Metric Entropy and Minimax Risk in Classification

David Haussler[1] and Manfred Opper[2]

[1] Computer Science, UC Santa Cruz, CA 95064, USA
[2] Dept. of Physics, Universität Würzburg, Germany

Abstract. We apply recent results on the minimax risk in density estimation to the related problem of pattern classification. The notion of loss we seek to minimize is an information theoretic measure of how well we can predict the classification of future examples, given the classification of previously seen examples. We give an asymptotic characterization of the minimax risk in terms of the metric entropy properties of the class of distributions that might be generating the examples. We then use these results to characterize the minimax risk in the special case of noisy two-valued classification problems in terms of the Assouad density and the Vapnik-Chervonenkis dimension.

1 Introduction

The most basic problem in pattern recognition is the problem of classifying instances consisting of vectors of measurements into a one of a finite number of types or *classes*. One standard example is the recognition of isolated capital characters, in which the instances are measurements on images of letters and there are 26 classes, one for each letter. Another example is the classification of aircraft according to features extracted from their radar images. Problems of this type are called *classification problems* in statistics. In this paper we derive theoretical bounds on the best performance that can be obtained for statistical methods that perform classification.

Let us denote the entire collection of measurements for a single instance by x. We will refer to x as an *instance* or a *feature vector*. The feature vector is a random quantity that varies from instance to instance, so we will also use the random variable X, or the random variables $X_1, \ldots, X_n$ to refer to a random instance, and n independently selected random instances, respectively. We use the notation $X_i = x$ to denote that the set of measurements of the ith random instance is x. The classes in our preselected family will be numbered $\{1, \ldots, K\}$. For a given instance x, the true class of the instance will be denoted by $y \in \{1, \ldots, K\}$, or by the random variable Y. The pair (x, y), or (X, Y), will be called an *example*. A sequence of n independent random examples will be denoted $(X_1, Y_1), \ldots, (X_n, Y_n)$ or $(x_1, y_1), \ldots, (x_n, y_n)$ and abbreviated by S_n.

One simple view of a method of classification is as a function that takes as input an instance x and outputs a classification $y \in \{1, \ldots, K\}$. However, in practice one cannot be certain of one's classification, so it is better, given an instance

x, to output a probability distribution $\{\hat{P}(Y = y|X = x) : 1 \leq y \leq K\}$ that specifies the estimated probability of each of the possible classes for the instance x. We refer to this distribution as the *predictive distribution*. The predictive distribution tells not only which class is deemed most likely, but how confident the system is in that classification, and what, if any, are good alternative classifications. If, as is nearly always the case, there are different costs associated with making different kinds of misclassifications, then a separate decision making module can use the predicted probabilities produced by the classification system to decide on the optimal action to take. The theory of making optimal decisions from given these probability distributions is quite simple, and is treated fully in standard texts such as that by Duda and Hart [18], so we will not elaborate on it here. Rather we will focus solely on the problem of obtaining accurate predictive distributions, which is the critical part of the problem.

The predictive distribution can be estimated by estimating the joint probability distribution over the random variables X and Y. This joint distribution is usually broken down into a *prior distribution* $\{P(Y = y) : 1 \leq y \leq K\}$ for the values of Y, specifying which of the classes are *a priori* more likely than others, and a *generative model* that gives a conditional probability $P(X = x|Y = y)$ for each x and y, which specifies a distribution over the instances for each class. The estimated predictive probability distribution $\{\hat{P}(Y = y|X = x) : 1 \leq y \leq K\}$ is then obtained by applying Bayes rule.

When a classification method is trained to estimate the joint distribution on X and Y, a set of independent random examples $S_n = (x_1, y_1), \ldots, (x_n, y_n)$ is used. We will refer to this as the *training set*. In this process, one cannot explore all possible joint distributions. Nor would one want to explore all possible distributions, since given only a finite training set, it is impossible, using only a moderate sized training set, to pick out a good distribution from the set of all possible distributions with any kind of statistical reliability in all but trivial cases. Rather, it is up to the designer of the system to use his knowledge to pick a particular class of joint distributions on X and Y from which to choose his statistical model. We will refer to this class as Θ, and let $\theta \in \Theta$ denote a particular model in this class. Formally, each θ is the name or index for a joint probability distribution P_θ on X and Y. We will denote probabilities under the particular distribution indexed by θ by conditioning on θ. For example, the probability that $Y = y$ given that $X = x$, using the joint distribution indexed by θ, will be denoted $P(Y = y|x, \theta)$. Note that we have also abbreviated by conditioning on x only, rather than conditioning on $X = x$. We will also do this henceforth, to shorten our notation.

When assessing the performance of a classification system, there are two key issues to address: How well does the best distribution in Θ approximate the true joint distribution on X and Y, and how close does the method of estimation get to finding the best distribution in Θ. The difference between the best model in Θ and the true joint distribution is called the *approximation error*, and the difference between the model that the estimation method finds and the best model in Θ is called the *estimation error*. There is usually a tradeoff between the approximation and estimation errors: the larger one makes Θ, the more the approximation error can be reduced, but the more the estimation error increases. In order to optimize this tradeoff, it is important that the designer of a classification method have good bounds on the approximation and estimation errors for the class Θ of models he is using. There is a good general theory on approximation error, starting with the fundamental theorems of approximation theory, as given, for example, in the classic book of Lorentz [33] (see also [31, 15]). While in specific cases this error depends strongly on the nature of the true distribution, which is unknown, one can still make statements about the general approximability of functions or distributions in one family by functions or distributions in another. On the other hand, the estimation error can be analyzed somewhat independently from the true distribution, as we will show below. We focus on the estimation error in this paper. We show that rigorous bounds on the estimation error of the best possible classification systems can be established making a minimal set of assumptions.

Because we analyze estimation error, all our performance bounds will be relative to the performance of the best possible model in Θ. For this reason we will refer to Θ as the *comparison class*. If a classification method takes the training examples and produces from them an estimated model $\hat{\theta} \in \Theta$, we will ask how well does $\hat{\theta}$ perform compared to the best model θ^* in the comparison class Θ. Performance will be assessed on further random examples drawn from the same joint distribution used to generate the training examples. However, we will also consider methods that use an estimated model $\hat{\theta}$ that is not a member of the comparison class Θ in order to make their predictions. For example, Bayes methods use a weighted mixture of models in Θ to make predictions, and often this mixture does not correspond to any single model in Θ. In a Bayes method, a prior distribution over the parameter space Θ is specified. Combined with the observed examples, this prior generates a posterior distribution on Θ. The predictive distribution is then obtained by integrating over all the conditional distributions in Θ, weighted according to this posterior distribution (see e.g. [23]). Some of the most successful classification methods are Bayes methods, or computationally efficient approximations to Bayes methods. We will discuss these methods further in the last section of this paper, after we have established the basic theory of estimation error.

This paper is organized as follows: In section 2 we give a minimax definition of the estimation error for a comparison model class Θ. We use relative entropy to measure the difference between the learner's predictive distribution $\hat{P}$ and the best distribution in Θ. Then in the following three sections we develop the theoretical tools needed to determine the rate at which this estimation error converges to zero as a function of the sample size n. The main concepts used are the Hellinger distance, and the metric entropy of Θ with respect to this distance. Then in section 6 we look at the problem of cumulative minimax risk for a series of predictions made on-line by an adaptive classification method. It turns out that tighter estimates of the convergence rate of the estimation error can be made in this case using the general theory. Following this, in section

7, we compare the classification results obtained this way to the results that can be obtained using the Vapnik-Chervonenkis theory [43]. Here we restrict ourselves to a special problem of two-class classification that has been called "noisy concept learning" in the AI and computational learning theory literature [40, 1, 10, 11, 21]. We show how fairly precise, general rates can be obtained for this problem based on a combinatorial parameter known as the Assouad density [2], which is related to the VC dimension [44]. Finally, we review the implications of these results in the closing section, section 8.

The main results given here are derived from results in [28] (see also [35, 27, 22, 34]), where a general theory of minimax estimation error using relative entropy is developed that applies not only to classification problems in the form that we have defined them, but to other important statistical problems, including regression and density estimation.

