A RATE-DISTORTION FRAMEWORK FOR SUPERVISED LEARNING

Matthew Nokleby, Ahmad Beirami, Robert Calderbank
Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA. Email: {matthew.nokleby, ahmad.beirami, robert.calderbank}@duke.edu

ABSTRACT
An information-theoretic framework is presented for bounding the number of samples needed for supervised learning in a parametric Bayesian setting. This framework is inspired by an analogy with rate-distortion theory, which characterizes tradeoffs in the lossy compression of random sources. In a parametric Bayesian environment, the maximum a posteriori classifier can be viewed as a random function of the model parameters. Labeled training data can be viewed as a finite-rate encoding of that source, and the excess loss due to using the learned classifier instead of the MAP classifier can be viewed as distortion. A strict bound on the loss—measured in terms of the expected total variation—is derived, providing a minimum number of training samples needed to drive the expected total variation to within a specified tolerance. The tightness of this bound is demonstrated on the classification of Gaussians, for which one can derive closed-form expressions for the bound.

Index Terms—Supervised Learning; Rate-Distortion Theory; Binary Classification; Bayesian Statistics.

1. INTRODUCTION
In supervised learning, the learning machine obtains a finite collection of labeled training samples, drawn from an unknown distribution, from which it must infer a classifier to be applied to further data samples drawn from the same distribution. Common approaches include support vector machines [1–3], random forests [4], and deep neural networks [5,6].

A central question is how many training samples is sufficient to learn an effective classifier. The dominant approach to this question is the probably approximately correct (PAC) framework, which provides almost sure bounds on the generalization performance of families of classifiers irrespective of the data distribution. Typical bounds are expressed in terms of the Vapnik-Chervonenkis (VC) dimension, which combinatorially measures the flexibility of the classifier family [7,8].

The VC dimension provides a rough measure of the bias-variance tradeoff in learning: A classifier family with high VC dimension can model complex data, but it will overfit if there are too few samples.

Although PAC bounds provide strict upper bounds and valuable rules of thumb on performance, they inherently characterize worst-case performance, which often leads to pessimistic predictions [9,10]. Refinements of the PAC framework abound, such as distribution-dependent performance bounds [11], average-case performance bounds [9,12], and margin-based bounds depending on the fat-shattering dimension [13]. Even these bounds are well-known to be pessimistic, however, and machine learning practitioners often resort to cross-validation to verify learning performance. Cross-validation is computationally expensive for complex learning machines and large data sets. Hence, there is a need for accurate, a priori bounds on the number of training needed in machine learning applications.

In this paper, we attack the supervised learning problem from the opposite direction, and the result is a complement to the PAC framework. Instead of considering families of classifiers, we consider parameterized families of data distributions over which we define a prior. This leads us to a Bayesian analysis of classifier performance. Drawing an analogy with rate-distortion theory, we prove lower bounds on the number of training samples needed to learn a maximum a posteriori classifier to within a specified tolerance measured in the average total variation. The resulting bounds are predictive even when the number of training samples is small and are tailored to the relevant family of data distributions.

The main result is the requirement that the mutual information between the training samples and the data distribution exceed the entropy of the MAP classifier plus a penalty term that depends on the error tolerance. The key quantity in the error term is the interpolation dimension, which is the minimum number of samples of the posterior distribution needed to interpolate the entire function. This quantity plays a role dual to the VC dimension in the PAC framework. It captures the inherent complexity of the data distribution just as the VC dimension captures the complexity of the classifier family. The higher the interpolation dimension, the more training samples are needed to avoid classifier mismatch.

We validate the proposed framework by evaluating the bounds for a binary Gaussian classification problem. Over this family of distributions, the optimum MAP classifier is a linear threshold function. Both the VC dimension and
interpolation dimension are linear in the signal dimension. Whereas the PAC bounds require the number of samples also to grow linearly in the signal dimension, our bounds apply for smaller sample size. Under certain regimes, the PAC bounds are loose, whereas the bounds of the proposed framework are relatively tight.

2. PRELIMINARIES

2.1. Background Review

Let each data sample \( X \in \mathbb{R}^d \) and its label \( Y \in \{0, 1\} \) be distributed according to \( p(X, Y; \theta) \), where \( \theta \in \Lambda \subset \mathbb{R}^b \). Therefore, the set \( \{p(X, Y; \theta)\}_{\theta \in \Lambda} \) defines a family of distributions parametrized by \( \theta \). We suppose that the distribution family is known to the learning machine, as is a prior distribution \( \theta \) parametrized by \( t \).

