EMPIRICAL DIVERGENCE MAXIMIZATION FOR QUANTIZER DESIGN: AN ANALYSIS OF APPROXIMATION ERROR

Michael A. Lexa

Institute for Digital Communications
The University of Edinburgh
Edinburgh, UK

ABSTRACT

Empirical divergence maximization is an estimation method similar to empirical risk minimization whereby the Kullback-Leibler divergence is maximized over a class of functions that induce probability distributions. We use this method as a design strategy for quantizers whose output will ultimately be used to make a decision about the quantizer's input. We derive this estimator's approximation error decay rate as a function of the resolution of a class of partitions known as recursive dyadic partitions. This result, coupled with earlier results, show that this estimator can converge to the theoretically optimal solution as fast as n^{-1}, where n is the number of training samples. This estimator also is capable of producing estimates that well-approximate optimal solutions that existing techniques cannot.

Index Terms— empirical quantizer design, empirical divergence maximization, Kullback-Leibler divergence, recursive dyadic partitions

1. INTRODUCTION

This paper extends the analysis of an empirical quantization design methodology proposed in [1] that is based on empirical divergence maximization. By leveraging recent results in statistical learning theory, this approach shows that fast convergence rates for these estimators are possible, in addition to explicitly showing the parameters on which these rates depend. Under the assumption that the underlying probability distributions are unknown, commonly used empirical risk estimators seek to find classifiers that minimize the empirical risk (empirical probability of error) by searching among a pre-specified class of candidate classifiers [2]. We adopt this approach here, but seek to maximize an empirical form of the Kullback-Leibler (KL) divergence over a given class of quantization rules. The KL divergence is a well-known quantity related to optimal detector performance [3]. Hence, the presumption is that the quantized samples will ultimately be used to make a decision about the quantizer's input signal.

Let P and Q be two probability measures defined on $[0,1]^d$ and let p and q denote their respective density functions, which we assume to be uniformly bounded, i.e., $c \leq p(\mathbf{x}), q(\mathbf{x}) \leq C$ for all $\mathbf{x} \in [0,1]^d$, $c > 0$, $C < \infty$. Any quantization rule $\phi : \mathbb{R}^d \mapsto \{0,\ldots,L-1\}$ that operates on a random vector $\mathbf{X}$ (distributed according to P or Q) induces the probability mass functions (pmfs), $p(\phi) = (p_0(\phi),\ldots,p_{L-1}(\phi))$ and $q(\phi) = (q_0(\phi),\ldots,q_{L-1}(\phi))$, where $p_i(\phi) = P(\phi(\mathbf{X}) = i)$ and similarly for $q_i(\phi)$. In this context, the KL divergence is defined as

$$D_{KL}(p(\phi),q(\phi)) = \sum_{i=0}^{L-1} -p_i(\phi)\log\left(\frac{q_i(\phi)}{p_i(\phi)}\right).$$

Stein's Lemma [3] states that the KL divergence equals an optimal detector's exponential error decay rate; thus, by constructing quantization rules $\widehat{\phi}_n$ that induce maximally divergent pmfs, we, in some sense, ensure the best possible performance of a follow-on detector.

To be clear, this problem concerns two types of rates, one which we deal with explicitly, the other implicitly. The rate with which the "best in class" estimate converges to the theoretically optimal quantization rule is explicitly analyzed. The other error rate associated with the KL divergence through Stein's Lemma characterizes a detector's performance *after* a quantization rule is designed and is in use.

We therefore analyze an estimator of the form

$$\widehat{\phi}_n = \arg\max_{\phi\in\Phi} D_n(\phi),$$

where Φ and $D_n(\phi)$ represent an, as yet unspecified, candidate class and empirical KL divergence estimate, respectively. In this paper, we examine the decay rate of the so-called approximation error associated with $\widehat{\phi}_n$; the estimation error is examined in [1].

The empirical nature of the strategy makes it an attractive choice for newly proposed continuous-time compressed sensing (CS) sampling schemes [4, 5]. For signals that have a sparse representation, that is for signals that can be represented, or well approximated, by a small number of basis vectors, CS suggests that it is possible to recover these signals with sampling rates far below Nyquist rates. While the quantization strategy proposed here does not depend on any notion of signal sparsity, these new schemes often process the continuous-time waveforms prior to sampling. Consider, for example, the random demodulator proposed in [4] that multiplies the signal by a *random* waveform taking values ± 1. In this case, even if an accurate probability model exists for the continuous-time random process, it may be difficult to propagate the model through the sampling process. Fast convergence is also advantageous for continuous-time CS sampling schemes because a primary goal of these systems is to sample at rates as slow as possible (sub-Nyquist). Thus, strategies that quickly converge hold significant advantage over strategies with slower convergence.