There is a large statistical literature on minimax rates for estimation error for general statistical problems. However, much of this work has been done using loss functions other than relative entropy, see e.g. the texts [17, 30]. Our line of investigation, based on relative entropy, has its roots in the early work by Ibragimov and Hasminskii, who showed that the cumulative relative entropy risk for Bayes methods for parametric density estimation on the real line is approximately $(d/2)\log n$, where d is the number of parameters and n is the sample size [29]. In this case they were even able to estimate the lower order additive terms in this approximation, which involve the Fisher information and the entropy of the prior. Further related results were given by Efroimovich [20] and Clarke [12]. Clarke and Barron gave a detailed analysis, with applications, of the risk of the Bayes strategy [13], discussing the relation of the cumulative relative entropy loss to the notion of redundancy in information theory, and giving applications to hypothesis testing and portfolio selection theory. These results were extended to the cumulative relative entropy Bayes and minimax risk in [14] (see also [5]). Related lower bounds, which are often quoted, were obtained by Rissanen [37], based on certain asymptotic normality assumptions.

Estimations of the relative entropy risk in nonparametric cases were obtained in [4, 6, 38, 45, 46]. General approaches, for loss functions other than the relative entropy, to minimax risk in nonparametric density estimation were pioneered by Le Cam, who introduced methods using metric entropy and Hellinger distance (see e.g. [32]). This approach is further developed in [7, 8, 25, 41, 9, 6, 45, 42, 28]. The results of sections 2 through 5 show how this theory can be applied to the classification problem.

2 Using relative entropy and minimax risk to define estimation error

Let us summarize the problem we are considering in its abstract setting. The training data is a sequence of examples $S_n = (x_1, y_1), \ldots, (x_n, y_n)$, where each instance x_t is an element of an arbitrary set X, the *instance space* and each outcome y_t is an element of a finite set Y, the *outcome space*. (Here and below we will use X to denote both a random instance, and the instance space from which it is drawn, and similarly for Y. The usage will be clear from the context.) Given this training data and a new instance $x \in X$, a classification method produces an estimated predictive distribution

$$\hat{P}(Y = y|x, S_n)$$

that specifies the estimated probability that the outcome will be y, for each possible outcome $y \in Y$, given that the instance is x, and given the previous training examples. Note that in this notation we explicitly show the dependence of this estimated distribution on the previous training examples, whereas this dependence was implicit in the previous section. To evaluate the performance of the classification method, we have a comparison model class Θ, where each $\theta \in \Theta$ denotes a joint distribution P_θ on X and Y, here viewed as random variables. We are concerned with the estimation error, which we have defined informally as the difference between the performance of the classification method in estimating the outcome Y, and the performance of the best model in Θ. We now formalize this notion.

2.1 General setting

When assessing performance, we need a function that measures how much the predictive distribution $\hat{P}(Y = y|x, S_n)$ differs from the distribution $P(Y = y|x, \theta^*)$ produced by the best model $\theta^* \in \Theta$. While there are several functions that are often used in the literature to measure the difference between two probability distributions, the most natural one to choose here is the *relative entropy* or *Kullback-Leibler divergence*. This measure has a deep and useful information-theoretic interpretation, and it is also arises naturally in related statistical contexts, where loglikelihood ratios play a fundamental role. For two discrete probability distributions $P = (p_1, \ldots, p_K)$ and $Q = (q_1, \ldots, q_K)$, the relative entropy between P and Q is defined by

$$D_{KL}(P||Q) = \sum_{i=1}^{K} p_i \log \frac{p_i}{q_i}.$$

This quantity is nonnegative, and is 0 if and only if $P = Q$. In information theory, $-\log p_i$ is the amount of information contained in the event i under the distribution P, or equivalently, the minimum number of bits (if logarithm base 2 is used) it takes to encode the event i in the optimal (block) code based on the the distribution P is used. The relative entropy $D_{KL}(P||Q)$ is the difference between the average number of bits to encode an event when the true probability distribution is P and the optimal code based on the distribution P is used, and the average number of bits when the true distribution is P, but the optimal code based on the distribution Q is used. This is called *redundancy* in information theory. It is a measure of the regret you have at using the distribution Q to define your code, instead of the optimal (true) distribution P.

We can use the relative entropy to define a regret that is suffered if we use some nonoptimal estimate $\hat{P}(Y = y|x, S_n)$ instead of the best distribution $P(Y = y|x, \theta^*)$. The relative entropy between these two distributions is

$$\sum_y P(Y = y|x, \theta^*) \log \frac{P(Y = y|x, \theta^*)}{\hat{P}(Y = y|x, S_n)},$$

which we can write for short as $D_{KL}(P(\cdot|x, \theta^*)||\hat{P}(\cdot|x, S_n))$. The loglikelihood ratio

$$\log \frac{P(Y = y|x, \theta^*)}{\hat{P}(Y = y|x, S_n)}$$

plays a fundamental role here, and will be referred to as the *loss* for the particular prediction for outcome y. Of course, if the predictive distribution gives higher probability than the distribution P_{θ^*} does to the outcome y that actually occurs, then this loss is negative, and thus may be interpreted as a gain. The relative entropy is the average loss, assuming that the outcome y is generated at random according to the best distribution P_{θ^*}, where θ^* in Θ. This distribution P_{θ^*} is often referred to as the *true distribution*, since it plays that role in this analysis.

The average regret is called the *risk* in statistics. Here, to define the risk, we average over possible training sets S_n and possible instances x. We also assume that these are generated according to the true distribution P_{θ^*}. Thus for all $n \geq 0$ we can define the risk as

$$r_{n+1,\hat{P}}(\theta^*) = \int_{(X\times Y)^n} dP^n_{\theta^*}(S_n) \int_X dP^{(marg)}_{\theta^*}(x) D_{KL}(P(\cdot|x, \theta^*)||\hat{P}(\cdot|x, S_n)).$$

Here $\int_{(X\times Y)^n} dP^n_{\theta^*}(S_n)$ denotes expectation with respect to the random choice of the training set S_n, chosen according to the n-fold product distribution on $X \times Y$ defined by the parameter θ^*, and $\int_X dP^{(marg)}_{\theta^*}$ denotes the expectation with respect to an additional random instance x, chosen according to the marginal distribution on X defined by the parameter θ^*.

From the properties of relative entropy, the risk $r_{n,\hat{P}}(\theta^*)$ is a nonnegative number for every n, and is 0 only when the estimated distribution is the same as the true distribution (with probability 1). Once a comparison class Θ is chosen, the goal in designing a classification method $\hat{P}$ is to make the risk $r_{n,\hat{P}}(\theta^*)$ as small as possible for each n and $\theta^* \in \Theta$. However, any method $\hat{P}$ will work better for some θ^* and worse for others, so there is always some "risk" in choosing a method $\hat{P}$, since one might get a true distribution P_{θ^*} that is unfavorable for that method. To deal with this, when evaluating methods we can look at the *minimax risk*, which is defined as the minimum over all classification methods of the maximum risk over all true distributions in Θ, i.e.

$$r_n^{minimax} = r_n^{minimax}(\Theta) = \inf_{\hat{P}} \sup_{\theta^*\in\Theta} r_{n,\hat{P}}(\theta^*).$$

We define the estimation error for Θ for training samples of size n to be this minimax risk $r_n^{minimax}(\Theta)$. It represents the best possible worst case performance that can be achieved for any classification method using n training examples, when the true distribution is in Θ.

Note that since Y is finite, the minimax risk $r_n^{minimax}$ is bounded by $\log |Y|$, the logarithm of the cardinality of the outcome space Y. To see this, note that we can always set $\hat{P}$ to just predict a uniform distribution on all outcomes in Y, and in this case, no matter what the true conditional distribution on Y given x is, the regret will be at most $\log |Y|$. This is because the relative entropy from any distribution on a finite set to the uniform distribution is at most the logarithm of the cardinality of the set.

2.2 Assumption of a common marginal distribution

It is difficult to obtain an accurate and completely general analysis of the minimax risk for arbitrary Θ. However, since our interest is only in predicting Y given X, and not in predicting X itself, it is reasonable to consider a case where all the joint distributions in Θ share the same marginal distribution on X, and the only difference among the various $\theta \in \Theta$ is in the conditional distribution on Y given X. In this case each joint distribution may be decomposed into a conditional distribution on Y given X, which we denote by $P_\theta(Y = y|x)$ or $P(Y = y|x, \theta)$, and a marginal distribution on X, which we denote $P_\phi(x)$, for a new, fixed parameter ϕ, since this marginal is the same for all θ. The joint distribution on $X \times Y$ will be denoted by $P_{\theta,\phi}(x, y) = P_\theta(Y = y|x)P_\phi(x)$, or, with some abuse of notation, simply by P_θ when the common marginal distribution on X is only implicitly defined. The comparison model class itself, consisting of the set of all joint distributions $\{P_{\theta,\phi}|\theta \in \Theta\}$ will henceforth be represented by the pair (Θ, ϕ) when we wish to make the common marginal distribution on X, P_ϕ, explicit, and otherwise it will be represented simply as Θ, leaving the common marginal distribution implicitly defined. Notation for risk and minimax risk will be similarly extended to include a specific subscript for ϕ when needed. Analysis of this special case focuses attention on the conditional distributions, which are what we are really trying to learn. We restrict our attention to this case for the remainder of this paper.

