Theorem 3:* There is a constant  $C$ , depending only on  $d, \lambda$ , and  $T$ , such that for all  $n \geq 1$  and  $\mu \in \mathcal{P}(T)$

$$\mathbb{E} \Delta_\lambda(\mu, q_n^*) - \Delta_\lambda^*(\mu) \leq \frac{C}{\sqrt{n}}.$$

The proof is an application of the technique that was used in [2] to prove a similar result for fixed-rate quantizers. The technique is based on Vapnik–Chervonenkis theory and, due to Lemma 3, is readily adaptable to our case. For the sake of completeness, we give a detailed proof.

*Proof:* In what follows we will use notions and results of Vapnik–Chervonenkis theory given, for example, in [20]–[22].

By bound (7) of Lemma 3, it suffices to give an appropriate upper bound on the expected value of

$$\sup_{q \in \mathcal{Q}_{N,L}(T)} |\Delta_\lambda(\mu_n, q) - \Delta_\lambda(\mu, q)|$$

where  $N = \lfloor 2^{5T^2/\lambda} \rfloor$  and  $L = \lfloor 5T^2/\lambda \rfloor$ . For

$$q \equiv (\alpha, \beta, \ell) \in \mathcal{Q}_{N,L}(T)$$

define the distortion function

$$f_{\lambda,q}(x) \triangleq \|x - q(x)\|^2 + \lambda \ell(\alpha(x)).$$

Then for all  $x \in B_T$

$$0 \leq f_{\lambda,q}(x) \leq 4T^2 + \lambda L = 4T^2 + \lambda \lfloor 5T^2/\lambda \rfloor \leq 9T^2.$$

Letting  $I_A$  denote the indicator of the event  $A$ , we have

$$\Delta_\lambda(\mu, q) = \mathbf{E}f_{\lambda, q}(X) = \int_0^{9T^2} \mathbf{P}\{f_{\lambda, q}(X) > u\} du$$

and

$$\begin{aligned} \Delta_\lambda(\mu_n, q) &= \frac{1}{n} \sum_{k=1}^n f_{\lambda, q}(X_k) \\ &= \int_0^{9T^2} \frac{1}{n} \sum_{k=1}^n I_{\{f_{\lambda, q}(X_k) > u\}} du \quad \text{a.s.} \end{aligned}$$

Hence,

$$\begin{aligned} &\sup_{q \in \mathcal{Q}_{N, L}(T)} |\Delta_\lambda(\mu_n, q) - \Delta_\lambda(\mu, q)| \\ &= \sup_{q \in \mathcal{Q}_{N, L}(T)} \left| \frac{1}{n} \sum_{k=1}^n f_{\lambda, q}(X_k) - \mathbf{E}f_{\lambda, q}(X) \right| \\ &= \sup_{q \in \mathcal{Q}_{N, L}(T)} \left| \int_0^{9T^2} \left( \frac{1}{n} \sum_{k=1}^n I_{\{f_{\lambda, q}(X_k) > u\}} - \mathbf{P}\{f_{\lambda, q}(X) > u\} \right) du \right| \\ &\leq 9T^2 \sup_{q \in \mathcal{Q}_{N, L}(T), u > 0} \left| \frac{1}{n} \sum_{k=1}^n I_{\{f_{\lambda, q}(X_k) > u\}} - \mathbf{P}\{f_{\lambda, q}(X) > u\} \right| \\ &= 9T^2 \sup_{A \in \mathcal{A}_N} |\mu_n(A) - \mu(A)| \quad \text{a.s.} \end{aligned} \quad (10)$$

where  $\mathcal{A}_N$  is the family of sets in  $\mathbb{R}^d$  defined by

$$\mathcal{A}_N \triangleq \{\{x: f_{\lambda, q}(x) > u\}; q \in \mathcal{Q}_{N, L}(T), u > 0\}. \quad (11)$$

For any class  $\mathcal{B}$  of subsets of  $\mathbb{R}^d$  and positive integer  $k$ , the *shatter coefficient*  $\mathfrak{S}_{\mathcal{B}}(k)$  is defined as the maximal number of different subsets of a set of  $k$  points in  $\mathbb{R}^d$  which can be obtained by intersecting it with elements of  $\mathcal{B}$ , that is,

$$\mathfrak{S}_{\mathcal{B}}(k) = \max_{x_1, \dots, x_k \in \mathbb{R}^d} |\{\{x_1, \dots, x_k\} \cap B: B \in \mathcal{B}\}|.$$

The Vapnik–Chervonenkis dimension  $V(\mathcal{B})$  of  $\mathcal{B}$  is the largest integer  $k$  such that

$$\mathfrak{S}_{\mathcal{B}}(k) = 2^k$$

(if  $\mathfrak{S}_{\mathcal{B}}(k) = 2^k$  for all  $k$ , then  $V(\mathcal{B}) = \infty$  by definition).

