Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension

David Haussler [*] U.C. Santa Cruz
Michael Kearns [†] AT&T Bell Labs
Robert Schapire [‡] AT&T Bell Labs

Abstract

In this paper we study a Bayesian or average-case model of concept learning with a twofold goal: to provide more precise characterizations of learning curve (sample complexity) behavior that depend on properties of both the prior distribution over concepts and the sequence of instances seen by the learner, and to smoothly unite in a common framework the popular statistical physics and VC dimension theories of learning curves. To achieve this, we undertake a systematic investigation and comparison of two fundamental quantities in learning and information theory: the probability of an incorrect prediction for an optimal learning algorithm, and the Shannon information gain. This study leads to a new understanding of the sample complexity of learning in several existing models.

1 Introduction

Consider a simple concept learning model in which the learner attempts to infer an unknown *target concept* f, chosen from a known *concept class* $\mathcal{F}$ of $\{0,1\}$-valued functions over an instance space X. At each trial i, the learner is given a point $x_i \in X$ and asked to predict the value of $f(x_i)$. If the learner predicts $f(x_i)$ incorrectly, we say the learner makes a *mistake*. After making its prediction, the learner is told the correct value.

Informally speaking, there are at least two natural measures of the performance of a learning algorithm in this setting:

The probability the algorithm makes a mistake on $f(x_{m+1})$, having already seen the examples $(x_1, f(x_1)), \ldots, (x_m, f(x_m))$. Regarded as a function of m, this familiar measure is known as the algorithm's *learning curve*.

[*]Author's address: Computer and Information Sciences, University of California at Santa Cruz, Santa Cruz, California 95064.

[†]Author's address: AT&T Bell Laboratories, 600 Mountain Ave., Room 2A-423, P.O. Box 636, Murray Hill, New Jersey, 07974-0636. This research was conducted while the author was at the Massachusetts Institute of Technology Laboratory for Computer Science and at the International Computer Science Institute.

[‡]Author's address: AT&T Bell Laboratories, 600 Mountain Ave., Room 2A-424, P.O. Box 636, Murray Hill, New Jersey, 07974-0636. This research was conducted while the author was at the Massachusetts Institute of Technology Laboratory for Computer Science and at Harvard University.

The total number of mistakes made by the algorithm on the first m trials $f(x_1), \ldots, f(x_m)$. This measure counts the *cumulative mistakes* of the algorithm.

These measures are clearly closely related to each other. In either measure, we are interested in the asymptotic behavior of a learning algorithm as m becomes large. Since the learning curve can be used to determine how large m must be before the probability of mistake drops below a desired value ϵ, the study of learning curves may also be viewed as the study of the *sample complexity* of learning.

The recent and intensive investigation of concept learning undertaken by the research communities of neural networks, artificial intelligence, cognitive science and computational learning theory has resulted in the development of at least two fairly general and successful viewpoints of the learning process in terms of learning curves and cumulative mistakes. One of these, arising from the study of Valiant's *distribution-free* or *probably approximately correct* model [32] and having roots in the pattern recognition and minimax decision theory literature, characterizes the *distribution-free, worst-case* sample complexity of concept learning in terms of a combinatorial parameter known as the *Vapnik-Chervonenkis (VC) dimension* [34, 4]. In contrast, the *average-case* sample complexity of learning in neural networks has recently been investigated from a standpoint that is essentially Bayesian[1], and is strongly influenced by ideas and tools from statistical physics, as well as by information theory [10, 31, 15, 29, 24]. While each of these theories has its own distinct strengths and drawbacks, there is little understanding of what relationships hold between them.

In this paper, we study an average-case or Bayesian model of learning with two primary goals. First, we are interested in ultimately developing a general framework that provides precise characterizations of learning curves and expected cumulative mistakes that extends and refines the VC dimension and statistical physics theories. The results presented here are a first step in this direction. Second, we would like this framework to smoothly incorporate both of these previous theories, thus yielding a unified viewpoint that can be used both for giving realistic estimates of average-case performance in the case that the distributions on the concept class and instance space are known, and for giving good worst-case estimates in the case that these distributions are not known.

In a setting where the target concept is drawn at random according to a fixed but arbitrary prior distribution $\mathcal{P}$, we undertake a systematic investigation and comparison of two fundamental quantities in learning and information theory: the probability of mistake (known as the *0-1 loss* in decision theory) for an optimal learning algorithm, and the Shannon information gain from the labels of the instance sequence. In doing so, we borrow from and contribute to the work on weighted majority and aggregating learning strategies [18, 20, 36, 11, 2, 19], as well as to the VC dimension and statistical physics work. This study leads to a new understanding of the sample complexity of learning in several existing models.