Now let us consider each labeled example (x, y) as if it were a single random variable $z = (x, y)$, with distribution defined by the parameters θ and ϕ. Consider the problem of estimating the distribution $P_{\theta,\phi}(z) = P_\theta(Y = y|x)P_\phi(x)$ from a random sample $S_n = (x_1, y_1), \ldots, (x_n, y_n) = z_1, \ldots, z_n$, drawn independently according to the unknown distribution $P_{\theta,\phi}(z)$, knowing that $\theta \in \Theta$, and knowing the common marginal distribution ϕ on X. Estimating the distribution of a random variable Z from independent observations $z_1, \ldots, z_n$ is a well studied problem, which we will call the problem of *density estimation*, even if the resulting estimate is in the form of a more general probability distribution, and not a simple density. Here we have defined a special type of density estimation problem. Intuitively, since we already know the marginal distribution on X, this special type of density estimation problem should just boil down to estimating the conditional distribution on Y given X, which is the pattern recognition problem we are studying in this paper. We show this formally below. This reduction allows us to use results derived for the more general density estimation problem when analyzing the pattern recognition problem.

To see how this reduction works, first let us define the risk function for the density estimation problem as in [28]. Denote the estimate for the distribution of Z given a sample S_n by $\hat{P}(z|S_n)$. The risk in density estimation is defined to be the average relative entropy between the true distribution and the predicted distribution, i.e.

$$r^{density}_{n+1,\hat{P}}(\theta^*) = r^{density}_{n+1,\hat{P},\phi}(\theta^*) = \int_{Z^n} dP^n_{\theta^*,\phi}(S_n) D_{KL}(P_{\theta^*,\phi}(\cdot)||\hat{P}(\cdot|S_n)).$$

This is analogous to the definition of the risk $r_{n+1,\hat{P}}(\theta^*)$ defined above for the pattern recognition problem.

Now suppose that $\hat{P}(Y = y|x, S_n)$ is a predictive distribution for the pattern recognition problem. Let us define the corresponding estimate for the density estimation problem by $\hat{P}_\phi(z|S_n) = \hat{P}(Y = y|x, S_n)P_\phi(x)$. We claim that with this choice, the risk of the pattern recognition problem is the same as the risk of the density estimation problem. Indeed

$$\begin{aligned}
r^{density}_{n+1,\hat{P},\phi}(\theta^*) &= \int_{Z^n} dP^n_{\theta^*,\phi}(S_n) \int_Z dP_{\theta^*,\phi}(z) \log \frac{dP_{\theta^*,\phi}(z)}{d\hat{P}_\phi(z|S_n)} \\
&= \int_{Z^n} dP^n_{\theta^*,\phi}(S_n) \int_X \sum_y P_{\theta^*}(Y = y|x) dP_\phi(x) \log \frac{P_{\theta^*}(Y = y|x) dP_\phi(x)}{\hat{P}(Y = y|x, S_n) dP_\phi(x)} \\
&= \int_{Z^n} dP^n_{\theta^*,\phi}(S_n) \int_X dP_\phi(x) \sum_y P_{\theta^*}(Y = y|x) \log \frac{P_{\theta^*}(Y = y|x)}{\hat{P}(Y = y|x, S_n)} \\
&= r_{n+1,\hat{P},\phi}(\theta^*).
\end{aligned}$$

To complete this reduction, we define the minimax risk for the density estimation problem as in [28] by

$$r^{minimax,density}_n(\Theta) = \inf_{\hat{P}} \sup_{\theta^* \in \Theta} r^{density}_{n,\hat{P}}(\theta^*),$$

where the infimum is over all possible estimators of the joint distribution on $Z = X \times Y$. We claim that

$$r^{minimax,density}_n(\Theta) = r^{minimax}_n(\Theta), \tag{1}$$

the minimax risk for the pattern recognition defined above. Indeed, since we have shown above that we can get the same risk for all $\theta^* \in \Theta$ for both the pattern recognition and density estimation problems by the density estimator $\hat{P}_\phi(z|S_n) = \hat{P}(Y = y|x, S_n)P_\phi(x)$, it is clear that $r^{minimax,density}_n(\Theta) \leq r^{minimax}_n(\Theta)$. Now suppose we choose any density estimator $\hat{Q}(z|S_n)$. We may decompose this estimator into $\hat{Q}(z|S_n) = \hat{Q}(Y = y|x, S_n)\hat{Q}(x|S_n)$. Then by the chain rule for relative entropy ([16]), the risk for the density estimation problem can be decomposed into

$$r^{density}_{n+1,\hat{Q},\phi}(\theta^*) = r_{n+1,\hat{Q},\phi}(\theta^*) + \int_{Z^n} dP^n_{\theta^*,\phi}(S_n) \int_X dP_\phi(x) \log \frac{dP_\phi(x)}{d\hat{Q}(x|S_n)}.$$

The last term is never negative, and is zero only when $\hat{Q}(x|S_n) = P_\phi(x)$ for all S_n. In this latter case, $\hat{Q}$ is an estimator of the type used in our reduction. It follows that the minimax risk in density estimation can be obtained by restricting ourselves to such estimators, and hence $r_n^{minimax,density}(\Theta) \geq r_n^{minimax}(\Theta)$. This establishes claim (1).

3 Covering numbers and metric entropy

We now study the asymptotic properties of the minimax risk $r_n^{minimax}$ as the sample size n grows. It is easy to verify that the minimax risk $r_n^{minimax}$ is nonincreasing for all n, and in most cases approaches 0 as n goes to infinity [28]. The rate at which $r_n^{minimax}$ approaches 0 depends primarily on the metric entropy properties of Θ, the topic to which we now turn.

The theory of packing and covering numbers, and the associated metric entropy, was introduced by Kolmogorov and Tikhomirov in [31], and is commonly used in the theory of empirical processes (see e.g. [19, 36, 24, 9, 42]). For the following definitions, let (S, ρ) be any metric space.

Definition 1. (Metric entropy, also called Kolmogorov ϵ-entropy [31]) A partition Π of S is a collection $\{\pi_i\}$ of subsets of S that are pairwise disjoint and whose union is S. The diameter of a set $A \subseteq S$ is given by $\mathrm{diam}(A) = \sup_{x,y \in A} \rho(x, y)$. The diameter of a partition is the supremum of the diameters of the sets in the partition. For $\epsilon > 0$, by $\mathcal{D}_\epsilon(S, \rho)$ we denote the cardinality of the smallest finite partition of S of diameter at most ϵ, or ∞ if no such finite partition exists. The metric entropy of (S, ρ) is defined by

$$\mathcal{K}_\epsilon(S, \rho) = \log \mathcal{D}_\epsilon(S, \rho).$$

We say S is *totally bounded* if $\mathcal{D}_\epsilon(S, \rho) < \infty$ for all $\epsilon > 0$.

Definition 2. (Packing and covering numbers) For $\epsilon > 0$, an *ϵ-cover* of S is a subset $A \subseteq S$ such that for all $x \in S$ there exists a $y \in A$ with $\rho(x, y) \leq \epsilon$. By $\mathcal{N}_\epsilon(S, \rho)$ we denote the cardinality of the smallest finite ϵ-cover of S, or ∞ if no such finite cover exists. For $\epsilon > 0$, an *ϵ-separated subset of S* is a subset $A \subseteq S$ such that for all distinct $x, y \in A$, $\rho(x, y) > \epsilon$. By $\mathcal{M}_\epsilon(S, \rho)$ we denote the cardinality of the largest finite ϵ-separated subset of S, or ∞ if arbitrarily large such sets exist.

The following lemma is easily verified [31].

Lemma 3. *For any $\epsilon > 0$,*

$$\mathcal{M}_{2\epsilon}(S, \rho) \leq \mathcal{D}_{2\epsilon}(S, \rho) \leq \mathcal{N}_\epsilon(S, \rho) \leq \mathcal{M}_\epsilon(S, \rho).$$

It follows that the metric entropy $\mathcal{K}_\epsilon$ (and the condition defining total boundedness) can also be defined using either the packing or covering numbers in place of $\mathcal{D}_\epsilon$, to within a constant factor in ϵ.