Our main tool is a sharpened version of the classical Vapnik–Chervonenkis inequality [20, Sec. 4.3] which states that if  $\mu_n$  is the empirical distribution of  $n$  independent and identically distributed (i.i.d.) samples, then for any class  $\mathcal{B}$  of measurable subsets of  $\mathbb{R}^d$

$$\mathbf{E} \left\{ \sup_{B \in \mathcal{B}} |\mu_n(B) - \mu(B)| \right\} \leq c \sqrt{\frac{V(\mathcal{B})}{n}}$$

where  $c$  is a universal constant. Applying this to the family of sets  $\mathcal{A}_N$  defined in (11) and combining with the bound of (10) gives

$$\mathbf{E} \left\{ \sup_{q \in \mathcal{Q}_{N, L}(T)} |\Delta_\lambda(\mu_n, q) - \Delta_\lambda(\mu, q)| \right\} \leq 9T^2 c \sqrt{\frac{V(\mathcal{A}_N)}{n}}. \quad (12)$$

This bound yields the theorem if we can show that  $V(\mathcal{A}_N) < \infty$ .

Since the encoder of each  $q \in \mathcal{Q}_{N, L}(T)$  uses the generalized nearest neighbor rule,  $f_{\lambda, q}(x) > u$  if and only if

$$\|x - \beta(i)\|^2 + \lambda \ell(i) > u \quad \text{for all } i \in \mathcal{I}$$

i.e.,

$$\{x: f_{\lambda, q}(x) > u\} = \bigcap_{i \in \mathcal{I}} \{x: \|x - \beta(i)\|^2 > u - \lambda \ell(i)\}.$$

Since  $|\mathcal{I}| \leq N$ , we obtain that either  $\{x: f_{\lambda, q}(x) > u\}$  is an intersection of the complements of at most  $N$  closed balls of possibly different radii in  $\mathbb{R}^d$  (if  $u - \lambda \ell(i) \geq 0$  for some  $i$ ), or  $\{x: f_{\lambda, q}(x) > u\} = \mathbb{R}^d$  (this is the case if  $u - \lambda \ell(i) < 0$  for all  $i$ ). Thus, if  $\bar{\mathcal{A}}_N$  denotes the family of all intersections of complements of  $N$  closed balls in  $\mathbb{R}^d$ , then we have

$$\mathcal{A}_N \subset \bar{\mathcal{A}}_N \cup \{\mathbb{R}^d\}.$$

It easily follows from the definition of shatter coefficients that  $\mathfrak{S}_{\bar{\mathcal{A}}_N \cup \{\mathbb{R}^d\}}(k) = \mathfrak{S}_{\bar{\mathcal{A}}_N}(k)$  for all  $k$ . Hence,  $\mathfrak{S}_{\mathcal{A}_N}(k) \leq \mathfrak{S}_{\bar{\mathcal{A}}_N}(k)$ , which in turn implies

$$V(\mathcal{A}_N) \leq V(\bar{\mathcal{A}}_N).$$

A standard argument in Lemma 4 in the Appendix shows that  $V(\bar{\mathcal{A}}_N)$  is bounded as

$$V(\bar{\mathcal{A}}_N) \leq 4N(d+1) \ln(N(d+1)).$$

Combining this with (12) and the bound (7) of Lemma 3 we obtain the theorem with

$$C = 18T^2 c \sqrt{4N(d+1) \ln(N(d+1))}$$

where  $N = \lfloor 2^{5T^2/\lambda} \rfloor$ .  $\square$

## V. CONNECTION WITH ENTROPY-CONSTRAINED QUANTIZATION

In variable-rate quantization, one can alternatively define the rate of a quantizer by the entropy of the quantizer output [10], [12]. This definition often simplifies matters since it does not tie the rate to a particular variable-length lossless code. The entropy-constrained rate of the vector quantizer with index set  $\mathcal{I}$ , encoder  $\alpha$ , and decoder  $\beta$  is

$$H(q(X)) = - \sum_{i \in \mathcal{I}} \mathbf{P}\{\alpha(X) = i\} \log_2 \mathbf{P}\{\alpha(X) = i\}.$$

Our formulation of variable-rate quantization needs only a slight modification to accommodate this rate definition. If we remove the restriction that the code-length function  $\ell$  has to take integer values, and only require that  $\ell(i) \geq 0$  for all  $i \in \mathcal{I}$  and  $\sum_{i \in \mathcal{I}} 2^{-\ell(i)} \leq 1$ , then the expected code length of a quantizer  $q \equiv (\alpha, \beta, \ell)$

$$r(\mu, q) = \sum_{i \in \mathcal{I}} \ell(i) \mathbf{P}\{\alpha(X) = i\}$$

is uniquely minimized by the length function

$$\ell(i) = -\log_2 \mathbf{P}\{\alpha(X) = i\}, \quad i \in \mathcal{I}.$$

With these optimal (noninteger) code lengths, the quantizer rate becomes  $r(\mu, q) = H(q(X))$ , the entropy of the quantizer output.