The typical strategy is to choose the classifier according to the i.i.d. training samples according to \( p(X, Y; \theta) \). The learning task is to select an optimum classifier based on the i.i.d. training samples \( Z^n \).

In the PAC framework, one considers a family of classifiers \( \mathcal{H} \), where each \( h \in \mathcal{H} \) is a mapping \( h : \mathbb{R}^d \to \{0, 1\} \). The typical strategy is to choose the classifier according to the empirical risk minimization (ERM) principle, which in this setting is simply the classifier \( h \in \mathcal{H} \) that makes the fewest classification errors (on the average) over the training samples \( Z^n \). One can bound the generalization error of the ERM classifier by the VC dimension of \( \mathcal{H} \).

**Definition 1 (VC dimension)** We say that \( \mathcal{H} \) shatters a finite set \( T \subset \mathbb{R}^d \) if there is a \( h \in \mathcal{H} \) that instantiates each of the \( 2^{|T|} \) labelings of \( T \). The VC dimension of \( \mathcal{H} \), denoted \( \text{VC}(\mathcal{H}) \), is the cardinality of the largest \( T \) that is shattered by \( \mathcal{H} \). If no such set exists, the VC dimension is infinite.

The VC dimension characterizes the bias-variance tradeoff of \( \mathcal{H} \). A rich family of classifiers can model complicated decision regions, which leads to low bias. However, if the number of training samples is small, the classifier will overfit to \( Z^n \), resulting in high variance. This intuition is made precise by a bound on the generalization error expressed in terms of the VC dimension and the number of training samples [8].

For any classifier \( h \) and the true distribution \( p(x, y; \theta) \) over the pairs \( (X, Y) \), let \( R(h) \) (or in short \( R(h) \)) be defined as the misclassification probability given by

\[
R(h) = \Pr[h(X) \neq Y],
\]

where the underlying probability measure is \( p(x, y; \theta) \). Further, let \( R_{\text{emp}}(h) \) be the empirical classification error for any \( h \in \mathcal{H} \) over the training set \( Z^n \). Then, with probability \( 1 - \delta \) for any small \( \delta > 0 \), the misclassification probability over subsequent data samples is bounded by [8]

\[
R(h) \leq R_{\text{emp}}(h) + \sqrt{\frac{\text{VC}(\mathcal{H}) (\log \frac{2n}{\text{VC}(\mathcal{H})} + 1) + \log(4/\delta)}{n}},
\]

for \( n \geq \text{VC}(\mathcal{H}) \). When the VC dimension is high, one can choose a classifier that achieves a small empirical risk, but the true risk may be high unless the number of training samples is sufficiently large. The upshot is that one expects to need at least as many training samples as the VC dimension in order to ensure that training performance generalizes to performance over data samples.

This result further leads to the structured risk minimization (SRM) principle. One can train a classifier on a hierarchy of class families \( \mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \) with increasing complexity, compute the empirical risk, and choose the classifier that minimizes the bound in (2) in order to achieve the sweet spot in the bias-variance tradeoff. The VC bounds, however, are well known to be pessimistic, and in practice it is more reliable to cross validate classifier performance instead.

2.2. Problem Setup

We bound the performance of Bayesian classifiers. In particular, we allow the learning machine to output a soft classifier that outputs the likelihood function of the data sample \( X \) rather than selecting one of the labels. We measure the performance of the learned classifier with respect to the optimum soft classifier, which is the posterior distribution \( p(y|x; \theta) \). For notational convenience, define

\[
W(x; \theta) = p(Y = 0|X = x; \theta).
\]

For each point \( x \) and parameter \( \theta \), \( W(x; \theta) \) specifies entirely the posterior distribution. Furthermore, \( W(x; \theta) \) is a random function of the parametric family with prior \( p(\theta) \).

Let \( \hat{W}(x; Z^n) \) denote the soft classifier output by the learning machine, which we view as an estimate of the posterior. Hard classifiers are merely a special case of the soft classifiers considered in this work. We characterize the sub-optimality of the learned classifier in terms of the total variation, which is defined as

\[
\text{TV}(W, \hat{W}) = 2E_{X, \theta}[|W(X; \theta) - \hat{W}(X; W^n)|].
\]

The total variation is the \( \ell_1 \) norm of the difference \( W - \hat{W} \), averaged over the data points \( x \in \mathbb{R}^d \), and is intimately related to the misclassification probability. The following theorem, found in [14] relates the two.