Kolmogorov and Tikhomirov also introduced abstract notions of the dimension and order of metric spaces in their seminal paper [31]. These can be used to measure the "massiveness" of both spaces indexed by finite dimensional parameter vectors and infinite dimensional function spaces. In the following, the metric ρ is omitted from the notation, being understood from the context.

Definition 4. The *upper* and *lower metric dimensions* [31] of S are defined by

$$\overline{\mathbf{dim}}(S) = \limsup_{\epsilon \to 0} \frac{\mathcal{K}_\epsilon(S)}{\log \frac{1}{\epsilon}}$$

and

$$\underline{\mathbf{dim}}(S) = \liminf_{\epsilon \to 0} \frac{\mathcal{K}_\epsilon(S)}{\log \frac{1}{\epsilon}},$$

respectively. When $\overline{\mathbf{dim}}(S) = \underline{\mathbf{dim}}(S)$, then this value is denoted $\mathbf{dim}(S)$ and called the *metric dimension* of S. Thus

$$\mathbf{dim}(S) = \lim_{\epsilon \to 0} \frac{\mathcal{K}_\epsilon(S)}{\log \frac{1}{\epsilon}}.$$

For totally bounded S, we say that S is finite dimensional if $\mathbf{dim}(S) < \infty$, else it is infinite dimensional. To measure the massiveness of infinite dimensional spaces, including typical function spaces, further indices were introduced by Kolmogorov and Tikhomirov. The *functional dimension* of S is defined similarly as

$$\mathbf{df}(S) = \lim_{\epsilon \to 0} \frac{\log \mathcal{K}_\epsilon(S)}{\log \log \frac{1}{\epsilon}},$$

with similar upper and lower versions, $\overline{\mathbf{df}}$ and $\underline{\mathbf{df}}$, when this limit does not exist. Finally, the *metric order* of S is defined as

$$\mathbf{mo}(S) = \lim_{\epsilon \to 0} \frac{\log \mathcal{K}_\epsilon(S)}{\log \frac{1}{\epsilon}},$$

with similar upper and lower versions, $\overline{\mathbf{mo}}$ and $\underline{\mathbf{mo}}$.

4 Hellinger distance

We can view the comparison model class Θ as a metric space, and calculate its metric entropy, by specifying a metric on this space. It turns out that the right metric to use is the *Hellinger distance*. If $P = (p_1, \ldots, p_k)$ and $Q = (q_1, \ldots, q_k)$ are two discrete probability distributions, then the Hellinger distance between P and Q is defined as

$$D_{HL}(P, Q) = \left(\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2 \right)^{1/2}.$$

That is, the Hellinger distance between P and Q is the Euclidean distance between $(\sqrt{p_1}, \ldots, \sqrt{p_k})$ and $(\sqrt{q_1}, \ldots, \sqrt{q_k})$. The Hellinger distance can be generalized to discrete distributions on countably infinite sets, and on continuous sets such as the real line, by using l_2 and L_2 norms, respectively, in place of Euclidean distance. The Hellinger distance is useful because it is a metric, and the squared Hellinger distance approximates the relative entropy distance, which is not a metric. The sense of this approximation is given, e.g., in [28]. This metric has been used to give bounds on the risk of estimation procedures in statistics by many authors, including LeCam [32], Birgé [7, 8], Hasminskii and Ibragimov [25], and van de Geer [41].

Now assume that θ and θ^* are two joint distributions on $X \times Y$ with a common marginal distribution on X. Then, when X is discrete, the Hellinger distance between these two distributions is

$$D_{HL}(\theta, \theta^*) = \left(\sum_{x,y} \left(\sqrt{P(x)P(y|x,\theta)} - \sqrt{P(x)P(y|x,\theta^*)} \right)^2 \right)^{1/2}$$
$$= \left(\sum_{x} P(x) \sum_{y} \left(\sqrt{P(y|x,\theta)} - \sqrt{P(y|x,\theta^*)} \right)^2 \right)^{1/2}.$$

This extends naturally to continuous X as well. Using this distance, if all the distributions in Θ are distinct (i.e. differ on a set of positive measure), which we may assume without loss of generality, then (Θ, D_{HL}) is a metric space, else it is a *pseudo* metric space, i.e. a metric space that possibly includes distinct points at distance 0.

5 Rates for minimax risk

We are now in a position to state the main theorem about rates for minimax risk. Let us define the *best exponent in the rate for the minimax risk* by

$$e(\Theta) = \sup\{t : \limsup_{n \to \infty} \frac{r_n^{minimax}(\Theta)}{n^{-t}} \leq 1\}.$$

By calculating this best exponent, we can distinguish the various rates at which the minimax risk approaches 0.

Theorem 5. *Assume Θ is a comparison model class in which all models have a common marginal distribution on X. Then the bounds on $e(\Theta)$ given in the following table are valid.*

size of Θ	*bound on exponent*
Θ *is finite*	$e(\Theta) = \infty$
$\mathbf{dim}(\Theta, D_{HL}) = 0$	$e(\Theta) \geq 1$
$\mathbf{dim}(\Theta, D_{HL}) = D$ *where* $0 < D < \infty$	$e(\Theta) = 1$
$\mathbf{df}(\Theta, D_{HL}) = \beta$ *where* $1 < \beta < \infty$	$e(\Theta) = 1$
$\mathbf{mo}(\Theta, D_{HL}) = \alpha$ *where* $0 < \alpha < \infty$	$e(\Theta) = \frac{2}{2+\alpha}$
$\mathbf{mo}(\Theta, D_{HL}) = \infty$	$e(\Theta) = 0$
(Θ, D_{HL}) *not totally bounded*	$e(\Theta) = 0$

Proof. This follows directly from Theorem 7 in [28], using claim (1). To see that the conditions of Theorem 7 in [28] hold, suppose $|Y| = K$ and let $\hat{P}(Y = y|x) = 1/K$ for all $x \in X$. Then note that in the density estimation problem that corresponds to the pattern recognition problem under consideration, because of the common marginal distribution, for any $\lambda > 0$ and any θ^*,

$$\begin{aligned}
\int (dP_{\theta^*,\phi})^{1+\lambda}(d\hat{P}_\phi)^{-\lambda} &= \int_X dP_\phi(x) \sum_y (P_{\theta^*}(Y = y|x))^{1+\lambda}(\hat{P}(Y = y|x))^{-\lambda} \\
&= K^\lambda \int_X dP_\phi(x) \sum_y (P_{\theta^*}(Y = y|x))^{1+\lambda} \\
&\leq K^\lambda \\
&< \infty.
\end{aligned}$$

Thus the minimax risk for the regret function $\int (dP_{\theta^*,\phi})^{1+\lambda}(d\hat{P}_\phi)^{-\lambda}$ is finite as required.

Thus for a comparison model class Θ of finite dimension or finite functional dimension, the rate of convergence of the minimax risk (i.e. estimation error) to zero as a function of sample size n is better than $\frac{1}{n^{1-\delta}}$ for all positive δ, but for (larger) model classes of finite metric order α, the best rate is something like $\frac{1}{n^{2/(2+\alpha)}}$, which is much slower for large metric order α. Going to further extremes, convergence is faster than any inverse polynomial for finite model classes Θ (it can be shown to be exponential in n [28]), but model classes of infinite metric order, or that are not even totally bounded, are essentially "unlearnable" with this definition of estimation error as minimax risk: for any learning method there is a choice of true distribution that makes the convergence slower than $\frac{1}{n^\delta}$ for all positive δ. The advantage of this theorem is that is gives a characterization of the best possible rate of convergence entirely in terms of the metric entropy of the model class Θ, without referring to any specific properties of the models themselves. However, it does not give the most precise convergence rates that can be stated for many common cases, especially finite dimensional ones. This is addressed in the following sections.

