Rate-Distortion Bounds on Bayes Risk in Supervised Learning

Matthew Nokleby*, Ahmad Beirami†, Robert Calderbank†
*Wayne State University, Detroit, MI, email: matthew.nokleby@wayne.edu
†Duke University, Durham, NC, email: \{ahmad.beirami, robert.calderbank\}@duke.edu

Abstract—An information-theoretic framework is presented for estimating the number of labeled samples needed to train a classifier in a parametric Bayesian setting. Ideas from rate-distortion theory are used to derive bounds for the average $L_1$ or $L_\infty$ distance between the learned classifier and the true maximum a posteriori classifier in terms of familiar information-theoretic quantities and the number of training samples available. The maximum a posteriori classifier is viewed as a random source, labeled training data are viewed as a finite-rate encoding of the source, and the $L_1$ or $L_\infty$ Bayes risk is viewed as the average distortion. The result is a framework dual to the well-known probably approximately correct (PAC) framework. PAC bounds characterize worst-case learning performance of a family of classifiers whose complexity is captured by the Vapnik-Chervonenkis (VC) dimension. The rate-distortion framework, on the other hand, characterizes the average-case performance of a family of data distributions in terms of a quantity called the interpolation dimension, which represents the complexity of the family of data distributions. The resulting bounds do not suffer from the pessimism typical of the PAC framework, particularly when the training set is small.

I. INTRODUCTION

A fundamental question in supervised learning is the number of labeled samples needed to train an effective classifier. The prevailing approach to this question is the probably approximately correct (PAC) framework, which provides almost sure bounds for families of classifiers irrespective of the data distribution. These bounds are expressed in terms of quantities such as the Vapnik-Chervonenkis (VC) dimension, which captures combinatorially the complexity of families of classifiers [1]–[3]. A typical result goes as follows: for a classifier family with VC dimension $h$ and given $n$ labeled samples, the excess classification error due to imperfect learning scales as $O(\sqrt{h}/n)$ with high probability.

A drawback of PAC bounds is that they characterize the performance under the worst-case distribution. This leads to pessimistic predictions [4], [5], and practitioners often resort to cross-validation or regularization techniques to validate classifier performance. Furthermore, PAC bounds are only predictive when the number of training samples is larger than the VC dimension. Finally, VC dimension bounds apply primarily to binary classification problems.

We present a framework for computing sample complexity bounds in a parametric Bayesian setting which are not pessimistic, which are predictive when the training set is small, and which accommodate multi-class settings. We derive lower bounds on the average $L_1$ and $L_\infty$ distance between the true posterior $p(y|x; \theta)$ and the posterior estimated from $n$ i.i.d. samples from $p(x, y; \theta)$. Because the bounds are averaged over the prior $q(\theta)$, they do not exhibit the pessimism of the worst-case PAC bounds. Finally, because they bound the $L_1$ and $L_\infty$ error in estimating the posterior, they readily apply to multi-class parametric families. The $L_1$ and $L_\infty$ errors are surrogates for the excess classification error [6], [7], thus bounds on them predict the performance of the learned classifier.

The proposed framework is inspired by a close relationship between supervised learning and rate-distortion theory. In the parametric Bayesian setting, the posterior distribution is a function of the random parameters $\theta$ and therefore is a random object. If we take the $L_1$ or $L_\infty$ distance as the distortion function, we can bound the number of bits needed to describe the posterior to within a specified tolerance. What follows is the main result of this paper: in order to drive the average $L_1$ or $L_\infty$ error below a threshold $\epsilon$, the mutual information between the training samples and the parameters $\theta$ must be at least as great as the differential entropy of the posterior plus a penalty term that depends on $\epsilon$ and a sample-theoretic quantity, called the interpolation dimension, which measures the number of data points from the posterior distribution needed to uniquely interpolate the entire function.