This paper also addresses a long-standing shortcoming in quantization for classification problems. It is well-known that the theoretically optimal quantization rule (assuming p and q are known) can always be constructed by thresholding the likelihood ratio [6]. Put differently, the quantization rule that maximizes the KL divergence of the pmfs it induces, can always be chosen to be a piecewise constant function defined on a partition whose boundary sets are level sets of the likelihood ratio $q(\mathbf{x})/p(\mathbf{x})$. These optimal likelihood ratio partitions can be very different from typical nearest neighbor regions/partitions that are associated with quantizers designed

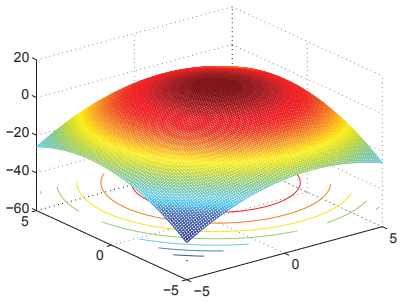

Fig. 1. Plot of a log-likelihood ratio function formed by dividing a bivariate Gaussian probability density function (pdf) by a bivariate Laplace pdf. The elliptical contours indicate that, for any number of quantization levels, the optimal likelihood ratio partition will consist of concentric ellipses. Such partitions are not nearest-neighbor partitions, nor can they be, generally speaking, well-approximated by nearest-neighbor partitions.

to minimize mean-squared error. Past related work by Gupta and Hero [7] and Lazebnik and Raginsky [8] forced a small-cell property in their design strategy. The resultant partitions thus resembled nearest-neighbor partitions. Consequently, their design strategies disallow close approximation to some optimal likelihood ratio partitions and/or quantization rules, for instance, partitions resembling those in Figure 1 or quantization rules with disjoint regions. The approach taken here largely alleviates this shortcoming.

In [8], Lazebnik and Raginsky study a conceptually similar quantization problem to the one considered here, but the differences between the approaches are substantial. For example, their information loss criterion is a difference of mutual informations, and while related to the KL divergence, this criterion measures a different quantity than the approximation loss studied here.

2. EMPIRICAL DIVERGENCE ESTIMATION

The form of $D_n(\phi)$ is taken from recent work by Nguyen et al. [9] and relies on rewriting the convex function $-\log(\cdot)$ appearing in the definition of the KL divergence.

Convex conjugates. The notion of a convex conjugate is based on the observation that a function can be described as either by its graph or by an envelope of tangents curves [10]. Formally, one can rewrite any (closed) convex function f such that for any point $x \in \mathbb{R}$, $f(x)$ is a supremum over a set of affine functions,

$$f(x) = \sup_{x^*} \{x^* x - f^*(x^*)\}, \tag{1}$$

where $f^*(x^*)$ is the convex conjugate of $f(x)$.

Now, suppose $\gamma : [0,1]^d \mapsto \{0, \ldots, L-1\}$ is an arbitrary quantization rule characterized by the partitioning sets $\{R_i\}_{i=0}^{L-1}$. Using (1), we write the divergence between the pmfs induced by γ as

$$D_{KL}(p(\gamma), q(\gamma)) = \sum_{i=0}^{L-1} p_i(\gamma) f\left(\frac{q_i(\gamma)}{p_i(\gamma)}\right) \tag{2a}$$

$$= \sum_{i=0}^{L-1} p_i(\gamma) \cdot \sup_{x^*} \Big\{ x^* \frac{q_i(\gamma)}{p_i(\gamma)} - f^*(x^*) \Big\}, \tag{2b}$$

where $f(x) = -\log(x), x > 0, +\infty$, otherwise. For this particular convex function, $f^*(x^*)$ equals

$$f^*(x^*) = \begin{cases} -1 - \log(-x^*) & \text{if } x^* < 0 \\ +\infty & \text{if } x^* \geq 0. \end{cases}$$