6 Rates for cumulative minimax risk

More precise bounds on the rate of convergence for the minimax risk can be obtained if we look at the cumulative risk. This is the total minimax average regret (risk) for the first n predictions in a sequential or *on line* prediction setting, in which the examples $(x_1, y_1), \ldots, (x_n, y_n)$ are presented to the learner one at a time, and for each t between 1 and n, after seeing the first $t-1$ examples $S_{t-1} = (x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ and the tth instance x_t, the learner must produce an estimated predictive probability distribution $\hat{P}(Y_t = y|x_t, S_{t-1})$. The loss in this sequential version of the prediction game is

$$\sum_{t=1}^{n} \log \frac{P(Y_t = y_t|x_t, \theta^*)}{\hat{P}(Y_t = y_t|x_t, S_{t-1})},$$

where P_{θ^*}, with $\theta^* \in \Theta$, is the true distribution. So the cumulative loss is simply the total loss for all individual predictions. When we average this over possible sequences of examples generated independently according to P_{θ^*}, we get the cumulative regret

$$R_{n,\hat{P}}(\theta^*) = \int_{(X \times Y)^n} dP^n_{\theta^*}(S_n) \sum_{t=1}^{n} \log \frac{P(Y_t = y_t|x_t, \theta^*)}{\hat{P}(Y_t = y_t|x_t, S_{t-1})}.$$

It is easily verified, using the linearity of expectation, that

$$R_{n,\hat{P}}(\theta^*) = \sum_{t=1}^{n} r_{t,\hat{P}}(\theta^*).$$

So the cumulative regret for the first n predictions is just the sum of the regrets for each of the predictions for sample sizes t between 1 and n. Finally, just as before, the *cumulative minimax risk* is defined as the minimum over all classification methods of the maximum cumulative regret over all true distributions in Θ, i.e.

$$R_n^{minimax} = R_n^{minimax}(\Theta) = \inf_{\hat{P}} \sup_{\theta^* \in \Theta} R_{n,\hat{P}}(\theta^*).$$

Looking at the cumulative minimax risk provides an alternate way to study the estimation error and its rate of convergence. When comparing the cumulative minimax risk to the minimax risk defined above, called the *instantaneous* minimax risk in [28] to contrast it with the cumulative minimax risk, note that from the definition of the individual minimax risks $r_t^{minimax}$, for $1 \leq t \leq n$, we see that for each separate t we are possibly looking at a different worst case true distribution P_{θ^*} when we compute the minimax risk $r_t^{minimax}$, whereas for the cumulative minimax risk $R_n^{minimax}$, the same true distribution P_{θ^*} must be used for all $1 \leq t \leq n$. Thus the cumulative minimax risk is in some ways a better measure of the sustained difficulty of the learning/prediction problem over a range of sample sizes, while the instantaneous minimax risk for a particular sample size n could in principle reflect the difficulty of the problem due to particular distributions in Θ that are "hard" for that particular sample size n. However, it turns out that this effect cannot be very strong. In particular, it can be shown in general that $R_n^{minimax}$ is nondecreasing in n, and

$$\sum_{t=1}^{n} r_t^{minimax} \geq R_n^{minimax} \geq n r_n^{minimax}$$

(see [3, 13, 6, 28].) It follows that $R_n^{minimax}$ grows at most linearly in n, since $r_t^{minimax} \leq \log |Y|$ for all t. These inequalities also give fairly tight bounds on $R_n^{minimax}$ in terms of $r_n^{minimax}$ when $r_n^{minimax}$ decreases slowly. For example[3], if $r_n^{minimax} \asymp 1/\sqrt{n}$, then $R_n^{minimax} \asymp \sqrt{n}$. However, if $r_n^{minimax} = D/n$ then the inequalities only tell us that $D \leq R_n^{minimax} \leq D \sum_{t=1}^{n} 1/t \leq D \log(n+1)$. We get a more precise analysis of the cumulative minimax risk $R_n^{minimax}$ by bounding it directly in terms of the metric entropy of Θ, as in the results on the minimax risk in the previous section. (Actually, the results on minimax risk are derived from the results on cumulative minimax risk given here.)

Theorem 6. *Assume Θ is a comparison model class in which all models have a common marginal distribution on X. Then*

1. *If Θ is finite then*
$$R_n^{minimax}(\Theta) \to \log |\Theta| \text{ as } n \to \infty.$$
2. *If $\mathbf{dim}(\Theta, D_{HL}) = 0$ then*
$$R_n^{minimax}(\Theta) \in o(\log n).$$
3. *If $\mathbf{dim}(\Theta, D_{HL}) = D$ where $0 < D < \infty$ then*
$$R_n^{minimax}(\Theta) \sim \frac{D}{2} \log n.$$
4. *If $\mathbf{df}(\Theta, D_{HL}) = \beta$ where $1 < \beta < \infty$ then*
$$\log R_n^{minimax}(\Theta) \sim \beta \log \log n.$$
5. *If $\mathbf{mo}(\Theta, D_{HL}) = \alpha$ where $0 < \alpha < \infty$ then*
$$\log R_n^{minimax}(\Theta) \sim \frac{\alpha}{2+\alpha} \log n.$$
6. *If $\mathbf{mo}(\Theta, D_{HL}) = \infty$ or (Θ, D_{HL}) is not totally bounded, then*
$$\log R_n^{minimax}(\Theta) \sim \log n.$$

Proof. Similar to the proof of Theorem 5, but using Theorem 4 of [28].

Furthermore, analogous results using upper and lower dimensions and orders also hold in the situation when the upper and lower dimensions/orders are different. For example, we can show

[3] For integer or real-valued functions f and g, we say $f \sim g$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$, $f \asymp g$ if $\liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0$ and $\limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty$.

Theorem 7. *Assume Θ is a comparison model class in which all models have a common marginal distribution on X. Then*

$$\limsup_{n\to\infty} \frac{R_n^{minimax}(\Theta)}{\log n} = \frac{\overline{\mathbf{dim}}(\Theta, D_{HL})}{2}$$

and

$$\liminf_{n\to\infty} \frac{R_n^{minimax}(\Theta)}{\log n} = \frac{\underline{\mathbf{dim}}(\Theta, D_{HL})}{2}.$$

Proof. Let $R_n^{minimax} = R_n^{minimax}(\Theta)$ and $\mathcal{K}(\epsilon) = \mathcal{K}_\epsilon(\Theta, D_{HL})$. By Lemma 7 of [28], there is some positive constant c such that for any n and any $\epsilon > 0$,

$$\min\{\mathcal{K}(\epsilon), n\epsilon^2/8\} - \log 2 \le R_n^{minimax} \le \mathcal{K}(\epsilon) + c\epsilon^2 n \log n + c.$$

Here we verify the conditions of the lemma again as in the proof of Theorem 5. Now, in the lower bound let $\epsilon = \frac{\log n}{\sqrt{n}}$ and in the upper bound let $\epsilon = \frac{1}{\sqrt{n}\log n}$. Note that if $\overline{\mathbf{dim}}(\Theta, D_{HL}) < \infty$ then $\mathcal{K}(\epsilon)$ is $O(\log(1/\epsilon))$, so $\min\{\mathcal{K}(\epsilon), n\epsilon^2/8\} = \mathcal{K}(\epsilon)$ for large n if $\epsilon = \frac{\log n}{\sqrt{n}}$. It follows that

$$\limsup_{n\to\infty} \frac{\mathcal{K}(\frac{\log n}{\sqrt{n}})}{\log n} \le \limsup_{n\to\infty} \frac{R_n^{minimax}}{\log n} \le \limsup_{n\to\infty} \frac{\mathcal{K}(\frac{1}{\sqrt{n}\log n})}{\log n}.$$

Looking at the upper bound, let $m = \sqrt{n}\log n$. Then

$$\limsup_{n\to\infty} \frac{\mathcal{K}(\frac{1}{\sqrt{n}\log n})}{\log n} = \limsup_{m\to\infty} \frac{\mathcal{K}(\frac{1}{m})}{\log \frac{m^2}{f(m)}},$$

where $f(m) \asymp \log^2 m$. Hence

$$\limsup_{m\to\infty} \frac{\mathcal{K}(\frac{1}{m})}{\log \frac{m^2}{f(m)}} = \frac{1}{2} \limsup_{m\to\infty} \frac{\mathcal{K}(\frac{1}{m})}{\log m} = \frac{\overline{\mathbf{dim}}(\Theta, D_{HL})}{2}.$$

Looking at the lower bound, a similar argument shows

$$\limsup_{n\to\infty} \frac{\mathcal{K}(\frac{\log n}{\sqrt{n}})}{\log n} = \frac{\overline{\mathbf{dim}}(\Theta, D_{HL})}{2}.$$

This establishes the first part of the theorem. The second part is established in a similar manner.