**Theorem 1** ([14]) Let \( h = \arg \max W(x; \theta) \) and \( \hat{h} = \arg \max \hat{W}(x; Z^n) \) be the MAP classifiers associated with the true posterior and the learned estimate with misclassification probabilities \( R(h) \) and \( R(\hat{h}) \) defined in (1), respectively. Then,

\[
E_{\theta}[R(\hat{h}) - R(h)] \leq \text{TV}(W, \hat{W}).
\]
This bound is tight in the sense that when \((R(\hat{h}) - R(h))\) so does \(TV(W, \tilde{W})\). Therefore the total variation is a suitable surrogate for the gap between the true MAP classifier and the learned classifier when the MAP classifier has low risk.

3. AVERAGE-CASE PERFORMANCE BOUNDS

We first define the interpolation dimension, which describes the complexity of the posterior distribution to be approximated by the learning machine.

**Definition 2 (interpolation set/dimension)** A countable set \(S \subset \mathbb{R}^d\) is said to be an interpolation set for the posterior described by \(W(x; \theta)\) if, for all \(\theta \in \Lambda\) and \(x \in \mathbb{R}^d\), \(\{W(x; \theta)\}_{x \in \mathbb{R}^d}\) is a deterministic function of \(\{W(x; \theta)\}_{x \in S}\). Furthermore, let the interpolation dimension be

\[
f(W) = \min \{|S| : S \text{ is an interpolation set for } W(x; \theta)\}.
\]

That is, \(f(W)\) is the cardinality of the smallest \(S\) that is an interpolation set for \(W\), and \(S\) is a minimal interpolation set. If no such interpolation set exists, \(f(W)\) is infinite.

**Definition 3 (minimal sufficient statistic)** If \(S\) is a minimal \(f\)-dimensional interpolation set for \(W(x; \theta)\), then the \(f\)-dimensional vector \(\alpha := \{W(x; \theta)\}_{x \in S}\) is called a minimal sufficient statistic for \(W(x; \theta)\).

One can view the interpolation set/dimension as a generalization of Nyquist sampling for arbitrary parametric sources. The interpolation dimension \(f(W)\) plays a similar role to the VC dimension in our analysis. Whereas the VC dimension characterizes the complexity of a classifier family by how many points it can shatter, the interpolation dimension characterizes the complexity of a distribution family by how many samples of the posterior are needed to uniquely determine it. In each case, one expects intuitively the dimensions to be equal to the dimension of the parameter space, and in each case one can construct counterexamples.

Here we prove our main result, which bounds the total variation in terms of the interpolation dimension, the entropy of the posterior, and the mutual information between the training samples and the parameter \(\theta\).

**Theorem 2** Let \(f\) be the interpolation dimension of \(W(x; \theta)\), and let \(S\) be any interpolation set that achieves the dimension. Then, there exists a learning rule \(W = \tilde{W}(x; z^n)\) such that \(TV(W, \tilde{W}) \leq \epsilon\) only if

\[
I(Z^n; \theta) \geq h(\{W(x; \theta)\}_{x \in S}) + f \log \left( \frac{1}{\epsilon} \right),
\]

where \(\{W(x; \theta)\}_{x \in S}\) is a minimal sufficient statistic, \(I(X; Y)\) is the mutual information between \(X\) and \(Y\), and \(h(X)\) is the differential entropy of \(X\).

**Proof.** First, observe the following Markov chain:

\[
\{W(x; \theta)\}_{x \in S} \rightarrow \theta \rightarrow Z^n \rightarrow \{\tilde{W}(x; Z^n)\}_{x \in S}.
\]

Therefore, invoking the data processing inequality yields

\[
I(Z^n; \theta) \geq I(\{W(x; \theta)\}_{x \in S}; \{\tilde{W}(x)\}_{x \in S}) \geq \inf_{p(W|\theta)} I(\{W(x; \theta)\}_{x \in S}; \{\tilde{W}(x)\}_{x \in S}).
\]

Without loss of generality, suppose \(\tilde{W}(x; Z^n) = W(x; \theta) + U(x)\) for some random variable \(U(x)\). Then, we have

\[
I(Z^n; \theta) \geq \inf_{p(U)} I(\{W(x; \theta)\}_{x \in S}; \{\tilde{W}(x)\}_{x \in S}) + \inf_{p(U)} h(\{W(x; \theta)\}_{x \in S} \mid \{W(x; \theta) + U(x)\}_{x \in S}) - h(\{U(x)\}_{x \in S}).
\]

Taking the infimum over \(p(U)\) is equivalent to choosing the random variables \(U(x)\) to maximize the entropy subject to the total variation constraint, which we can rewrite as

\[
E_X[|U(X)|] \leq \epsilon.
\]

Via Lagrange multipliers it follows that the maximizing choice is for each \(U(x)\) to be independently and uniformly distributed across \([-\epsilon/2, \epsilon/2]\) leading to the desired result.