Substituting this expression into (2b), we have the following expressions for the KL divergence

$$\begin{aligned} &D_{KL}(p(\gamma), q(\gamma)) \\ &= \sum_{i=0}^{L-1} p_i(\gamma) \sup_{x_i^* \in \mathbb{R}^-} \Big\{ x_i^* \frac{q_i(\gamma)}{p_i(\gamma)} + 1 + \log(-x_i^*) \Big\} \\ &= \sum_{i=0}^{L-1} p_i(\gamma) \sup_{c_{R_i} \in \mathbb{R}^+} \Big\{ \log(c_{R_i}) - c_{R_i} \frac{q_i(\gamma)}{p_i(\gamma)} + 1 \Big\} \\ &= 1 + \sum_{i=0}^{L-1} \sup_{c_{R_i} \in \mathbb{R}^+} \Big\{ P(R_i) \log(c_{R_i}) - c_{R_i}\, Q(R_i) \Big\}, \end{aligned}$$

where in the second step we let $c_{R_i} = -x_i^*$, and in the last step used the fact that $p_i(\gamma) = P(R_i)$ with $R_i = \{\mathbf{x} : \gamma(\mathbf{X}) = i\}$. The validity of last expression is easily verified by differentiating it with respect to c_{R_i} and solving for the maximizers. By defining the piecewise constant function

$$\phi(\mathbf{x}) = \sum_{i=0}^{L-1} c_{R_i} \mathbf{1}_{R_i}(\mathbf{x}), \quad c_{R_i} \in \mathbb{R}^+, \tag{3}$$

we can write $D_{KL}(p(\gamma), q(\gamma))$ in integral form:

$$1 + \sup_{\phi} \left\{ \int_{[0,1]^d} \log(\phi)\, dP - \int_{[0,1]^d} \phi\, dQ \right\}, \tag{4}$$

where $\mathbf{1}$ denotes the indicator function, and the supremum is taken over all functions of the form (3).

Empirical estimator. Let $\{R_i\}_{i=0}^{L-1}$ be a generic partition of $[0,1]^d$. Then for positive constants $m > 0$ and $M < \infty$, we define the candidate function class

$$\Phi(m, M, L, R_i) = \left\{ \phi(\mathbf{x}) = \sum_{i=0}^{L-1} c_{R_i} \mathbf{1}_{R_i}(\mathbf{x}) : m \leq c_{R_i} \leq M \right\}.$$

Φ is a set of piecewise constant functions with L levels that are bounded and positive. In Section 3, we consider a specific class of partitions that admits fast convergence rates.

Define the function $D_n(\phi)$ as the empirical counterpart of (4),

$$D_n(\phi) = 1 + \frac{1}{n} \sum_{i=1}^{n} \log \phi(X_i^p) - \frac{1}{n} \sum_{i=1}^{n} \phi(X_i^q), \tag{5}$$

and consider the following empirical estimator

$$\widehat{\phi}_n = \underset{\phi \in \Phi(m,M,L)}{\arg\max}\ D_n(\phi), \tag{6}$$

where $\{X_i^p\}_{i=1}^n$ and $\{X_i^q\}_{i=1}^n$ in (5) are observations distributed according to p and q respectively. $\widehat{\phi}_n$ is an *empirical divergence maximization* estimator akin to the familiar empirical risk minimization estimators [2]. Note $D_n(\phi)$ is not in general a KL divergence; it can in fact be negative for some $\phi \in \Phi$. It is a consistent estimator, however, converging to the "best in class" estimator as $n \to \infty$ [9].

The best in class estimate ϕ^* is that element in Φ that maximizes $D(\phi)$,

$$\phi^* = \underset{\phi \in \Phi(m,M,L)}{\arg\max} \; D(\phi), \quad \text{where} \tag{7}$$
$$D(\phi) = 1 + \int_{[0,1]^d} \log(\phi)\, dP - \int_{[0,1]^d} \phi \, dQ.$$

Note that $D(\phi)$, as opposed to $D_n(\phi)$, is not an empirical quantity, but uses knowledge of the distributions P and Q.

We take the theoretically optimal quantization rule γ^* to be the rule that maximizes the divergence over a broad class of piecewise constant functions whose only restriction, essentially, is an assumed boundary regularity. To define this class, we need the following concept: a function $f : [0,1]^d \mapsto \mathbb{R}$ is *locally constant* at a point $x \in [0,1]^d$ if there exists a ball about x with positive radius in which f is constant.