The resulting framework is a dual to the PAC framework. Whereas the PAC framework considers families of classifiers and provides performance bounds that hold for any data distribution, the rate-distortion framework considers families of data distributions and provides performance bounds that hold for any classifier. Whereas the VC dimension characterizes the combinatorial complexity of a family of classifiers, the interpolation dimension characterizes the sample complexity of a parametric family of data distributions. The larger the interpolation dimension, the more training samples are needed to guarantee classifier performance.

The rest of this paper is organized as follows. In Section II we lay out the problem formally. In Section III we present the rate-distortion framework and state the main results. In Section IV, we evaluate the bounds in a binary and multi-class Gaussian setting. For proofs and additional results, we refer the reader to the full-length version of this paper [8].

Notation: Let $| \cdot |$ denote the cardinality of a set. For a function $f(x)$ and a finite set $S$, let $\{f(x)\}_S$ denote the $|S|$-length vector of function evaluations of $f$ at the points in $S$, suppressing the arguments when clarity permits. Let $[M] = \{1, \ldots, M\}$ for integer $M$. We use the natural log throughout.
II. Problem Statement

Let each data point \( X \in \mathbb{R}^d \) and its label \( Y \in [M] \) be distributed according to \( p(x, y ; \theta) \), where \( \theta \in \Lambda \subset \mathbb{R}^k \) indexes a smooth parametric family of distributions \( \mathcal{D} = \{ p(x, y ; \theta) : \theta \in \Lambda \} \). The learning machine obtains a sequence of \( n \) samples, denoted \( Z^n = (X^n, Y^n) \), where each pair \( Z_i = (X_i, Y_i), \ 1 \leq i \leq n \) is drawn i.i.d. according to \( p(x, y ; \theta) \). The learning task is to select a classifier \( \hat{y} = w(x) \) from the training samples \( Z^n \). The classifier may be any function that maps \( \mathbb{R}^d \) to \([M]\); we do not restrict our attention to a particular family of classifiers (e.g., the family of linear classifiers) as in the PAC framework. If \( \theta \) were known, one could choose the the maximum a posteriori classifier, which minimizes the classification error:

\[
w_{MAP}(x) = \arg \max_y p(y|x; \theta),
\]

where \( p(y|x; \theta) \) is calculated according to Bayes’ rule. We suppose that the learning machine only knows the family of distributions \( \mathcal{D} \), but not the specific distribution \( p(y|x; \theta) \). The MAP classifier \( w_{MAP} \) is unavailable, and the learning machine has to learn a classifier \( w \) whose misclassification probability will necessarily be no smaller than that of \( w_{MAP} \).

Our objective is to characterize the performance of the learned classifier \( w(x) \) as a function of the number of training samples \( n \). A natural performance metric is the gap between the misclassification error of the learned classifier and that of \( w_{MAP} \):

\[
L_c(x; \theta; w, w_{MAP}) = \Pr(Y \neq w(x)) - \Pr(Y \neq w_{MAP}(x)),
\]

where the probabilities are computed according to \( p(y|x; \theta) \). As discussed in the introduction, the minimax loss with respect to \( L_c \) is characterized by the PAC framework. For a family of classifiers containing the MAP classifier and having with VC dimension \( h \), \( L_c(x; \theta; w, w_{MAP}) = O(\sqrt{h/n}) \) for any distribution over \( X \) and \( Y \). The minimax loss, however, is pessimistic compared to typical loss.

Therefore, we analyze the Bayes risk. Let \( q(\theta) \) be a prior distribution over the parametric family. We consider performance averaged over the family of distributions according to \( q(\theta) \). The misclassification error gap turns out to be difficult to study under the proposed rate-distortion framework, and we instead study the \( L_1 \) and \( L_\infty \) losses. These losses are defined in terms of estimates of the posterior, which are then used to classify according to the MAP rule. For notational convenience, let \( \hat{W}(y|x; \theta) = p(y|x; \theta) \) be the posterior. Also let \( \delta(Z^n) = \hat{W}(y|x) \) be a learning rule that maps the training samples to the estimate of the posterior. Then, for every \( x \), the losses are the \( L_1 \) and \( L_\infty \) distances between the \( M \)-dimensional vector formed by \( \hat{W}(.|x; \theta) \) and \( \hat{W}(.|x) \):

\[
L_1(x; W, \hat{W}) = \sum_{y=1}^{M} |W(y|x; \theta) - \hat{W}(y|x)|,
\]

\[
L_\infty(x; W, \hat{W}) = \max_{1 \leq y \leq M} 2|W(y|x; \theta) - \hat{W}(y|x)|.
\]