The content of Theorem 2 is that the information that the training samples \(Z^n\) provide about the parameter \(\theta\) must be at least as great as the entropy of \(W(x)\) plus a penalty term involving the interpolation dimension of \(W(x)\) and the total variation bound. The higher the interpolation dimension, or the lower the tolerance, the more training samples are necessary to satisfy the total variation constraint. Observe that the bounds derived above are technically correct for any set \(S\) whether or not it is an interpolation set for \(W(x)\). Choosing a smaller \(S\), however, gives a looser bound with a smaller penalty term. Deriving the result with the interpolation set ensures the largest correct penalty term. We can further tighten the bound by choosing the interpolation set \(S\) that maximizes the joint entropy \(h(\{W(x; \theta)\}_{x \in S})\).
Remark. The proof of Theorem 2 bears similarity to that of the classic rate-distortion theory in information theory (see [15]). In rate-distortion theory, we have a source signal which we want to compress. An ideal compression uses a small number of bits but provides a high-fidelity representation of the source. The rate-distortion theorem specifies the optimum tradeoff between these two desiderata, relating the number of bits required to compress the source, the amount of distortion introduced by compression, and the entropy of the source. The bounds in Theorem 2 take a similar form, where the mutual information $I(\mathcal{C};\mathcal{S})$ plays the role of the number of bits used to compress, the entropy of the posterior at the interpolation set $h(\{W(x;\theta)\}_{x\in\mathcal{S}})$ takes the role of the source entropy, and the total variation tolerance $\epsilon$ takes the role of the distortion.

Let $\mathcal{S}$ be a minimal $f$-dimensional interpolation set for $W(x;\theta)$. Denote $\mathcal{I}(\alpha)$ as the Fisher information matrix associated with the minimal sufficient statistic $\alpha = \{W(x;\theta)\}_{x\in\mathcal{S}}$, i.e.,

$$
\mathcal{I}_n(\alpha)_{i,j} \equiv \frac{1}{n} \log e \mathbb{E}_{X,Y} \left[ \frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \log \left( \frac{1}{p(X^n,Y^n;\theta)} \right) \right].
$$

(15)

Roughly speaking, Fisher information quantifies the amount of information, on the average, that each symbol in a training sample of size $n$ from the source conveys about the source parameters. We assume that the source is ergodic such that the Fisher information matrix asymptotically converges on the limit $\mathcal{I}(\alpha) := \lim_{n\to\infty} \mathcal{I}_n(\alpha)$.

Then, we can bound the mutual information in terms of the number of training samples $n$.

Theorem 3 Assume a source parametrized with $\theta \in \Lambda$ with minimal sufficient statistic $\alpha \in \mathbb{R}^f$. Then,

$$
I(\mathcal{Z}^n;\theta) \leq \frac{f}{2} \log \frac{n}{2\pi e} + \int_{\mathbb{R}^f} p(d\alpha) \log |\mathcal{I}(\alpha)|^{\frac{1}{2}} + h(\theta) + o(1).
$$

Proof. This is straightforward in light of the celebrated redundancy-capacity theorem (see [16–18]) and averaging the bounds derived by Clarke and Barron [19].

To evaluate the performance bounds, one needs to compute three quantities. First, we need to compute the interpolation dimension $f$ and find an interpolation set $\mathcal{S}$. Second, we need to compute the differential entropy $h(\{W(x;\theta)\}_{x\in\mathcal{S}})$ of the sufficient statistic. Finally, we need to compute the Fisher information integral specified in Theorem 3. These three quantities are sufficient to establish necessary relationships between the number of training samples and the total variation of the learned classifier.