7 Vapnik-Chervonenkis entropy and dimension

In this section we examine how the results given here relate to results that can be obtained by another approach that has often been used to analyze the convergence rates of classification methods, namely, the Vapnik-Chervonenkis dimension [44]. For simplicity, in this comparison we restrict ourselves to a special class of classification problems that we call *noisy two-class learning* (also called "noisy concept learning" in the computational learning theory and AI machine learning literature [1]). In noisy two-class learning, the outcome space Y has just two values, which we may designate as $+1$ and -1, instead of an arbitrary finite set of values, as we have been assuming up to this point. Furthermore, not only do the joint distributions in Θ all have the same marginal distribution on X, but the conditional distributions on Y given X all have a special form described as follows:

It is assumed that there is a fixed *noise rate* $0 < \lambda < 1/2$, and for each distribution $\theta \in \Theta$ there is a function $f_\theta : X \to Y$ such that for all instances $x \in X$,

$$P(Y \neq f_\theta(x)|x, \theta) = \lambda.$$

You can view this conditional distribution as being generated by an underlying functional relationship between X and Y, namely, $Y = f_\theta(X)$, composed with an independent noise process that flips the sign of Y independently with probability λ. Thus in this case our examples $(x_1, y_1), \ldots, (x_n, y_n)$ are really a noise corrupted version of an underlying set of random examples $(x_1, f_\theta(x_1)), \ldots, (x_n, f_\theta(x_n))$ of the function f_θ, for some unknown $\theta \in \Theta$. The instances $x_1, \ldots, x_n$ are generated independently at random according to the marginal distribution ϕ on X.

Let us define $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$. Vapnik-Chervonenkis theory provides a way of bounding the estimation error of Θ in terms of certain combinatorial properties of the class of functions $\mathcal{F}_\Theta$. The key element of this theory is the *growth function*. For the following definitions, let $\mathcal{F}$ be a family of $\{\pm 1\}$-valued functions on a set X.

Definition 8. For each sequence $x^n = x_1, \ldots, x_n$ in X^n, let $\mathcal{F}_{|x^n} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$. The growth function $\Pi_{\mathcal{F}}(n)$ is defined by

$$\Pi_{\mathcal{F}}(n) = \max_{x^n \in X^n} |\mathcal{F}_{|x^n}|,$$

where $|S|$ denotes the cardinality of the set S. Thus $\Pi_{\mathcal{F}}(n)$ is the maximum number of distinct functions that can be obtained by restricting the domain of the functions in $\mathcal{F}$ to n points.

From the growth function we can define the Assouad density of $\mathcal{F}$ [2], and the Vapnik-Chervonenkis (VC) dimension of $\mathcal{F}$. We treat the Assouad density first, relating it to a certain supremum over the metric dimension, and return to the VC dimension later.

Definition 9. The *Assouad density* of $\mathcal{F}$ is defined by

$dens(\mathcal{F}) = \inf\{d > 0 :$ there exists $C > 0$ such that for all $n \geq 1, \Pi_{\mathcal{F}}(n) \leq Cn^d\}$.

It is easily verified that

$$dens(\mathcal{F}) = \limsup_{n \to \infty} \frac{\log \Pi_{\mathcal{F}}(n)}{\log n} \tag{2}$$

To see this, note that if $r > \limsup_{n\to\infty} \frac{\log \Pi_{\mathcal{F}}(n)}{\log n}$, then there exists $r_0 < r$ and n_0 such that for all $n \geq n_0$, $\frac{\log \Pi_{\mathcal{F}}(n)}{\log n} \leq r_0$, which implies $\Pi_{\mathcal{F}}(n) \leq n^{r_0}$. Hence $r > dens(\mathcal{F})$. On the other hand, if $r < \limsup_{n\to\infty} \frac{\log \Pi_{\mathcal{F}}(n)}{\log n}$ then there exists $r_0 > r$ such that $\Pi_{\mathcal{F}}(n) > n^{r_0}$ infinitely often, and thus $r < dens(\mathcal{F})$. Equation (2) follows.

Now let P_ϕ be the probability distribution on X. For $f, g \in \mathcal{F}_\Theta$, define $D_\phi(f,g) = P_\phi(f(x) \neq g(x))$. Then $(\mathcal{F}_\Theta, D_\phi)$ is a (pseudo) metric space. This metric space is related to the metric space (Θ, D_{HL}) that was central to the results in the previous sections when Θ is the model class for a noisy two-class learning problem. Let the noise rate be λ and let $c_\lambda = 2(\sqrt{\lambda} - \sqrt{1-\lambda})^2$. Then

$$D_{HL}^2(\theta, \theta^*) = \int_X dP_\phi(x) \sum_{y \in Y} \left(\sqrt{P(Y=y|x,\theta)} - \sqrt{P(Y=y|x,\theta^*)}\right)^2 = c_\lambda D_\phi(f_\theta, f_{\theta^*}).$$

Hence the metric entropies of these two spaces are related by

$$\mathcal{K}_\epsilon(\Theta, D_{HL}) = \mathcal{K}_{\epsilon^2/c_\lambda}(\mathcal{F}_\Theta, D_\phi),$$

and thus if Θ is finite dimensional,

$$\mathbf{dim}(\Theta, D_{HL}) = 2\mathbf{dim}(\mathcal{F}_\Theta, D_\phi), \tag{3}$$

and similarly for $\overline{\mathbf{dim}}$ and $\underline{\mathbf{dim}}$. Similar relations can be derived when Θ is infinite dimensional. In this manner, for the noisy two-class learning problem, the results of the previous sections can be restated in terms of the scaling as $\epsilon \to 0$ of the metric entropy of the metric space $(\mathcal{F}_\Theta, D_\phi)$, rather than the metric space (Θ, D_{HL}). A result of Assouad's, given in a monograph by Dudley [19], relates this scaling, in the worst case over distributions P_ϕ, to the Assouad density of $\mathcal{F}_\Theta$. For the following definitions and results, let $\mathcal{F}$ be any class of $\{\pm 1\}$-valued functions on a set X and P_ϕ be any distribution on X.

Definition 10. Let

$s(\mathcal{F}) = \inf\{d > 0 :$ there is a $C > 0$ such that for every P_ϕ and $0 < \epsilon \leq 1, \mathcal{M}_\epsilon(\mathcal{F}, D_\phi) \leq C\epsilon^{-d}\}$.

Theorem 11. *(Theorem 9.3.1 of [19])*

$$dens(\mathcal{F}) = s(\mathcal{F})$$

Using a similar method, we can also relate $s(\mathcal{F})$ directly to the upper dimension of $(\mathcal{F}, D_\phi)$.

Theorem 12.

$$s(\mathcal{F}) = \limsup_{\epsilon \to 0} \sup_{P_\phi} \frac{\mathcal{K}_\epsilon(\mathcal{F}, D_\phi)}{\log \frac{1}{\epsilon}} = \sup_{P_\phi} \limsup_{\epsilon \to 0} \frac{\mathcal{K}_\epsilon(\mathcal{F}, D_\phi)}{\log \frac{1}{\epsilon}} = \sup_{P_\phi} \overline{\mathbf{dim}}(\mathcal{F}, D_\phi)$$

Proof. The first equality is similar to (2), and the last equality follows directly from the definition of $\overline{\mathbf{dim}}(\mathcal{F}, D_\phi)$. So we need only consider the middle equality. Let $l = \limsup_{\epsilon \to 0} \sup_{P_\phi} \frac{\mathcal{K}_\epsilon(\mathcal{F}, D_\phi)}{\log \frac{1}{\epsilon}}$ and $u = \sup_{P_\phi} \limsup_{\epsilon \to 0} \frac{\mathcal{K}_\epsilon(\mathcal{F}, D_\phi)}{\log \frac{1}{\epsilon}}$. For any function $f(n, m)$,

$$\limsup_{n} \sup_{m} f(n, m) \geq \sup_{m} \limsup_{n} f(n, m),$$

so it suffices to show that $l \leq u$. As this inequality is trivial when $l = 0$, we will assume $0 < l \leq \infty$. Let $\{\epsilon_n\}_{n \geq 1}$ be a sequence of positive numbers such that $\epsilon_n \leq 2^{-n}$ and $\{\phi_n\}_{n \geq 1}$ be a sequence of distributions on X such that

$$l = \lim_{n \to \infty} \frac{\log \mathcal{M}_{\epsilon_n}(\mathcal{F}, D_{\phi_n})}{\log \frac{1}{\epsilon_n}}$$

Using Lemma 3 it is clear that such sequences can be found. Suppose $0 < r < t < l$, and $t < \infty$. Let the distribution P_ϕ be defined by

$$dP_\phi(x) = \frac{1}{S} \sum_{n=1}^{\infty} n^{-t/r} dP_{\phi_n}(x),$$

where $S = \sum_{n=1}^{\infty} n^{-t/r} < \infty$. We claim that

$$r \leq \limsup_{\epsilon \to 0} \frac{\mathcal{K}_\epsilon(\mathcal{F}, D_\phi)}{\log \frac{1}{\epsilon}}.$$

Since r can be chosen arbitrarily close to l, or arbitrarily large if $l = \infty$, and the right hand side above is less than or equal to u, this shows that $l \leq u$.