One of our main motivations for this research arises from the frequent claims of machine learning practitioners that sample complexity bounds derived via the VC dimension are overly pessimistic in practice [5, 26]. This pessimism can be traced to three assumptions that are implicit in results that are based on the VC dimension. The first pessimistic assumption is that only the worst-case performance over possible target concepts counts. This is the minimax pessimism. We may think of an adversary choosing the hardest possible concept

[1]More general Bayesian approaches to learning in neural networks are described in the recent papers [21, 6].

for the learner, rather than the Bayesian approach which incorporates prior beliefs regarding which concepts might be "more likely".

The second pessimistic assumption is that even though VC dimension analysis allows a distribution $\mathcal{D}$ over the instance space X, this distribution is also assumed to be the hardest possible for learning the class $\mathcal{F}$. Thus, the VC dimension is also based on a worst-case assumption over instance space distributions. In addition to the VC dimension, Vapnik and Chervonenkis have a distribution-specific formulation that overcomes this limitation [35], but apart from Natarajan's work [23], it has not been used much in computational learning theory. We extend this idea further in Section 9.

The third and perhaps most subtle pessimistic assumption can be seen by noting that the VC dimension provides upper bounds on the learning curves of *any* consistent learning algorithm. Thus, even the hypothetical algorithm that always manages to find a hypothesis that is consistent with the examples so far but that has the largest possible error with respect to $\mathcal{D}$ is covered by VC dimension analysis. (This is the *uniform convergence* property of the VC dimension). In practice it seems unlikely that one would encounter such algorithms — reasonable algorithms should manage to find an "average" consistent hypothesis (in terms of error on $\mathcal{D}$) rather than the "worst" consistent hypothesis.

In this paper we attempt to address each of these pessimistic assumptions in the hopes of obtaining a more realistic picture of sample complexity. To relax the worst-case assumption over the concept class $\mathcal{F}$, we adopt a Bayesian framework that places a prior distribution $\mathcal{P}$ over $\mathcal{F}$. If we also assume that the target concept is drawn according to $\mathcal{P}$, then this allows us to derive bounds on learning curves and cumulative mistakes that depend on properties of the particular prior $\mathcal{P}$.

Our solution to the worst-case assumption over the instance space distribution $\mathcal{D}$ is twofold. For most of the paper, we in fact do not need to assume that there is a distribution governing the generation of sample points, and instead *fix* an arbitrary sequence of instances $\mathbf{x} = x_1, \ldots, x_m, x_{m+1}, \ldots$ that is seen by the learner. We do not assume that this sequence is worst-case (distinguishing this setting from the various adversary-based on-line learning models that count worst-case mistake bounds), or that it is drawn randomly (distinguishing this setting from the VC dimension and statistical physics theories). Thus our bounds on learning curves and cumulative mistakes also depend on properties of $\mathbf{x}$. Two advantages that come from allowing $\mathbf{x}$ to be a parameter are that we incorporate time-dependent instance sequences, and we model the fact that a learning algorithm does in fact have the training data in its possession, and may be able to exploit this knowledge. For some of our later results, particularly for comparing our bounds with those derived via the VC dimension, we will need to revert to the assumption that the instances in $\mathbf{x}$ are generated independently at random according to an instance space distribution $\mathcal{D}$ (but here again, our bounds will depend on properties of the particular $\mathcal{D}$ in contrast to worst-case bounds).

Finally, to address the pessimism implicit in demanding uniform convergence, we will study *particular* learning algorithms of interest rather than giving bounds for any consistent algorithm. In addition to analyzing the learning curve and cumulative mistakes of the optimal prediction algorithm (the *Bayes algorithm*), we simultaneously study the algorithm that outputs a *random* consistent hypothesis (the *Gibbs algorithm*). The motivation for this latter algorithm is exactly that of relaxing the uniform convergence demand while still making realistic assumptions about practical learning algorithms, since this algorithm will output a consistent hypothesis whose error with respect to the instance space distribution

$\mathcal{D}$ is the average (over $\mathcal{P}$), not the worst.

One appealing aspect of our approach is the elementary nature of most of the proofs, which rest almost entirely on well-known or easily derived algebraic expressions for the information gain and the probability of mistake, and employ simple inequalities relating these expressions. The additivity of the Shannon information is invoked repeatedly in order to obtain easy and useful bounds on otherwise complicated sums. For instance, our results include a short and transparent derivation of an upper bound on the expected total number of mistakes in terms of the VC dimension that is tight to within a constant factor.

Perhaps the main strength of this research is the unifying framework it provides for several previously unrelated theories and results. By beginning in a model that averages over both the concept class and the instance space, then gradually removing the averaging in favor of combinatorial parameters that upper bound certain expectations, we can move smoothly from the information theoretic bounds of the Bayesian and statistical physics theory to bounds based on the VC dimension. Thus, our bounds can be used both for average-case analyses of particular distributions, or for worst-case bounds in situations where the prior or instance space distribution is arbitrary.