It is a simple matter to check that the proofs of the results in Sections III and IV carry over without change if the restriction that  $\ell$  is integer valued is removed, leading to analogous results for entropy-constrained quantization. In particular, the entropy-constrained counterparts of Theorems 2 and 3 show the consistency and bound the finite sample performance, respectively, of the empirical Lagrangian design of entropy-constrained vector quantizers.

## APPENDIX

*Lemma 4:* Let  $\mathcal{A}$  denote the collection of all complements of closed balls of nonnegative radius in  $\mathbb{R}^d$ , and for any positive integer  $N$  define

$$\bar{\mathcal{A}}_N \triangleq \{A_1 \cap \dots \cap A_N: A_i \in \mathcal{A}, i = 1, \dots, N\}. \quad (13)$$

Then the Vapnik–Chervonenkis dimension of  $\bar{\mathcal{A}}_N$  is upper-bounded as

$$V(\bar{\mathcal{A}}_N) \leq 4N(d+1)\ln(N(d+1)).$$

*Proof:* We have  $\bar{\mathcal{A}}_1 = \mathcal{A}$ , and

$$\bar{\mathcal{A}}_N = \{A \cap B : A \in \mathcal{A}, B \in \bar{\mathcal{A}}_{N-1}\}, \quad \text{for } N \geq 2.$$

A well-known property of shatter coefficients (see, e.g., [20]) implies that for  $N \geq 2$  and  $k \geq 1$

$$\mathfrak{S}_{\bar{\mathcal{A}}_N}(k) \leq \mathfrak{S}_{\mathcal{A}}(k)\mathfrak{S}_{\bar{\mathcal{A}}_{N-1}}(k).$$

Thus,  $\mathfrak{S}_{\bar{\mathcal{A}}_N}(k) \leq \mathfrak{S}_{\mathcal{A}}(k)^N$  by induction. Define  $\mathcal{D} = \{A^c : A \in \mathcal{A}\}$ . It follows immediately from the definition that  $\mathfrak{S}_{\mathcal{A}}(k) = \mathfrak{S}_{\mathcal{D}}(k)$ . Hence, we obtain

$$\mathfrak{S}_{\bar{\mathcal{A}}_N}(k) \leq \mathfrak{S}_{\mathcal{D}}(k)^N. \quad (14)$$

Since  $\mathcal{D}$  is the collection of all closed balls in  $\mathbb{R}^d$ , we have  $V(\mathcal{D}) = d+1$  by a result of Dudley [23]. Next, we use a well-known consequence of Sauer’s lemma which states that for any class of sets  $\mathcal{B}$  and all integers  $k \geq V(\mathcal{B})$

$$\mathfrak{S}_{\mathcal{B}}(k) \leq \left(\frac{ke}{V(\mathcal{B})}\right)^{V(\mathcal{B})}$$

(see, e.g., [20, Corollary 4.1]). This and (14) imply that for all  $k \geq d+1$

$$\mathfrak{S}_{\bar{\mathcal{A}}_N}(k) \leq \left(\frac{ke}{V(\mathcal{D})}\right)^{NV(\mathcal{D})} = \left(\frac{ke}{d+1}\right)^{N(d+1)}. \quad (15)$$

An upper bound to  $V(\bar{\mathcal{A}}_N)$  can now be obtained by finding a  $k$  for which the right-hand side is less than  $2^k$ . It is easy to check that if  $d \geq 2$ , then  $k = 4N(d+1)\ln(N(d+1))$  satisfies this requirement. Since for  $d = 1$  we obviously have  $V(\bar{\mathcal{A}}_N) \leq 2N$ , we obtain that for all  $N, d \geq 1$ ,