A well-known fact (see, e.g. [6]) is that \( L_1 \), averaged over \( X \), bounds the classification loss \( L(w, w_{MAP}) \) from above. A somewhat less well-known fact is that this relationship holds pointwise. For any point \( x \), a MAP classifier that uses \( \hat{W}(y|x) \) as the posterior will have a classification loss no larger than \( L_1 \). Also somewhat less well-known is that \( L_\infty \) also bounds the classification loss pointwise [7]. Indeed, the following relationship holds both on the average and pointwise:

\[
L \leq L_\infty \leq L_1.
\]  

The main contribution of this paper is a framework for computing lower bounds on the Bayes risk with respect to the loss functions \( L_1 \) and \( L_\infty \). We consider two risk functions that average over different random variables. In the first risk function, we make high-probability guarantees over the test point \( X \) and the training set \( Z^n \). These guarantees are akin to high-probability guarantees of the PAC framework, where the loss bounds hold everywhere except for a set of sufficiently low probability. Our guarantees still hold on the average with respect to \( \theta \) rather than the worst-case. To make these high-probability guarantees, we consider subsets \( \mathcal{Z} \) and \( \mathcal{X} \) of possible training sets and test points, respectively, having probability greater than \( (1 - \gamma) \) for \( \gamma > 0 \). Then, we consider the minimax Bayes risk, where we take the minimum risk over all high-probability subsets, the maximum risk over test points and training sets in the high-probability subsets, and average over \( \theta \). Concretely, define the \( \gamma \)-Bayes risk:

\[
\hat{R}_1(\delta, \gamma) = \inf_{\mathcal{X} \subset \mathcal{R}^d} \sup_{(x, z^n) \in \mathcal{X} \times \mathcal{Z}^n} E_{\theta}[L_1(x, \theta; W, \hat{W})],
\]

\[
\hat{R}_\infty(\delta, \gamma) = \inf_{\mathcal{X} \subset \mathcal{R}^d} \sup_{(x, z^n) \in \mathcal{X} \times \mathcal{Z}^n} E_{\theta}[L_\infty(x, \theta; W, \hat{W})],
\]

\[
\hat{R}_c(\delta, \gamma) = \inf_{\mathcal{X} \subset \mathcal{R}^d} \sup_{(x, z^n) \in \mathcal{X} \times \mathcal{Z}^n} E_{\theta}[L_c(x, \theta; w, w_{MAP})].
\]

In other words, we require that the Bayes risk be bounded for all test points and training sequences in a set of sufficiently high probability.

Just as the worst-case bounds from the PAC framework may be pessimistic, the high-probability guarantees associated with \( \hat{R} \) may be too restrictive to describe typical behavior. Therefore, we define a second risk function, which we simply call the Bayes risk, in which we average over all of the random variables:

\[
R_1(\delta) = E_{X, Z^n, \theta}[L_1(X, \theta; W, \hat{W})]
\]

\[
R_\infty(\delta) = E_{X, Z^n, \theta}[L_\infty(X, \theta; W, \hat{W})]
\]

\[
R_c(\delta) = E_{X, Z^n, \theta}[L_c(X, \theta; w, w_{MAP})].
\]

In other words, the Bayes risk \( R \) is simply the average performance.