The aim of this paper is to demonstrate the applicability of information theory tools in an average-case learning model, and to show how some important results in the VC dimension theory can be reconstructed from these simple mechanisms. Towards ease of exposition and technical simplicity and clarity, we have chosen the simplest concept learning model that is still of general interest; clearly this model is far from being a perfect model of the real world. In a later companion paper, we hope to develop our methods further and apply them to more varied and realistic models; some of this ongoing work is outlined in Section 12. Many beautiful results on the performance of Bayesian methods are also given in the statistics literature, see e.g. [8, 7] and references therein.

2 Summary of results

Following a brief introduction of some notation in Section 3, our results begin in Section 4. Here we define the Shannon information gain of an example, and introduce the two learning algorithms we shall study. The primary purpose of this section is to derive expressions for the information gain and the probabilities of mistake for the two learning algorithms in terms of an important random variable known as the *volume ratio*.

In Section 5 we prove that the probabilities of mistake for our two learning algorithms can be bounded above and below by simple functions of the expected information gain. As in the paper of Tishby, Levin and Solla [31], we upper bound the probability of mistake by the information gain. We also provide an information-theoretic lower bound on the probability of mistake, which can be viewed as a special case of Fano's inequality [9, 14]. Together these bounds provide a general characterization of learning curve behavior that is accurate to within a logarithmic factor.

In Section 6 we exploit the learning curve bounds of Section 5 and the additivity of information to obtain upper and lower bounds on the cumulative mistakes of our algorithms that are simple functions of the total information gain. These bounds are again tight to within a logarithmic factor. The total information gain is naturally expressed here as an appropriate entropy expression. This entropy forms the crucial link between the Bayesian

approach and the VC dimension bounds. This link is investigated in detail in Section 9.

In Section 7 we investigate the important variation of the basic Bayesian model in which the target concept f is drawn according to a *true prior* $\mathcal{Q}$ that may differ from the learner's *perceived prior* $\mathcal{P}$. We again bound learning curves by information gain and cumulative mistakes by an entropy depending only on $\mathcal{Q}$ plus an additive "penalty term" measuring the distance between $\mathcal{P}$ and $\mathcal{Q}$.

In Section 8 we prove that if the instances are chosen randomly according to an instance space distribution $\mathcal{D}$ then the instantaneous information gain is a non-increasing function of m. This result is used in Section 9, where we demonstrate that some important results in the VC dimension theory of learning curves and cumulative mistakes can in fact be recovered from the simple information-theoretic results in the Bayesian model. This is primarily accomplished by gradually removing averaging over the instance space and the target class in favor of combinatorial parameters that upper bound certain expectations. The main technical tool required is the Sauer/VC combinatorial lemma. In Section 10 we extend these ideas to show how the VC dimension can be used to obtain improved bounds in the case that the perceived prior and true prior differ.

In Section 12 we draw some conclusions and mention extensions of the results presented here.

3 Notational conventions

Before presenting our results, we establish a few notational conventions. Let X be the *instance space*. A *concept class* $\mathcal{F}$ over X is a (possibly infinite) collection of subsets of X. We will find it convenient to view a concept $f \in \mathcal{F}$ as a function $f : X \to \{0, 1\}$, where we interpret $f(x) = 1$ to mean that $x \in X$ is a *positive example* of f, and $f(x) = 0$ to mean x is a *negative example* of f.

The symbols $\mathcal{P}$, $\mathcal{Q}$ and $\mathcal{D}$ are used to denote probability distributions. The distributions $\mathcal{P}$ and $\mathcal{Q}$ are over $\mathcal{F}$, and $\mathcal{D}$ is over X. When $\mathcal{F}$ and X are countable we assume that these distributions are defined as probability mass functions. For uncountable $\mathcal{F}$ and X they are assumed to be probability measures over some appropriate σ-algebra. All of our results hold for both countable and uncountable $\mathcal{F}$ and X.

We use the notation $\mathbf{E}_{f \in \mathcal{P}}[\chi(f)]$ for the expectation of the random variable χ under the distribution $\mathcal{P}$, and $\mathbf{Pr}_{f \in \mathcal{P}}[cond(f)]$ for the probability under the distribution $\mathcal{P}$ of the set of all f satisfying the predicate $cond(f)$. Everything that needs to be measurable is assumed to be measurable.

4 Instantaneous information gain and mistake probabilities

In this section we begin the analysis of the three quantities that form the backbone of the theory developed here: the Shannon information gain from a labeled example, and the probability of mistake for the *Bayes* and *Gibbs* learning algorithms. Our immediate goal is to define these algorithms and quantities, and to derive expressions for the behavior of each in terms of an important random variable that we shall call the *volume ratio*.