4. APPLICATION: A BINARY GAUSSIAN CLASSIFIER

To illustrate our framework, we consider a binary Gaussian classifier. Let $p(Y) = 1/2$, and let $\Lambda = \mathbb{R}^d$ be the parameterization of the class-conditional densities. The class-conditional densities are the Gaussians

$$
p(X|Y = 0; \theta) = \mathcal{N}(\theta, \sigma^2 I)
p(X|Y = 1; \theta) = \mathcal{N}(-\theta, \sigma^2 I),
$$

(16)

where $\sigma^2 > 0$ is the variance. We take $\theta \sim \mathcal{N}(0,(1/d)I)$. We choose this problem because it is simple enough for closed-form analysis, but rich enough to illustrate the advantages of our framework.

The optimum classifier for this problem is a hyperplane passing through the origin and normal to $\theta$. The classifier family for this problem is the set of hyperplanes in $\mathbb{R}^d$ passing through the origin. The VC dimension of this family is $d$, suggesting that at least $d$ training samples are necessary for good classification performance. However, this is often pessimistic. For $\sigma \approx 0$ and sufficiently large allowed classification error $\epsilon$, a single labeled sample is sufficient to obtain $\theta$ and recover the optimum classifier regardless of $d$. A more comprehensive performance analysis should expose the dependence on both $d$ and $\sigma^2$.

Here we show that our framework yields such an analysis. To evaluate the bounds provided by Theorem 2, we need to evaluate the mutual information $I(\mathcal{Z}^n;\theta)$ and the entropy $h(\{W(x;\theta)\}_{x\in\mathcal{S}})$ for an interpolation set $\mathcal{S}$. We calculate the mutual information by invoking Theorem 3. Observe that in this particular problem, it is straightforward to verify that $\theta$ itself is a minimal sufficient statistic and the interpolation dimension is $d$. Note that $p(x,y;\theta) = p(y;\theta)p(x;y;\theta)$ where $p(y;\theta) = (1/2)$ in this case and $p(x;y;\theta)$ is Gaussian as given by (16).

Hence, log $p(x,y;\theta) = \log p(x;y;\theta) - 1$. Therefore,

$$
\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log \left( \frac{1}{p(x,y;\theta)} \right) = \begin{cases} 0 & \text{if } i \neq j \\ 1/\sigma^2 & \text{if } i = j 
\end{cases}.
$$

The Fisher information matrix is diagonal with equal diagonal entries. Therefore, the second term in Theorem 3 is

$$
\int_{\mathbb{R}^d} p(d\theta) \log |\mathcal{I}(\theta)|^{\frac{1}{2}} = -\frac{d}{2} \log d\sigma^2.
$$

Next, we bound the entropy of the posterior. Applying Bayes’ rule, we obtain

$$
W(x;\theta) = p(Y = 1|X;\theta)
$$

(17)

$$
= \frac{\mathcal{N}(\theta, \sigma^2 I) + \mathcal{N}(-\theta, \sigma^2 I)}{1 + \exp(-2/\sigma^2 x^T \theta)}.
$$

(18)

Observe that $W(x;\theta)$ is an invertible function of the inner product $x^T \theta$. To recover $\{W(x;\theta)\}_{x\in\mathbb{R}^d}$, it is necessary and sufficient to sample at a basis for the $d$-dimensional space. Furthermore, if we choose the points to be orthogonal, the samples $W(x;\theta)$ are statistically independent. Therefore, we choose as our interpolation set the orthogonal basis
Observe that $h$ where $h$ find that the entropy is maximized when the norm is $\theta; \in \{0, 1\}$. Taking the derivative and setting it equal to zero, we obtain the tightest possible bound from Theorem 2.

$S = \{c_1, \ldots, c_d\}$, for arbitrary constant $c$ to be chosen later. To compute its entropy, we first derive the distribution of $W(x; \theta)$ for each $x \in S$. Define

$$Z = 2/\sigma^2 x^T \theta.$$  

(20)

For any $x \in S$, $Z$ is a Gaussian random variable with zero mean and variance $(4c^2)/(d\sigma^4)$. Performing a change of variables on the density of $Z$, we obtain the following expression for the density of $W(x; \theta)$:

$$p(w) = \sigma^2 \frac{1}{2c\sqrt{2\pi}1/dw(1-w)} \exp \left( -\frac{d\sigma^4}{c^2} \log^2 \left( \frac{w}{1-w} \right) \right).$$

The entropy of each $W(x; \theta)$ is therefore

$$h(W(x; \theta)) = -E[\log(p(W(x; \theta)))]$$

$$= E \left[ \log \left( \sqrt{2\pi}1/d\sigma^2 W(x; \theta)(1-W(x; \theta)) \right) \right]$$

$$- E \left[ \frac{d\sigma^4}{8c^2} \log^2 \left( \frac{W(x; \theta)}{1-W(x; \theta)} \right) \right].$$  