Definition (PC class)**.** *A function $f : [0,1]^d \mapsto \{c_i\}_{i=0}^{L-1}$, $c_i \in \mathbb{R}^+$ is a positive-valued piecewise constant function with L levels if it is locally constant at any point $\mathbf{x} \in [0,1,]^d \setminus B(f)$, where $B(f) \subset [0,1]^d$ is a boundary set satisfying $N(r) \leq \beta r^{-(d-1)}$ for all $r > 0$. Here, $\beta > 0$ is a constant and $N(r)$ is the minimal number of balls of diameter r that covers $B(f)$. Furthermore, let f be uniformly bounded on $[0,1]^d$, that is $m \leq f(\mathbf{x}) \leq M$ for all $\mathbf{x} \in [0,1]^d$, where $m > 0$ and $M < \infty$. The set of all piecewise constant functions f satisfying the above conditions is denoted by* $\mathrm{PC}(\beta, m, M, L)$.

We consider $\mathrm{PC}(\beta, m, M, L)$ to be a class of likelihood ratio quantization rules that have well-behaved boundaries [11]. Alternately, we say that we only consider densities p and q whose likelihood ratio function has well-behaved level sets.

Define the theoretically optimal quantization rule as

$$\gamma^* = \underset{\gamma \in \mathrm{PC}(\beta,m,M,L)}{\arg\max} \; D(\gamma). \tag{8}$$

Let $\{A_i^*\}_{i=0}^{L-1}$ denote the partition associated with γ^*, then the L levels (constant values) of γ^* equal $P(A_i^*)/Q(A_i^*), i = 0, \ldots, L-1$, and γ^* takes the form

$$\gamma^*(\mathbf{x}) = \sum_{i=0}^{L-1} c_{A_i^*} \mathbf{1}_{A_i^*}(\mathbf{x}), \quad c_{A_i^*} = \frac{P(A_i^*)}{Q(A_i^*)}. \tag{9}$$

Estimation and Approximation Errors. We gauge the quality of $\widehat{\phi}_n$ by characterizing the decay rates of the so-called estimation and approximation errors. *Estimation* error is defined as the difference $D(\phi^*) - D(\widehat{\phi}_n)$ and quantifies the error caused by computing $\widehat{\phi}_n$ without knowledge of p and q. *Approximation error* is defined as $D(\gamma^*) - D(\phi^*)$ and arises from differences in the partitions of γ^* and ϕ^*.

Investigations of these errors typically begin with two basic inequalities that follow from the definitions of ϕ^* and $\widehat{\phi}_n$: $D(\phi^*) - D(\widehat{\phi}_n) \geq 0$ and $D_n(\phi^*) - D_n(\widehat{\phi}_n) \leq 0$. They imply that the estimation error is upper bounded by a difference of empirical processes

$$\begin{aligned} 0 &\leq D(\phi^*) - D(\widehat{\phi}_n) \\ &\leq -[(D_n(\phi^*) - D(\phi^*)) - (D_n(\widehat{\phi}_n) - D(\widehat{\phi}_n))] \\ &= -(\nu_n(\phi^*) - \nu_n(\widehat{\phi}_n))/\sqrt{n}, \end{aligned}$$

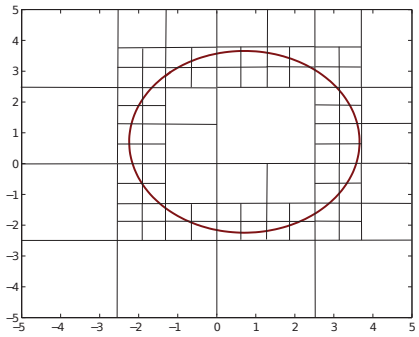

Fig. 2. An example RDP adapted to a level set of the graph in Fig. 1 ($J = 3, d = 2$). Here, all cells intersecting the level set have maximal depth $J = 3$.

where the second inequality results from adding and subtracting $D_n(\phi^*)$ and $D_n(\widehat{\phi}_n)$, and $\nu_n(\gamma) = \sqrt{n}(D_n(\gamma) - D(\gamma))$. By adding the approximation error to both sides of the above inequality, we have that the total error is bounded by the two component errors.

$$\begin{aligned} 0 &\leq \underbrace{D(\gamma^*) - D(\widehat{\phi}_n)}_{\text{total error}} \\ &\leq \underbrace{-(\nu_n(\phi^*) - \nu_n(\widehat{\phi}_n))/\sqrt{n}}_{\text{upper bound on est. error}} + \underbrace{D(\gamma^*) - D(\phi^*)}_{\text{approx. error}}. \end{aligned} \tag{10}$$

We analyze the decay rate of the approximation error below.