Let $\mathcal{F}$ be a concept class over the instance space X. Fix a *target concept* $f \in \mathcal{F}$ and an infinite sequence of instances $\mathbf{x} = x_1, \ldots, x_m, x_{m+1}, \ldots$ with $x_m \in X$ for all m. For now we assume that the fixed instance sequence $\mathbf{x}$ is known in advance to the learner, but that the target concept f is not. Let $\mathcal{P}$ be a probability distribution over the concept class $\mathcal{F}$. We think of $\mathcal{P}$ in the Bayesian sense as representing the *prior beliefs* of the learner about which target concept it will be learning.

In our setting, the learner receives information about f incrementally via the label sequence $f(x_1), \ldots, f(x_m), f(x_{m+1}), \ldots$. At time m, the learner receives the label $f(x_m)$. For any $m \geq 1$ we define (with respect to $\mathbf{x}, f$) the mth *version space*

$$\mathcal{F}_m(\mathbf{x}, f) = \{\hat{f} \in \mathcal{F} : \hat{f}(x_1) = f(x_1), \ldots, \hat{f}(x_m) = f(x_m)\}$$

and the mth *volume* $V_m^{\mathcal{P}}(\mathbf{x}, f) = \mathcal{P}[\mathcal{F}_m(\mathbf{x}, f)]$. We define $\mathcal{F}_0(\mathbf{x}, f) = \mathcal{F}$ for all $\mathbf{x}$ and f, so $V_0^{\mathcal{P}}(\mathbf{x}, f) = 1$. The version space at time m is simply the class of all concepts in $\mathcal{F}$ consistent with the first m labels of f (with respect to $\mathbf{x}$), and the mth volume is the measure of this class under $\mathcal{P}$. For the first part of the paper, the infinite instance sequence $\mathbf{x}$ and the prior $\mathcal{P}$ are fixed, thus we simply write $\mathcal{F}_m(f)$ and $V_m(f)$. Later, when we need to discuss distributions other than $\mathcal{P}$, or when the sequence $\mathbf{x}$ is chosen randomly, we will reintroduce these dependencies explicitly.

For each $m \geq 0$ let us define the mth *posterior distribution* $\mathcal{P}_m$ by restricting $\mathcal{P}$ to the mth version space $\mathcal{F}_m(f)$, that is, for all (measurable) $S \subset \mathcal{F}$, $\mathcal{P}_m[S] = \mathcal{P}[S \cap \mathcal{F}_m(f)]/\mathcal{P}[\mathcal{F}_m(f)] = \mathcal{P}[S \cap \mathcal{F}_m(f)]/V_m(f)$. Note that $\mathcal{P}_m$ has an implicit dependence on $\mathbf{x}$ and f that we have omitted for notational brevity. The posterior probability distribution $\mathcal{P}_m$ can be interpreted as the subjective probability distribution over various possible target concepts, given the labels $f(x_1), \ldots, f(x_m)$ of the first m instances.

Digressing momentarily from the problem of learning f, in this setting we may now ask the following question: Having already seen $f(x_1), \ldots, f(x_m)$, how much information (assuming the prior $\mathcal{P}$) does the learner gain by seeing $f(x_{m+1})$? (We think of this as the *instantaneous information gain*, since we address the gain only on the $m+1$st label.) The classic answer provided by information theory is that the information carried by $f(x_{m+1})$ is given by the quantity

$$\begin{aligned}
\mathcal{I}_{m+1}^{\mathcal{P}}(\mathbf{x}, f) &= \mathcal{I}_{m+1}(f) \\
&= -\log \mathbf{Pr}_{\hat{f} \in \mathcal{P}_m}[\hat{f}(x_{m+1}) = f(x_{m+1}) | \hat{f}(x_i) = f(x_i), 1 \leq i \leq m] \\
&= -\log \frac{V_{m+1}(f)}{V_m(f)} \\
&= -\log \chi_{m+1}(f)
\end{aligned}$$

where we define the $m+1$st *volume ratio* by

$$\chi_{m+1}^{\mathcal{P}}(\mathbf{x}, f) = \chi_{m+1}(f) = V_{m+1}(f)/V_m(f)$$

We shall be primarily interested in the *expected* information gain when f is chosen randomly according to $\mathcal{P}$, which may now be expressed

$$\mathbf{E}_{f \in \mathcal{P}}[\mathcal{I}_{m+1}(f)] = \mathbf{E}_{f \in \mathcal{P}}[-\log \chi_{m+1}(f)] \tag{1}$$

We now return to our learning problem, which we define to be that of predicting the label $f(x_{m+1})$ given only the previous labels $f(x_1), \ldots, f(x_m)$. The first learning algorithm we consider is called the *Bayes optimal classification algorithm* [12], or the *Bayes* algorithm for short. It is a special case of the weighted majority algorithm [20]. For any m and $b \in \{0,1\}$, define $\mathcal{F}_m^b(\mathbf{x}, f) = \mathcal{F}_m^b(f) = \{\hat{f} \in \mathcal{F}_m(\mathbf{x}, f) : \hat{f}(x_{m+1}) = b\}$. Then the Bayes algorithm behaves as follows:

If $\mathcal{P}_m[\mathcal{F}_m^1(f)] > \mathcal{P}_m[\mathcal{F}_m^0(f)]$, it predicts that $f(x_{m+1}) = 1$.