III. Main Results

We first define a few important concepts. The posterior \( \hat{W}(y|x; \theta) \) is an uncountable collection of random variables,
one for each point \((x, y) \in \mathbb{R}^d \times [M]\). The mutual information between, or the joint entropy of, uncountably many random variables is difficult to analyze directly, which makes it difficult to carry out the rate-distortion analysis on the Bayes risk. Therefore, we will analyze the information-theoretic quantities of a \(\text{sampled}\) version of \(W(y|x; \theta)\), which acts as a sufficient statistic for the entire function. We capture this notion by defining the interpolation set and the interpolation dimension.

**Definition 1 (interpolation set/dimension):** A finite set \(S \subset \mathbb{R}^m \times [M]\) is called an interpolation set for the posterior \(W(y|x; \theta)\) if there is a deterministic mapping from the set \(\{W(y|x; \theta)\}_S\) to the function \(W(y|x; \theta)\) for any \(\theta \in \mathbb{R}^d\). Furthermore, let the interpolation dimension \(d_I(W)\) be the cardinality of the smallest interpolation set. If there is no such set, \(d_I(W) = \infty\).

In the sequel, we will need a quantity that we call the multiplicity of the interpolation set.

**Definition 2 (multiplicity):** The multiplicity of an interpolation set \(S\) is the minimum over the number of times a point \(x \in \mathbb{R}^d\) appears in \(S\).

\[
\mu(S) = \min_{x \in \mathbb{R}^m} |\{(x, y) \in S : x = x'\}|.
\]

The maximum multiplicity is \(M - 1\). This is because the posterior is normalized. For a point \(x \in \mathbb{R}\) appearing in \(S\), the evaluation of \(W(y|x; \theta)\) at \(x\) for \(M - 1\) values of \(y\) suffices to determine \(W\) for the remaining value of \(y\).

The interpolation dimension characterizes the number of distinct evaluations of the posterior function needed to reconstruct the entire posterior function, and it is thus the analogue to the Nyquist sampling rate for families of parametric distributions. The interpolation dimension also serves as a measure of complexity similar to the VC dimension. The VC dimension characterizes the complexity of a family of classifiers by how many points it can shatter, whereas the interpolation dimension characterizes the complexity of a family of distributions by how many sample points of the posterior are needed to reconstruct it.

### A. Bounds on \(\hat{R}(\delta)\)

The first result is a bound on the \(\gamma\)-Bayes risk in terms of the interpolation dimension, the entropy of the posterior, and the mutual information between the training samples and the parameter \(\theta\).

**Theorem 1:** Let \(d_I(W) < \infty\) be the interpolation dimension of \(W(y|x; \theta)\), and let \(S\) be any interpolation set with \(|S| = d_I(W)\). Then, there exists a learning rule \(\delta(Z^n)\) such that

\[
\hat{R}_2(\delta, \gamma) \leq \epsilon \quad \text{only if}
\]

\[
I(Z^n; \theta) \geq \inf_{p(z) \geq 1-\gamma} \sup_{S \subset [X \times \{1, \ldots, M\}]} h(\{W\}_S) - \frac{1}{M} \gamma \log \left( \frac{M}{\mu(S)} \right) + H(\gamma) + (1 - \gamma) \log \left( \frac{\epsilon}{\mu(S)} \right),
\]

whenever \(h(\{W\}_S)\) exists and is finite. Furthermore, there exists a learning rule \(\delta(Z^n)\) such that \(\hat{R}_2(\delta, \gamma) \leq \epsilon \quad \text{only if}

\[
I(Z^n; \theta) \geq \inf_{p(z) \geq 1-\gamma} \sup_{S \subset [X \times \{1, \ldots, M\}]} h(\{W\}_S) - \frac{1}{M} \gamma \log \left( \frac{M}{\mu(S)} \right) + H(\gamma) + (1 - \gamma) \log \left( \frac{\epsilon}{\mu(S)} \right),
\]

whenever \(h(\{W\}_S)\) exists and is finite. Furthermore, there exists a learning rule \(\delta(Z^n)\) such that \(\hat{R}_2(\delta, \gamma) \leq \epsilon \quad \text{only if}