3. APPROXIMATION ERROR

Recursive Dyadic Partitions. To characterize the approximation error, we consider a particular class of partitions $\{R_i\}_{i=0}^{L-1}$ derived from underlying *Recursive Dyadic Partitions* (RDPs). RDPs are partitions composed of quasi-disjoint sets (two sets are quasi-disjoint if and only if their intersection has Lebesque measure zero) whose union equals the entire space $[0,1]^d$. We use RDPs because of their proven effectiveness in adapting to the boundaries of piecewise constant functions [11]. A RDP is any partition that can be constructed using only the following rules:

1. $\{[0,1]^d\}$ is a RDP.
2. Let $\pi = \{S_0, \ldots, S_{k-1}\}$ be a RDP, where $S_i = [a_{i1}, b_{i1}] \times \ldots \times [a_{id}, b_{id}]$. Then $\pi' = \{S_1, \ldots, S_{i-1}, S_i^{(0)}, \ldots, S_i^{(2^d-1)}, S_{i+1}, \ldots, S_k\}$ is a RDP, where $\{S_i^{(0)}, \ldots, S_i^{(2^d-1)}\}$ is obtained by dividing the hypercube S_i into 2^d quasi-disjoint hypercubes of equal size. Formally, let $q_1 q_2 \ldots q_d$ be the binary representation of $q \in \{0, \ldots, 2^{d-1}\}$. Then

$$\begin{aligned} S_i^{(q)} = &\left[a_{i1} + \frac{b_{i1} - a_{i1}}{2} q_1, b_{i1} + \frac{a_{i1} - b_{i1}}{2}(1 - q_1)\right] \times \\ &\ldots \times \left[a_{id} + \frac{b_{id} - a_{id}}{2} q_d, b_{id} + \frac{a_{id} - b_{id}}{2}(1 - q_d)\right]. \end{aligned}$$

Figure 2 illustrates an example RDP. We say a RDP has maximal depth J if the side length of its smallest hypercube equals 2^{-J}.

We now further specify Φ as the class of quantization rules whose partitioning cells R_i are unions of RDP cells of a fixed

maximal depth, i.e. for any $\phi \in \Phi$, we restrict R_i to have the form $R_i = \bigcup_{m \in I_i} S_m$, where the cells S_m belong to a RDP of fixed maximal depth, $\{I_i\}$ is a set of disjoint index sets and $\cup_i Ri = [0,1]^d$.

Main Result. *Let ϕ^* and γ^* be as defined in* (7) *and* (8) *and let the candidate class $\Phi(m, M, L, R_i)$ be based on RDPs of maximal depth J. Suppose further that p and q are uniformly bounded, $c \leq p(\mathbf{x}), q(\mathbf{x}) \leq C$ for all $\mathbf{x} \in [0,1]^d$, $c > 0$, $C < \infty$. Then the approximation error is bounded as*

$$D(\gamma^*) - D(\phi^*) \leq \text{const}(\beta, c, C, m, M, L)\, 2^{-J}. \qquad (11)$$

The significance of (11) is that this rate can now be balanced with the estimation error rate found in [1] to obtain a convergence rate for the total expected error $D(\gamma^*) - \mathbf{E}D(\widehat{\phi}_n)$. This means that for a given problem, an appropriate RDP depth can be calculated that ensures an overall convergence rate equal to that reported in [1], which in particular, can be as fast as n^{-1}.

Proof Outline of Main Result. The proof proceeds by constructing a quantization rule ϕ' from a RDP whose behavior on the boundary of the ideal partition $\{A_i^*\}$ can be characterized. It can then be shown that the approximation error is bounded above by the difference $D(\gamma^*) - D(\phi')$ and that this quantity is in turn bounded by $\|\gamma^* - \phi'\|_{L_1}$. The result then follows from showing that $\|\gamma^* - \phi'\|_{L_1}$ is bounded by a quantity that decays with the depth of the original RDP. We provide some of the details below, beginning with the following lemma and the definition of ϕ'.