If $\mathcal{P}_m[\mathcal{F}_m^1(f)] < \mathcal{P}_m[\mathcal{F}_m^0(f)]$, it predicts that $f(x_{m+1}) = 0$.

If $\mathcal{P}_m[\mathcal{F}_m^1(f)] = \mathcal{P}_m[\mathcal{F}_m^0(f)]$, it flips a fair coin and uses the outcome to predict $f(x_{m+1})$.

When the target concept f is drawn at random according to the prior distribution $\mathcal{P}$, then the Bayes algorithm is optimal in the sense that it minimizes the probability that $f(x_{m+1})$ is predicted incorrectly.

Despite the optimality of the Bayes algorithm, it suffers the philosophical (and potentially practical) drawback that its *hypothesis* at any time m may not be a member of the target class $\mathcal{F}$. (Here we define the hypothesis of an algorithm at time m to be the (possibly probabilistic) mapping $f : X \to \{0,1\}$ obtained by letting $f(x)$ be the prediction of the algorithm when $x_{m+1} = x$.) This drawback is absent in our second learning algorithm, which is called the *Gibbs* algorithm [24], and behaves as follows:

Given the labels $f(x_1), \ldots, f(x_m)$, a hypothesis concept $\hat{f}$ is chosen randomly according to the posterior distribution $\mathcal{P}_m$.

Given x_{m+1}, the algorithm then predicts that $f(x_{m+1}) = \hat{f}(x_{m+1})$.

Thus, the Gibbs algorithm simply chooses a hypothesis randomly (according to $\mathcal{P}$) from $\mathcal{F}$ among those that are consistent with the labels seen so far. The Gibbs algorithm is the "zero-temperature" limit of the learning algorithm studied in several recent papers [10, 31, 15, 29].

It is important to note that both the Bayes and Gibbs algorithms are quite different from the well-known *maximum a posteriori* algorithm, which chooses the hypothesis $\hat{f}$ that maximizes the posterior probability $\mathcal{P}_m[\hat{f}]$. While this algorithm maximizes the probability of *exactly identifying* the target concept, it may do quite poorly in the instantaneous mistake (learning curve) measure. In contrast, the Bayes algorithm has the optimal learning curve, and we shall see shortly that the Gibbs algorithm has a nearly optimal learning curve.

We now wish to derive expressions for the probability that $f(x_{m+1})$ is predicted incorrectly by these two algorithms. These are the *instantaneous mistake probabilities*, since they only address the probability of a mistake on the $m+1$st label. As was the case for the expected information conveyed by $f(x_{m+1})$ with respect to $\mathcal{P}$ given by Equation (1), we would like these probabilities to be expressed in terms of the volume ratio $\chi_{m+1}(f)$.

For the Bayes algorithm, note that a mistake in predicting $f(x_{m+1})$ is made with probability 1 if $V_{m+1}(f) < \frac{1}{2}V_m(f)$, with probability $\frac{1}{2}$ if $V_{m+1}(f) = \frac{1}{2}V_m(f)$, and with probability 0 otherwise. Thus we may express the Bayes mistake probability on $f(x_{m+1})$ as

$$Bayes_{m+1}^{\mathcal{P}}(\mathbf{x}, f) = Bayes_{m+1}(f) = \Theta\left(\frac{1}{2} - \chi_{m+1}(f)\right)$$

where $\Theta(x) = 1$ if $x > 0$, $\Theta(0) = \frac{1}{2}$, and $\Theta(x) = 0$ otherwise. The probability of mistake when f is chosen randomly according to $\mathcal{P}$ is thus simply

$$\mathbf{E}_{f\in\mathcal{P}}[Bayes^{\mathcal{P}}_{m+1}(\mathbf{x}, f)] = \mathbf{E}_{f\in\mathcal{P}}\left[\Theta\left(\frac{1}{2} - \chi_{m+1}(f)\right)\right] \tag{2}$$