\[
I(Z^n; \theta) \geq \inf_{p(z) \geq 1-\gamma} \sup_{S \subset [X \times \{1, \ldots, M\}]} h(\{W\}_S) - \frac{1}{M} \gamma \log \left( \frac{M}{\mu(S)} \right) + H(\gamma) + (1 - \gamma) \log \left( \frac{\epsilon}{\mu(S)} \right),
\]

whenever \(h(\{W\}_S)\) exists and is finite. Furthermore, there

### B. Bounds on \(R(\delta)\)

To prove bounds on \(R(\delta)\), we require additional machinery. In particular, we consider averages over interpolation sets, from which arises the notion of an interpolation map.

**Definition 3 (interpolation map):** Let the map \(S : \mathbb{R}^d \times [M] \rightarrow (\mathbb{R}^d \times [M])^d\) send sample point \((x, y)\) to an interpolation set \(S(x, y)\), such that \((x, y) \in S(x, y)\). We call \(S\) an interpolation map if \(S(x, y)\) if the following conditions hold, except for a set of Lebesgue measure zero: (1) \(S(x, y)\) is defined, (2) each point \((x', y') \in \mathbb{R}^d \times [M]\) appears in a constant, nonzero, and finite number of sets \(S(x, y)\), and (3) for each \((x', y') \in S(x, y)\), \(p(x') = p(x)\). Furthermore, let \(S(x, y) = \{s_1(x, y), \ldots, s_{d_I(s, y)}(x, y)\}\) denote the mapping pointwise, where \(s_1(x, y) = (x, y)\) without loss of generality.

An interpolation map may not always exist, but the conditions for existence are not too restrictive. For example, in Section IV we consider a binary Gaussian scenario in which any basis in \(\mathbb{R}^d\) is an interpolation set. In this case, \(S(x, y)\) maps to an orthogonal basis containing \((x, y)\). The map fails
only to map the point \( x = 0 \) to an interpolation set.

To account for the exclusion of a measure-zero set from the conditions of the interpolation map, we define an effective interpolation dimension of an interpolation map.

**Definition 4 (interpolation subdimension):** For an interpolation map \( S(x, y) \), let \( W \subseteq \mathbb{R}^d \times [M] \) be the measure-zero set on which the conditions in Definition 3 do not hold. Then, define the **interpolation subdimension of** \( S \):

\[
    d_S(S) = \min_{(x, y) \in S(x, y) \setminus W} |S(x, y)\setminus W|,
\]

where the minimization is over the \((x, y)\) pairs for which \( S(x, y) \) is defined.

When an interpolation map exists, we can bound the Bayes risk with expressions similar to those of Theorem 1. 

When an interpolation map exists, we can bound the Bayes risk with expressions similar to those of Theorem 1.

**Theorem 2:** Suppose the density \( p(x) \) is Riemann integrable, and suppose an interpolation map \( S(x, y) \) exists with interpolation dimension \( d_I \) and interpolation subdimension \( d_S \). There exists a learning rule \( \delta(Z^n) \) with Bayes risk \( \tilde{R}_1 \leq \epsilon \) only if

\[
    I(Z^n; \theta) \geq \max_y E_X[h(W)_{S(x, y)}]
\]

\[
    d_S \left( \log \left( \frac{M}{2\epsilon} \right) - 1 \right) - (d_I - d_S) \log(2) + 1.
\]

Similarly, there exists a learning rule \( \delta(Z^n) \) with Bayes risk \( \tilde{R}_\infty \leq \epsilon \) only if

\[
    I(Z^n; \theta) \geq \max_y E_X[h(W)_{S(x, y)}]
\]

\[
    d_S \left( \log \left( \frac{1}{2\epsilon} \right) - 1 \right) - (d_I - d_S) \log(2) + 1.
\]