Lemma 1 ([11], Lemma 5, p. 121)**.** *There is a RDP such that the cells intersecting $B(\gamma^*)$ are at depth J and all the other cells are at depths no greater than J. Denote the smallest such RDP by π_J^*. Then π_J^* has at most $2^{2d}\beta 2^{(d-1)J}$ cells intersecting $B(\gamma^*)$.*

Let ϕ' denote a L-level piecewise constant function defined on π_J^*,

$$\phi'(\mathbf{x}) = \sum_{i=0}^{L-1} c_{R_i'} \mathbf{1}_{R_i'}(\mathbf{x}), \quad c_{R_i'} = \frac{P(R_i')}{Q(R_i')}, \qquad (12)$$

with the added condition that cells $S \in \pi_J^*$ contained in $R_i'/B(\gamma^*)$ are also contained in A_i^*. More concisely, we assume

$$S \subseteq R_i'/B(\gamma^*) \Rightarrow S \subseteq A_i^*/B(\gamma^*).$$

This condition implies that each disjoint region R_i' is a union of cells from π_J^* and that the partitions $\{R_i'\}$ and $\{A_i^*\}$ coincide except possibly on the boundary $B(\gamma^*)$.

Observe that $D(\gamma^*) - D(\phi^*) \leq D(\gamma^*) - D(\phi')$ since the divergence between the pmfs induced by ϕ' must necessarily be less than or equal to the that induced by the best in class quantization rule ϕ^*. It then follows from the bounds on ϕ, p, and q and from the inequality $\log x \leq x - 1$, for $x > 0$ that

$$D(\gamma^*) - D(\phi') \leq \frac{C + cm}{m} \|\gamma^* - \phi'\|_{L_1}. \qquad (13)$$

This norm can be decomposed as a sum of the norms on each cell R_i' and over the set of cells $S \subseteq R_i'$ that do, and do not, intersect the boundary $B(\gamma^*)$.

$$\|\gamma^* - \phi'\|_{L_1} = \sum_{i=0}^{L-1} \left[\sum_{S \subseteq R_i'/B(\gamma^*)} \int_S |\gamma^*(\mathbf{x}) - \phi'(\mathbf{x})|\, d\mathbf{x} + \sum_{S \subseteq R_i'(B(\gamma^*))} \int_S |\gamma^*(\mathbf{x}) - \phi'(\mathbf{x})|\, d\mathbf{x} \right] \qquad (14)$$

Here, $S \subseteq R_i'/B(\gamma^*)$ means all cells S that are a subset of R_i' which do not intersect the boundary $B(\gamma^*)$. Similarly, $S \subseteq R_i'(B(\gamma^*))$ means all cells S that are subsets of R_i' which do intersect $B(\gamma^*)$.

By the boundedness assumptions on γ^* and ϕ', the second integrand on the right hand side of (14) can be upper bounded by M. Therefore,

$$\sum_{S \subseteq R_i'(B(\gamma^*))} \int_S |\gamma^*(\mathbf{x}) - \phi'(\mathbf{x})|\, d\mathbf{x} \leq M \sum_{S \subseteq R_i'(B(\gamma^*))} \text{Vol}(S) \leq M 2^{-dJ} \beta 2^{2d} 2^{(d-1)J}, \qquad (15)$$

where the second inequality follows from Lemma 1.

Similarly, it can also be shown that the first term on the right hand side of (14) can be bounded as

$$\sum_{i=0}^{L-1} \sum_{S \subseteq R_i'/B(\gamma^*)} \int_S |\gamma^*(\mathbf{x}) - \phi'(\mathbf{x})|\, d\mathbf{x} \leq (C/c)^2 \beta' 2L 2^{-J} \qquad (16)$$

The result follows by combining (13), (14), (15), and (16). □

4. REFERENCES

[1] M.A. Lexa, "Empirical quantization for sparse sampling systems," *Proc. IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing*, pp. 3942–3945, Mar 2010.

[2] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[3] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[4] J. Tropp, J. Laska, M. Duarte, J. Romberg, and R. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Info. Th.*, vol. 56, no. 1, pp. 520–544, 2010.

[5] M. Mishali and Y.C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE J. Sel. Topics Sig. Process.*, vol. 4, no. 2, pp. 375 –391, 2010.

[6] J.N. Tsitsiklis, "Extremal properties of likelihood-ratio quantizers," *IEEE Trans. Comm.*, vol. 41, no. 4, pp. 550–558, Apr 1993.

[7] R. Gupta and A. O. Hero, "High-rate vector quantization for detection," *IEEE Trans. Info. Th.*, vol. 49, no. 8, pp. 1951–1969, Aug 2003.

[8] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1294–1309, Jul 2009.

[9] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," Technical report, Department of Statistics, University of California, Berkeley, Sep 2008.

[10] R. Tyrrell Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.

[11] Rui Castro, *Active Learning and Adaptive Sampling for Nonparametric Inference*, Ph.D., Rice University, Houston TX, U.S.A., Aug 2007.