For the Gibbs algorithm, note that the prediction of $f(x_{m+1})$ is *correct* if and only if the randomly chosen hypothesis $\hat{f}$ is in $\mathcal{F}_{m+1}(f)$. Since $\hat{f}$ is chosen randomly according to the posterior $\mathcal{P}_m$, and the probability of $\mathcal{F}_{m+1}(f)$ under $\mathcal{P}_m$ is exactly $V_{m+1}(f)/V_m(f) = \chi_{m+1}(f)$, we may write the probability that $f(x_{m+1})$ is predicted incorrectly for fixed $\mathbf{x}, f$ and $\mathcal{P}$ as

$$Gibbs^{\mathcal{P}}_{m+1}(\mathbf{x}, f) = Gibbs_{m+1}(f) = 1 - \chi_{m+1}(f)$$

In the case that f is drawn according to $\mathcal{P}$ we have

$$\mathbf{E}_{f\in\mathcal{P}}[Gibbs^{\mathcal{P}}_{m+1}(\mathbf{x}, f)] \;=\; \mathbf{E}_{f\in\mathcal{P}}[1 - \chi_{m+1}(f)] \tag{3}$$

Note that by the definition of the Gibbs algorithm, Equation (3) is exactly the probability of mistake of a random consistent hypothesis in $\mathcal{F}$, using the distribution on $\mathcal{F}$ defined by the prior. Thus if we also average over $\mathbf{x}$, bounds on this expectation provide an interesting contrast to those obtained via VC dimension analysis, which always gives bounds on the probability of mistake of the *worst* consistent hypothesis.

5 Bounding the mistake probabilities by the information gain

Now that we have obtained expressions for the information gain and mistake probabilities in terms of the volume ratio, in this section we use these expressions to show that the mistake probabilities can always be bounded above and below by simple functions of the information gain.

First we extend our notation $\mathcal{F}_m(f)$ and $V_m(f)$ to allow a sequence of bits, $\mathbf{y} = \langle y_1, \ldots, y_n\rangle$ $(n \geq m)$, representing labels of $x_1, \ldots, x_n$, to replace the argument f. Thus, we define $\mathcal{F}_m(\mathbf{x}, \mathbf{y}) = \mathcal{F}_m(\mathbf{y}) = \{\hat{f} \in \mathcal{F} : \hat{f}(x_1) = y_1, \ldots, \hat{f}(x_m) = y_m\}$. Note that in the case that $n > m$, the last $n - m$ bits of the sequence are ignored, in the same way that in the notation $\mathcal{F}_m(f)$ only the first m values of f on $\mathbf{x}$ are relevant. Similarly, we define $V^{\mathcal{P}}_m(\mathbf{x}, \mathbf{y}) = V_m(\mathbf{y}) = \mathcal{P}[\mathcal{F}_m(\mathbf{y})]$ and thus $\chi^{\mathcal{P}}_{m+1}(\mathbf{x}, \mathbf{y}) = \chi_{m+1}(\mathbf{y}) = V_{m+1}(\mathbf{y})/V_m(\mathbf{y})$.

Let $\mathcal{G}$ be an *arbitrary* real-valued function of one argument, and let us examine the expectation $\mathbf{E}_{f\in\mathcal{P}}[\mathcal{G}(\chi_{m+1}(f))]$. Note that by Equations (1), (2) and (3), we may write the expectations (over the random choice of f according to $\mathcal{P}$) of $\mathcal{I}_{m+1}(f)$, $Bayes_{m+1}(f)$ and $Gibbs_{m+1}(f)$ in this form. Since

$$V_{m+1}(\mathbf{y}) \;=\; \mathbf{Pr}_{f\in\mathcal{P}}[f(x_1) = y_1 \wedge \cdots \wedge f(x_{m+1}) = y_{m+1}]$$

we have

$$\mathbf{E}_{f\in\mathcal{P}}[\mathcal{G}(\chi_{m+1}(f))] = \sum_{\mathbf{y}\in\{0,1\}^{m+1}} V_{m+1}(\mathbf{y})\mathcal{G}(\chi_{m+1}(\mathbf{y}))$$

$$
\begin{aligned}
&= \sum_{\mathbf{y}\in\{0,1\}^m} [V_{m+1}(\langle \mathbf{y}, 0\rangle)\mathcal{G}(\chi_{m+1}(\langle \mathbf{y}, 0\rangle)) + V_{m+1}(\langle \mathbf{y}, 1\rangle)\mathcal{G}(\chi_{m+1}(\langle \mathbf{y}, 1\rangle))] \\
&= \sum_{\mathbf{y}\in\{0,1\}^m} V_m(\mathbf{y})\,[\chi_{m+1}(\langle \mathbf{y}, 0\rangle)\mathcal{G}(\chi_{m+1}(\langle \mathbf{y}, 0\rangle)) + \chi_{m+1}(\langle \mathbf{y}, 1\rangle)\mathcal{G}(\chi_{m+1}(\langle \mathbf{y}, 1\rangle))] \\
&= \sum_{\mathbf{y}\in\{0,1\}^{m+1}} V_{m+1}(\mathbf{y})\,[\chi_{m+1}(\mathbf{y})\mathcal{G}(\chi_{m+1}(\mathbf{y})) + \chi_{m+1}(\mathbf{y}')\mathcal{G}(\chi_{m+1}(\mathbf{y}'))]
\end{aligned}
$$