**IV. NUMERICAL EXAMPLES**

A. Binary Gaussian Classifier

First, we consider a binary Gaussian setting. Let \( M = 2 \) and \( p(y) = 1/2, y \in \{1, 2\} \), and let \( \Lambda = \mathbb{R}^d \) parameterize the data distributions. The class-conditional densities are Gaussian with antipodal means:

\[
    p(x|y = 1; \theta) = \mathcal{N}(\theta, \sigma^2 I),
\]

\[
    p(x|y = 2; \theta) = \mathcal{N}(-\theta, \sigma^2 I),
\]

where \( \sigma > 0 \) is the known variance. We choose the prior \( q(\theta) = \mathcal{N}(0, (1/d)I) \).

The MAP classifier for this problem is a hyperplane passing through the origin and normal to \( \theta \). The VC dimension for this family of classifiers is \( d \), suggesting that \( d \) samples are necessary for good performance. This may be pessimistic. For \( \sigma \ll d \), a single labeled sample is sufficient to obtain an accurate estimate \( \hat{\theta} \) and recover a near-optimum classifier.

In this setting, we can derive the rate-distortion bounds on the L1 and \( L_\infty \) Bayes risk in closed form. First, we can evaluate the mutual information between \( \theta \) and \( Z^n \) in closed form. Simple calculation reveals that

\[
    I(Z^n; \theta) = \frac{d}{2} \log \left( 1 + \frac{n}{d\sigma^2} \right).
\]

Next, we bound the entropy of the posterior, which by Bayes rule is the sigmoid of the inner product of \( \theta \) with \( x \):

\[
    W(y = 1|x; \theta) = \frac{1}{1 + \exp(-2/\sigma^2 x^T \theta)}. \tag{8}
\]

To interpolate \( W \), it is sufficient to sample \( W \) at any basis of \( \mathbb{R}^d \) for any value of \( y \). Therefore, we choose the interpolation map

\[
    S(x, y) = \{x, x_2, \ldots, x_d \} \times \{y\}, \tag{9}
\]

where \( \{x, x_2, \ldots, x_d \} \) is an orthogonal basis and where \( \|x_i\| = \|x\| \) for every \( i \). The interpolation dimension and the interpolation subdimension are both equal to \( d \). Choosing orthogonal points makes the samples statistically independent, and in the following theorem we present a bound on the expected differential entropy.

**Theorem 3:** The expected differential entropy of the posterior evaluated at the interpolation map is

\[
    E_X[h(W)_{S(x, y)}] = \sum_{y} p(y) E_x[h(W)_{S(x, y)}] \geq \frac{d}{2} \psi(d/2) + \frac{d}{2} \log \left( \frac{16\pi(1/(da^2) + 1)}{d/2} \right)
\]

\[
    = \frac{\Gamma(d+1)/2}{\Gamma(1/2)} \sqrt{3(1/(da^2) + 1) - 3d/2} - 2d \log(2),
\]

where \( \Gamma(\cdot) \) is the Gamma function, and \( \psi(\cdot) \) is the digamma function, and where \( \nu(d, \sigma^2) \) is defined as the lower bound divided by \( d \).

Combining these results, and recalling that the \( L_1 \) and \( L_\infty \) errors are identical for \( M = 2 \), we obtain the following bound on the \( L_1 \) Bayes risk:

\[
    R_1(\delta) = R_\infty(\delta) \geq \sqrt{\frac{\sigma^2 d}{\sigma^2 d + n}} \exp(\nu(d, \sigma^2) - 1). \tag{10}
\]

Therefore, we expect expect the \( L_1 \) and \( L_\infty \) Bayes risk to be small when \( n \gg d\sigma^2 \). When the signal dimension is high, it will in general require more samples to train a classifier, but this requirement can be mitigated by small model noise variance. We emphasize that the preceding bound holds for all \( n \); one does not need to ensure that \( n > d \), as is necessary in the PAC framework.