To see that this claim holds, first note that for any set $A \subseteq X$ and any n, $P_\phi(A) \geq \frac{P_{\phi_n}(A)}{S n^{t/r}}$. Thus any ϵ-separated set of $\mathcal{F}$ under the metric D_{ϕ_n} is an $\frac{\epsilon}{S n^{t/r}}$-separated set under the metric D_ϕ. For each n let $M(n) = \mathcal{M}_{\epsilon_n}(\mathcal{F}, D_{\phi_n})$. Now set $\gamma_n = \frac{\epsilon_n}{S n^{t/r}}$. It follows that $\mathcal{M}_{\gamma_n}(\mathcal{F}, D_\phi) \geq M(n)$. Since $\lim_{n \to \infty} \frac{\log M(n)}{\log \frac{1}{\epsilon_n}} = l > t$, there is an n_0 such that for all $n \geq n_0$, $M(n) > \epsilon_n^{-t}$. Hence for large n, $\mathcal{M}_{\gamma_n}(\mathcal{F}, D_\phi) \geq M(n) > \epsilon_n^{-t}$. However, $\gamma_n \to 0$, and

$$\epsilon_n^{-t} = \epsilon_n^{-r} \epsilon_n^{r-t} = \gamma_n^{-r} S^{-r} n^{-t} \epsilon_n^{r-t} > \gamma_n^{-r}$$

for large n, since $\epsilon_n \leq 2^{-n}$ and $r - t < 0$. It follows that $\mathcal{M}_{\gamma_n}(\mathcal{F}, D_\phi) > \gamma_n^{-r}$ for large n, and hence

$$r \leq \limsup_{\epsilon \to 0} \frac{\mathcal{K}_\epsilon(\mathcal{F}, D_\phi)}{\log \frac{1}{\epsilon}}.$$

This establishes the claim.

As a corollary of Theorems 11 and (12), we have

$$dens(\mathcal{F}) = \sup_{P_\phi} \overline{\mathbf{dim}}(\mathcal{F}, D_\phi). \tag{4}$$

It should be noted that this relationship between the growth rate of the maximum size of $\mathcal{F}$ restricted to n points and the metric entropy of $(\mathcal{F}, D_\phi)$ requires that one take the supremum over all distributions P_ϕ. If one does not, then there is no close relationship between these two quantities, even if we use the expected size of $\mathcal{F}$ restricted to n random points. For example, if we let $X = [0,1]$, P_ϕ be the uniform distribution on X and $\mathcal{F}$ be the set of all $\{\pm 1\}$-valued functions that are $+1$ on at most d points, then $\int_{X^n} dP_\phi^n(x^n)|\mathcal{F}_{|x^n}| \asymp n^d$ but $\mathcal{K}_\epsilon(\mathcal{F}, D_\phi) = 0$, since all functions differ only on a set of measure 0, and hence are distance 0 apart under the metric D_ϕ.

The Assouad density is the exponent of the smallest polynomial function that upper bounds the growth function $\Pi_{\mathcal{F}}(n)$. The growth function has a curious combinatorial property: either it is bounded by some polynomial in n, and hence the Assouad density is finite, or it is equal to 2^n for all n (and hence the Assouad density is most decidedly infinite). In fact, if we let $\mathbf{dim}_{VC}(\mathcal{F})$ be the largest n such that $\Pi_{\mathcal{F}}(n) = 2^n$, then if $\mathbf{dim}_{VC}(\mathcal{F}) = d < \infty$, then $\Pi_{\mathcal{F}}(n) \le \sum_{i=0}^{d} \binom{n}{i} \le (en/d)^d$ for all $n \ge d \ge 1$. This result, often cited as Sauer's Lemma [39], was proven independently by Vapnik and Chervonenkis [44] (first in a slightly weaker version). $\mathbf{dim}_{VC}(\mathcal{F})$ is called the *VC dimension* of $\mathcal{F}$. It follows that

$$dens(\mathcal{F}) \le \mathbf{dim}_{VC}(\mathcal{F}). \tag{5}$$

This inequality is often tight, but not always tight. Indeed, for any finite $\mathcal{F}$, $dens(\mathcal{F}) = 0$, yet there are finite $\mathcal{F}$ with arbitrarily large VC dimension. However, for all $\mathcal{F}$, $dens(\mathcal{F})$ is finite if and only if $\mathbf{dim}_{VC}(\mathcal{F})$ is finite.

Finally, we can put the above results together with the results of the previous section to obtain the following characterization of the estimation error for noisy two-class learning problems, defined as the cumulative minimax risk $R_n^{minimax}(\Theta, \phi)$.

Theorem 13. *If Θ is any set of conditional distributions for the noisy two-class learning problem with any noise rate $0 < \lambda < 1/2$, then if $\mathbf{dim}_{VC}(\mathcal{F}_\Theta)$ is finite we have*

$$\sup_{P_\phi} \limsup_{n\to\infty} \frac{R_n^{minimax}(\Theta, \phi)}{\log n} = dens(\mathcal{F}_\Theta) \le \mathbf{dim}_{VC}(\mathcal{F}_\Theta)$$

and if $\mathbf{dim}_{VC}(\mathcal{F}_\Theta)$ is infinite then $\sup_{P_\phi} R_n^{minimax}(\Theta, \phi)$ grows linearly in n.

Proof. If $\mathbf{dim}_{VC}(\mathcal{F}_\Theta)$ is finite then using Theorem 7, the $\overline{\mathbf{dim}}$ version of Equation (3), Equation (4), and Equation (5) in that order, we have

$$\begin{aligned} \sup_{P_\phi} \limsup_{n\to\infty} \frac{R_n^{minimax}(\Theta, \phi)}{\log n} &= \sup_{P_\phi} \overline{\mathbf{dim}}(\mathcal{F}_\Theta, D_\phi) \\ &= dens(\mathcal{F}_\Theta) \\ &\le \mathbf{dim}_{VC}(\mathcal{F}_\Theta). \end{aligned}$$

If $\mathbf{dim}_{VC}(\mathcal{F}_\Theta)$ is infinite then for any n we can choose a distribution P_ϕ that is uniform on a large finite set $X_0 \subseteq X$ that is *shattered* in the sense that $|\mathcal{F}_{|X_0}| = 2^{|X_0|}$. Suppose a function f is chosen uniformly at random from $\mathcal{F}_{|X_0}$. If $|X_0|$ is large enough then the first n instances $x_1, \ldots, x_n$ will be distinct with probability near 1, and all labelings of these points with ± 1 values $y_1, \ldots, y_n$ will be equally likely to occur. Under such conditions, the average instantaneous regret in predicting y_t given $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ and x_t is a positive constant for all $1 \leq t \leq n$ for any noise rate $0 < \lambda < 1/2$ and any prediction method, so $R_n^{minimax}(\Theta, \phi)$ grows linearly in n. Since $R_n^{minimax}(\Theta, \phi)$ cannot grow faster than linear in n for any ϕ, as was remarked in Section 6, it follows that $\sup_{P_\phi} R_n^{minimax}(\Theta, \phi)$ grows linearly in n.

This theorem shows that $\sup_{P_\phi} R_n^{minimax}(\Theta, \phi)$ either grows logarithmically or slower, or it grows linearly. There is no rate in between. Results of this type are also available from the standard Vapnik-Chervonenkis theory [44, 11]. However, what is novel here is that in the case of logarithmic growth, the best possible constant in front of the logarithm is identified here to be the Assouad density. It is difficult to identify such constants with the standard Vapnik-Chervonenkis theory, which relies on uniform convergence of empirical estimates, and therefore gives only indirect bounds on the minimax risk.

Some tighter upper bounds are known for Theorem 13. In particular, in [26] it was shown that

Theorem 14. *If Θ is any set of conditional distributions for the noisy two-class learning problem with any noise rate $0 < \lambda < 1/2$, and $\mathcal{F} = \mathcal{F}_\Theta$, then for all marginal distributions P_ϕ on X*

$$R_n^{minimax}(\Theta, \phi) \leq \int_{X^n} dP_\phi^n(x^n) \log |\mathcal{F}_{|x^n}| \leq \log \Pi_{\mathcal{F}}(n).$$

It is an open problem to obtain tighter lower bounds.