$$V(\bar{\mathcal{A}}_N) \leq 4N(d+1)\ln(N(d+1)). \quad \square$$

REFERENCES

[1] D. Pollard, “Quantization and the method of  $k$ -means,” *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 199–205, Mar. 1982.  
 [2] T. Linder, G. Lugosi, and K. Zeger, “Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 1728–1740, Nov. 1994.  
 [3] P. Bartlett, T. Linder, and G. Lugosi, “The minimax distortion redundancy in empirical quantizer design,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 1802–1813, Sept. 1998.  
 [4] D. Pollard, “Strong consistency of  $k$ -means clustering,” *Ann. Statist.*, vol. 9, no. 1, pp. 135–140, 1981.  
 [5] —, “A central limit theorem for  $k$ -means clustering,” *Ann. Probab.*, vol. 10, no. 4, pp. 919–926, 1982.  
 [6] E. A. Abaya and G. L. Wise, “On the existence of optimal quantizers,” *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 937–940, Nov. 1982.  
 [7] A. J. Zeevi, “On the performance of vector quantizers empirically designed from dependent sources,” in *Proc. Data Compression Conf. (DCC’98)*, J. Storer and M. Cohn, Eds. Los Alamitos, CA: IEEE Computer Society Press, 1998, pp. 73–82.  
 [8] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*. Berlin, Heidelberg, Germany: Springer-Verlag, 2000.  
 [9] A. György and T. Linder, “On the structure of optimal entropy-constrained scalar quantizers,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 416–427, Feb. 2002.  
 [10] P. A. Chou, T. Lookabaugh, and R. M. Gray, “Entropy-constrained vector quantization,” *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, pp. 31–42, Jan. 1989.  
 [11] R. M. Gray, T. Linder, and J. Li, “A Lagrangian formulation of Zador’s entropy-constrained quantization theorem,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 695–707, Mar. 2002.

[12] R. M. Gray and D. L. Neuhoff, “Quantization,” *IEEE Trans. Inform. Theory*, (Special Commemorative Issue), vol. 44, pp. 2325–2383, Oct. 1998.  
 [13] T. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.  
 [14] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1992.  
 [15] A. György and T. Linder, “A note on the existence of optimal entropy-constrained vector quantizers,” in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, 2002.  
 [16] P. Papantoni-Kazakos and R. M. Gray, “Robustness of estimators on stationary observations,” *Ann. Probab.*, vol. 7, pp. 989–1002, 1979.  
 [17] T. Linder, V. Tarokh, and K. Zeger, “Existence of optimal prefix codes for infinite source alphabets,” *IEEE Trans. Inform. Theory*, vol. 43, pp. 2026–2028, Nov. 1997.  
 [18] T. Linder, “On the training distortion of vector quantizers,” *IEEE Trans. Inform. Theory*, vol. 46, pp. 1617–1623, July 2000.  
 [19] P. A. Chou and B. J. Betts, “When optimal entropy-constrained quantizers have only a finite number of codewords,” in *Proc. IEEE Int. Symp. Information Theory*, Cambridge, MA, Aug. 16–21, 1998, p. 97.  
 [20] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. New York: Springer-Verlag, 2000.  
 [21] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.  
 [22] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 1999.  
 [23] R. Dudley, “Balls in  $R^k$  do not cut all subsets of  $k+2$  points,” *Adv. Math.*, vol. 31, no. 3, pp. 306–308, 1979.

New Bounds for the Marcum  $Q$ -Function

Giovanni E. Corazza, *Member, IEEE*, and Gianluigi Ferrari, *Student Member, IEEE*

**Abstract**—New bounds are proposed for the Marcum  $Q$ -function, which is defined by an integral expression where the 0th-order modified Bessel function appears. The proposed bounds are derived by suitable approximations of the 0th-order modified Bessel function in the integration region of the Marcum  $Q$ -function. They prove to be very tight and outperform bounds previously proposed in the literature. In particular, the proposed bounds are noticeably good for large values of the parameters of the Marcum  $Q$ -function, where previously introduced bounds fail and where exact computation of the function becomes critical due to numerical problems.

**Index Terms**—Marcum  $Q$ -function, modified Bessel function of the first kind, upper and lower bounds.

I. INTRODUCTION

Calculation of the generalized Marcum  $Q$ -function of order  $M$ , usually referred to as  $Q_M(a, b)$ , and particularly the popular case ( $M = 1$ ) indicated as Marcum  $Q$ -function  $Q(a, b)$ , is important in many problems of signal detection [1], [2]. Immediate examples are the computation of error probability in transmission over fading channels or detection probability for code acquisition in a direct-sequence code-division multiple-access (DS-CDMA) system. Several algorithms have

Manuscript received September 5, 2001; revised May 6, 2002.  
 G. E. Corazza is with the Department of Electronics, Computer Science and Systems (D.E.I.S.), University of Bologna, 40136 Bologna, Italy (e-mail: gecorazza@deis.unibo.it).  
 G. Ferrari is with the Department of Information Engineering, University of Parma, 43100 Parma, Italy (e-mail: gferrari@tlc.unipr.it).  
 Communicated by A. Kavčić, Associate Editor for Detection and Estimation. Digital Object Identifier 10.1109/TIT.2002.804113.