(21)

Observe that

$$W(x; \theta)(1-W(x; \theta)) = \frac{\exp(-Z)}{1+\exp(-Z)}^2,$$

and

$$\frac{W(x; \theta)}{1-W(x; \theta)} = \exp(-Z)$$

thus

$$h(W(x; \theta)) = \frac{1}{2} \log \left( \frac{8\pi 4c^2}{d\sigma^4} \right) - E[-Z]$$

$$- 2E[\log(1+\exp(-Z))] + \frac{d\sigma^4}{8c^2} E[Z^2]$$

$$\geq \frac{1}{2} \log \left( \frac{8\pi 4c^2}{d\sigma^4} \right) - 2 \exp[-Z] + \frac{1}{2}$$

$$= \frac{1}{2} \log \left( \frac{8\pi 4c^2}{d\sigma^4} \right) - 2 \exp \left( \frac{2c^2}{d\sigma^4} \right) + \frac{1}{2}.$$  

The entropy $h(W(x; \theta))$ is a function of $c^2$. In order to obtain the tightest possible bound from Theorem 2, we choose the basis vectors in the interpolation set $S$ to maximize the entropy. Taking the derivative and setting it equal to zero, we find that the entropy is maximized when the norm is

$$||x||^2 = W_0(1/4) \frac{d\sigma^4}{2},$$  

(22)

where $W_0(\cdot)$ is the Lambert W function. Substituting this value, we obtain

$$h(\{W(x; \theta)\}_{x \in S}) \geq \frac{d}{2} \log(4\pi W_0(1/4)) - 2 \exp(W_0(1/4)) + \frac{1}{2},$$

(23)

Combining terms and invoking Theorem 2, we obtain the following necessary relationship between the total variation and the number of training samples

$$\frac{d}{2} \log \left( \frac{n}{d\sigma^2} \right) + o(1) \geq$$

$$\frac{d}{2} \log(4\pi W_0(1/4)) - 2 \exp(W_0(1/4)) + \frac{1}{2} - d \log(e).$$

(24)

Algebraic manipulation then yields the following bound on the total variation:

$$TV(W, \hat{W}) \geq \frac{\sqrt{4\pi W_0(1/4)d\sigma^2}}{n} \times$$

$$\exp(1/2 - 2 \exp(W_0(1/4)) + o(1/d)).$$

(25)

In other words, we expect the total variation, and by extension the excess classification error, to be small when $n \gg d\sigma^2$. This result is intuitive. 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, to evaluate performance.

Next, we validate the bounds empirically. We select the signal dimension $d \in \{10, 50\}$ and the noise power $\sigma^2 \in \{0.5, 1\}$. We let the number of training samples run from $1 \leq n \leq 500$. For each combination, we average the classifier performance over $10^3$ draws of $\theta$ and $Z^n$. The learning machine computes a maximum-likelihood estimate of $\theta$ from the samples in $Z^n$, from which it estimates the posterior distribution and thus the MAP classifier. For each draw, we estimate the misclassification probability and the total variation by averaging over a further $10^3$ data samples.

In Figs. 1 and 2 we plot the empirical classification error against the bounds. Using the VC dimension, we compute the upper bound on the classification error directly. Using the total variation bounds, we compute an approximate lower bound on the classification error. Although it is not completely rigorous, we suppose that total variation is equal to, rather than a bound on, the excess classifier error. Then, we can compute the misclassification probability of the MAP classifier, add it to the bound on the total variation, and obtain an approximate on the misclassification probability of the learned classifier.

The total variation bounds are much tighter those derived from the VC dimension. They are particularly useful when the number of samples is small relative to $d$. Observe that the PAC bounds are trivial when $n$ is smaller than $d$. In their predictive regimes, the PAC bounds correctly predict the shape of the curves. For smaller $\sigma^2$, the PAC bounds are particularly loose, presumably because they consider the worst-case data distribution, whereas here the distribution is favorable for training.
By contrast, the bounds from our proposed framework are relevant for all $n$. Because they are lower bounds, they underestimate the total variation. The tightness of the bounds follows a trend opposite of the PAC bounds. When $\sigma^2$ is small, the bounds are tight, capturing the effect of the data distribution on training performance. For larger $\sigma^2$, the performance is nearer to that of the worst-case data distribution, and the PAC bounds are tighter.

5. REFERENCES