8 Conclusions

We have looked at the performance of the best possible classification method in terms of the minimax relative entropy risk, obtained by comparing the predictive distribution produced by the particular classification method to the best possible distribution in a comparison model class Θ. We are able to characterize the best performance that can be achieved in terms of the metric entropy of the model class Θ.

One important question that remains is: what classification method gives this best possible performance? Recall that a Bayes method is one that employs a prior distribution over the model class Θ, and computes its predictive distribution by averaging over all conditional distributions $P(Y|x, \theta)$, weighted according to the posterior probability of θ given the training examples. It turns out that by careful choice of the prior, we can find Bayes methods that get asymptotically close to the best minimax performance. The priors to use can be found by examining the proof of the lower bound given in Lemma 7 of [28], upon which the lower bound in the result given in Theorem 6 above is based. These place a uniform prior distribution on a finite subset of Θ, chosen to be a maximal ϵ-separated subset with respect to the Hellinger distance for some suitable ϵ. The best value of ϵ to use decreases as the sample size n grows. This idea is quite intuitive: one picks a representative set of models in Θ, uses a uniform, "noninformative", prior on this set, lets the training data focus attention on the best model by computing a posterior distribution over this representative set of models, which will place much higher weight on those models that perform well on the training data, then finally uses an average of these well-performing models to form the predictive distribution for future outcomes. To get more and more accuracy as the number of training examples grows, one chooses larger and larger representative model sets, obtained by using a finer "mesh", i.e. a smaller separation ϵ between representative models. This leads to a kind of sieve method, as discussed in the introduction.

In some cases, in the limit as the sample size n goes to infinity, and hence the separation ϵ goes to zero, the uniform distribution over the maximal ϵ-separated set of models approaches something like a Jeffreys' prior over the model class Θ, which is known already to be asymptotically minimax (called "asymptotically least favorable" for technical reasons) for relative entropy risk in smooth parametric cases [14]. It is an interesting open problem to determine to what extent this holds for more general Θ, and what characterizes asymptotically minimax "generalized Jeffreys' priors".

As a practical classification method, the Bayes method using a uniform prior on an ϵ-separated set has two drawbacks

1. The size of the ϵ-separated set may grow too large too quickly as $\epsilon \rightarrow 0$. This happens for higher finite dimensional Θ, and for infinite dimensional Θ. This is kind of a "curse of dimensionality."
2. To compute the ϵ-separated set, it is required that one know the common marginal distribution P_ϕ on the the instance space X that it shared by the models in Θ. Often this distribution is not known, and one wants to do classification using a class of conditional distributions on the outcome Y given X, leaving the marginal distribution on X unspecified.

The first problem is a deep one. In cases where the asymptotically minimax prior is known and the posterior for this prior can be efficiently computed then this prior can be used in place of the priors on individual ϵ-separated sets. In applications of Bayes methods where such computations are not tractable it is common to employ Markov Chain Monte Carlo methods. However, it is difficult to give precise theoretical bounds on the performance of such methods.

The second problem can be handled by either trying to estimate the marginal distribution on the instance space, or by developing a method that works even for the worst case marginal distribution. In section 7 we have outlined what can be achieved with the latter approach in the special case of noisy two-class classification problems, relating our theory to the Vapnik-Chervonenkis theory.

It remains an important open problem to extend this analysis to arbitrary comparison model classes with a common marginal distribution. A related problem is to extend the whole of the theory used in this paper to handle the case where the models do not share a common marginal distribution.

9 Acknowledgements

David Haussler would like to thank Vincent Mirelli for suggesting some of the problems investigated here, and for valuable comments on an earlier draft of this paper.

References

1. D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.
2. P. Assouad. Densité et dimension. *Annales de l'Institut Fourier*, 33(3):233–282, 1983.
3. A. Barron. In T. M. Cover and B. Gopinath, editors, *Open Problems in Communication and Computation*, chapter 3.20. Are Bayes rules consistent in information?, pages 85–91. 1987.
4. A. Barron. The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, Dept. of Statistics, U. Ill. Urbana-Champaign, 1987.
5. A. Barron, B. Clarke, and D. Haussler. Information bounds for the risk of Bayesian predictions and the redundancy of universal codes. In *Proc. International Symposium on Information Theory.*
6. A. Barron and Y. Yang. Information theoretic lower bounds on convergence rates of nonparametric estimators, 1995. unpublished manuscript.
7. L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift fuer Wahrscheinlichkeitstheorie und verwandte gebiete*, 65:181–237, 1983.
8. L. Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probability theory and related fields*, 71:271–291, 1986.
9. L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.
10. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, 1987.
11. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.
12. B. Clarke. *Asymptotic cumulative risk and Bayes risk under entropy loss with applications.* PhD thesis, Dept. of Statistics, University of Ill., 1989.
13. B. Clarke and A. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.
14. B. Clarke and A. Barron. Jefferys' prior is asymptotically least favorable under entropy risk. *J. Statistical Planning and Inference*, 41:37–60, 1994.
15. G. F. Clements. Entropy of several sets of real-valued functions. *Pacific J. Math.*, 13:1085–1095, 1963.

16. T. Cover and J. Thomas. *Elements of Information Theory.* Wiley, 1991.
17. L. Devroye and L. Györfi. *Nonparametric density estimation, the L_1 view.* Wiley, 1985.
18. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, 1973.
19. R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097:2–142, 1984.
20. S. Y. Efroimovich. Information contained in a sequence of observations. *Problems in Information Transmission*, 15:178–189, 1980.
21. A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989.
22. M. Feder, Y. Freund, and Y. Mansour. Optimal universal learning and prediction of probabilistic concepts. In *Proc. of IEEE Information Theory Conference*, page 233. IEEE, 1995.
23. A. Gelman. *Bayesian Data Analysis.* Chapman and Hall, NY, 1995.
24. E. Giné and J. Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12:929–989, 1984.
25. R. Hasminskii and I. Ibragimov. On density estimation in the view of Kolmogorov's ideas in approximation theory. *Annals of statistics*, 18:999–1010, 1990.
26. D. Haussler and A. Barron. How well do Bayes methods work for on-line prediction of $\{+1, -1\}$ values? In *Proceedings of the Third NEC Symposium on Computation and Cognition.* SIAM, 1992.
27. D. Haussler and M. Opper. General bounds on the mutual information between a parameter and n conditionally independent observations. In *Proceedings of the Seventh Annual ACM Workshop on Computational Learning Theory*, 1995.
28. D. Haussler and M. Opper. Mutual information, metric entropy, and risk in estimation of probability distributions. Technical Report UCSC-CRL-96-27, Univ. of Calif. Computer Research Lab, Santa Cruz, CA, 1996.
29. I. Ibragimov and R. Hasminskii. On the information in a sample about a parameter. In *Second Int. Symp. on Information Theory*, pages 295–309, 1972.
30. A. J. Izenman. Recent developments in nonparametric density estimation. *JASA*, 86(413):205–224, 1991.
31. A. N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Translations (Ser. 2)*, 17:277–364, 1961.
32. L. LeCam. *Asymptotic methods in statistical decision theory.* Springer, 1986.
33. G. Lorentz. *Approxiamtion of Functions.* Holt, Rinehart, Winston, 1966.
34. R. Meir and N. Merhav. On the stochastic complexity of learning realizable and unrealizable rules. Unpublished manuscript, 1994.
35. M. Opper and D. Haussler. Bounds for predictive errors in the statistical mechanics of in supervised learning. *Physical Review Letters*, 75(20):3772–3775, 1995.
36. D. Pollard. *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics.* Institute of Math. Stat. and Am. Stat. Assoc., 1990.
37. J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100, 1986.
38. J. Rissanen, T. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Trans. Info. Th.*, 38:315–323, 1992.
39. N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (Series A)*, 13:145–147, 1972.

40. L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–42, 1984.
41. S. van deGeer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Annals of Statistics*, 21:14–44, 1993.
42. A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes.* Springer, NY, 1996.
43. V. N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, 1982.
44. V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–80, 1971.
45. W. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates for sieve MLE's. *Annals of Statistics*, 23(2):339–362, 1995.
46. B. Yu. Lower bounds on expected redundancy for nonparametric classes. *IEEE Trans. Info. Th.*, 42(1), 1996.

This article was processed using the LaTeX macro package with LLNCS style