In Figure 1, we evaluate the bounds numerically. We select \( d = 50 \) and \( \sigma^2 = 0.5 \), and estimate the relevant quantities from \( 200 \times 200 \) draws of \( \theta \), the training set \( Z^n \), and test points \( (x, y) \). We plot the bound on \( R \), the empirical average of the losses, the empirical average of the excess classification error \( L_\text{e}(w, w_{\text{MAP}}) \), and the PAC bound derived from the VC dimension. For the empirical quantities, we find the MAP estimate of \( \theta \) from the \( Z^n \) and plug the result into the posterior \( W(y|x; \theta) \). The PAC bound significantly overestimates the classification error, whereas the the Bayes risk bound is on the same order as \( R_c \).

B. Multi-class Gaussian

Next, we consider a multi-class Gaussian setting. Let \( p(x|y; \theta) = \mathcal{N}(\theta_y, \sigma^2 I) \), where \( \theta \in \mathbb{R}^{dM} \), and each \( \theta_y \in \mathbb{R}^d \) is the \( y \)th sub-vector of length \( d \) and denotes the mean of the \( y \)th class. Let \( p(y) = 1/M \) and \( q(\theta) = \mathcal{N}(0, 1/dI) \). In this
Bayes Risk

Fig. 1. Bayes risk bounds and empirical performance for binary Gaussian classification.

In this case, the mutual information is

\[ I(Z^n; \theta) = \frac{Md}{2} \log \left( 1 + \frac{n}{d \sigma^2 M} \right). \] (11)

It is possible to show that a valid interpolation map is

\[ S(x, y) = \{0, x, x_2, \ldots, x_d\} \times [M - 1], \] (12)

where \( x, x_2, \ldots, x_d \) are an orthogonal basis, and \( \|x_i\| = \|x\| \) for every \( i \). The measure-zero set \( W = \{0\} \times [M - 1] \) appears in infinitely many interpolation sets, in violation of the conditions of Definition 3. Therefore, the interpolation subdimension is \( d_S = (M - 1)d \). In this case, we cannot solve for the differential entropy of the posterior in closed form. We estimate \( E_X[h(\{W\}_{S(X, y)})] \) numerically as described in [11].

In Figures 2-3, we evaluate the bounds numerically. We choose \( d = 100, \sigma^2 = 0.1 \), and here we take \( M \in \{5, 10\} \). We plot the bounds and empirical Bayes risk and the empirical average of \( L(w, w_{MAP}) \), again over \( 200 \times 200 \) samples of \( \theta \), the training set, and the test point, and again using the MAP estimate of \( \theta \) from \( Z^n \) to compute the empirical quantities. We also consider the PAC bound given by the Natarajan dimension [12]. Following [13], we estimate the Natarajan dimension as \( M \cdot d \), and we plot the resulting bound.

We observe a few phenomena. First, the gap between the predicted \( L_1 \) and \( L_\infty \) risk grows with increasing \( M \), as predicted. As \( M \) increases, the bound on the \( L_\infty \) Bayes risk becomes optimistic by a few orders of magnitude, while the bound on the \( L_1 \) risk remains a reasonable prediction of the classification risk for all values of \( M \). The PAC bound, on the other hand, is pessimistic in all regimes. Finally, the empirical \( L_1 \) and \( L_\infty \) errors are closer together than predicted by theory, although the gap does grow in \( M \). This is because the estimation error of the posterior \( W(y|x; \theta) \) tends empirically to be “peaky” in \( y \), concentrating around a single value, leading to similar errors regardless or norm. The bounds on \( \hat{R}_1 \) and \( \hat{R}_\infty \), on the other hand, optimistically suppose the error is evenly spread among the classes, in which case the norms differ substantially. The extent to which this result generalizes beyond this scenario is a topic for further investigation.

Fig. 2. Rate-distortion bounds vs. empirical error for \( M = 5 \).

Fig. 3. Rate-distortion bounds vs. empirical error for \( M = 10 \).

REFERENCES