where $\mathbf{y}'$ is the vector of labels obtained from $\mathbf{y}$ by flipping the last label. Since $\chi_{m+1}(\mathbf{y}') = 1 - \chi_{m+1}(\mathbf{y})$, it follows that

$$
\begin{aligned}
&\mathbf{E}_{f\in\mathcal{P}}[\mathcal{G}(\chi_{m+1}(f))] \\
&\quad = \sum_{\mathbf{y}\in\{0,1\}^{m+1}} V_{m+1}(\mathbf{y})\,[\chi_{m+1}(\mathbf{y})\mathcal{G}(\chi_{m+1}(\mathbf{y})) + (1-\chi_{m+1}(\mathbf{y}))\mathcal{G}(1-\chi_{m+1}(\mathbf{y}))] \\
&\quad = \mathbf{E}_{f\in\mathcal{P}}[\chi_{m+1}(f)\mathcal{G}(\chi_{m+1}(f)) + (1-\chi_{m+1}(f))\mathcal{G}(1-\chi_{m+1}(f))] \qquad (4)
\end{aligned}
$$

The form of the expression inside the expectation of Equation (4) is $p\mathcal{G}(p)+(1-p)\mathcal{G}(1-p)$ (using the substitution $p = \chi_{m+1}(f)$), and is suggestive of a binary "entropy", in which we interpret $p \in [0,1]$ as a probability, and $\mathcal{G}(p)$ to be the "information" conveyed by the occurrence of an event whose probability is p.

We now apply Equation (4) to the three forms of $\mathcal{G}$ we have been considering, namely $\mathcal{G}(p) = -\log p$ (from Equation (1)), $\mathcal{G}(p) = \Theta(\frac{1}{2} - p)$ (from Equation (2)), and $\mathcal{G}(p) = 1 - p$ (from Equation (3)). From these three equations and some simple algebra we obtain

$$
\mathbf{E}_{f\in\mathcal{P}}[\mathcal{I}_{m+1}(f)] = \mathbf{E}_{f\in\mathcal{P}}[-\log \chi_{m+1}(f)] = \mathbf{E}_{f\in\mathcal{P}}[\mathcal{H}(\chi_{m+1}(f))] \qquad (5)
$$

for the expected information gain from $f(x_{m+1})$, where $\mathcal{H}$ is the familiar binary entropy function

$$
\mathcal{H}(p) = -p\log p - (1-p)\log(1-p)
$$

Note that since $0 \le \chi_{m+1}(f) \le 1$, this implies that on average, at most 1 bit of Shannon information can be obtained from a labeled example.

For the probability of mistake of the Bayes algorithm, we obtain

$$
\begin{aligned}
\mathbf{E}_{f\in\mathcal{P}}[Bayes_{m+1}(f)] &= \mathbf{E}_{f\in\mathcal{P}}\left[\Theta\left(\frac{1}{2} - \chi_{m+1}(f)\right)\right] \\
&= \mathbf{E}_{f\in\mathcal{P}}[\min(\chi_{m+1}(f), 1-\chi_{m+1}(f))] \qquad (6)
\end{aligned}
$$

For the probability of mistake of the Gibbs algorithm, we have

$$
\begin{aligned}
\mathbf{E}_{f\in\mathcal{P}}[Gibbs_{m+1}(f)] &= \mathbf{E}_{f\in\mathcal{P}}[1-\chi_{m+1}(f)] \\
&= \mathbf{E}_{f\in\mathcal{P}}[2\chi_{m+1}(f)(1-\chi_{m+1}(f))] \qquad (7)
\end{aligned}
$$

Now it is easily verified that for any $p \in [0,1]$,

$$
\min(p, 1-p) \le 2p(1-p) \le \frac{1}{2}\mathcal{H}(p) \qquad (8)
$$

Let us now define an inverse to $\mathcal{H}$ by letting $\mathcal{H}^{-1}(q)$, for $q \in [0,1]$, be the unique $p \in [0, 1/2]$ such that $\mathcal{H}(p) = q$. Note that

$$
\mathcal{H}^{-1}(\mathcal{H}(p)) = \min(p, 1-p)
$$

Then from Equations (5)–(8), and Jensen's inequality, we may conclude

$$\begin{aligned}
\mathcal{H}^{-1}(\mathbf{E}_{f\in\mathcal{P}}[\mathcal{I}_{m+1}(f)]) &= \mathcal{H}^{-1}(\mathbf{E}_{f\in\mathcal{P}}[\mathcal{H}(\chi_{m+1}(f))]) \\
&\leq \mathbf{E}_{f\in\mathcal{P}}[\mathcal{H}^{-1}(\mathcal{H}(\chi_{m+1}(f)))] \\
&= \mathbf{E}_{f\in\mathcal{P}}[\min(\chi_{m+1}(f), 1-\chi_{m+1}(f))] \\
&= \mathbf{E}_{f\in\mathcal{P}}[Bayes_{m+1}(f)] \\
&\leq \mathbf{E}_{f\in\mathcal{P}}[Gibbs_{m+1}(f)] \\
&\leq \frac{1}{2}\mathbf{E}_{f\in\mathcal{P}}[\mathcal{H}(\chi_{m+1}(f))] \\
&= \frac{1}{2}\mathbf{E}_{f\in\mathcal{P}}[\mathcal{I}_{m+1}(f)] \qquad (9)
\end{aligned}$$

Thus we see that the probabilities of mistake for both the Bayes and the Gibbs algorithms are between $\mathcal{H}^{-1}(\mathbf{E}_{f\in\mathcal{P}}[\mathcal{I}_{m+1}(f)])$ and $\frac{1}{2}\mathbf{E}_{f\in\mathcal{P}}[\mathcal{I}_{m+1}(f)]$. These upper and lower bounds are equal (and therefore tight) at both extremes $\mathbf{E}_{f\in\mathcal{P}}[\mathcal{I}_{m+1}(f)] = 1$ (maximal information gain) and $\mathbf{E}_{f\in\mathcal{P}}[\mathcal{I}_{m+1}(f)] = 0$ (minimal information gain). As the information gain becomes smaller, the difference between the upper and lower bounds shrinks, but the ratio of the two bounds grows logarithmically in the inverse of the information gain. In particular, it can be shown that there is a constant $c_0 > 0$ such that for all $p > 0$, $\mathcal{H}^{-1}(p) \geq c_0 p/\log(2/p)$, so we may also write the chain of inequalities ending with Equation (9) as

$$\begin{aligned}
\frac{c_0\mathbf{E}_{f\in\mathcal{P}}[\mathcal{I}_{m+1}(f)]}{\log(2/\mathbf{E}_{f\in\mathcal{P}}[\mathcal{I}_{m+1}(f)])} &\leq \mathbf{E}_{f\in\mathcal{P}}[Bayes_{m+1}(f)] \\
&\leq \mathbf{E}_{f\in\mathcal{P}}[Gibbs_{m+1}(f)] \\
&\leq \frac{1}{2}\mathbf{E}_{f\in\mathcal{P}}[\mathcal{I}_{m+1}(f)] \qquad (10)
\end{aligned}$$

Note that the upper and lower bounds given in both versions depend on properties of the particular prior $\mathcal{P}$, and on properties of the particular fixed sequence $\mathbf{x}$.

Finally, if all that is wanted is a direct comparison of the performances of the Gibbs and Bayes algorithms, a tighter relationship can be obtained from Equations (6), (7), (8), and the simple observation $p(1-p) \leq \min(p, 1-p)$, giving

$$\mathbf{E}_{f\in\mathcal{P}}[Bayes_{m+1}(f)] \leq \mathbf{E}_{f\in\mathcal{P}}[Gibbs_{m+1}(f)] \leq 2\mathbf{E}_{f\in\mathcal{P}}[Bayes_{m+1}(f)] \qquad (11)$$

6 Bounding the cumulative mistakes by the partition entropy

So far we have been primarily interested in analyzing the expectations of $Bayes_{m+1}(f)$ and $Gibbs_{m+1}(f)$. These expectations may be thought of as the *instantaneous* mistake probabilities: they are the probabilities a mistake is made in predicting $f(x_{m+1})$, and as such do not explicitly address what happened on the predictions of $f(x_1), \ldots, f(x_m)$. Similarly, $\mathcal{I}_{m+1}(f)$ is the *instantaneous information*, the information conveyed by $f(x_{m+1})$, without explicit regard for the information conveyed by the previous labels. A natural alternative measure is a *cumulative* bound — namely, the expected total information gained from the first m labels, or the expected number of mistakes made in the first m trials. While direct

analysis of the expressions for the expected number of mistakes for the Bayes and Gibbs algorithms is difficult due to the lack of a simple closed-form expression, the situation for the cumulative information gain is quite different due to the additivity of information. More precisely, we may write

$$\begin{aligned}
\mathbf{E}_{f\in\mathcal{P}}[\sum_{i=1}^{m}\mathcal{I}_i(f)] &= \mathbf{E}_{f\in\mathcal{P}}[\sum_{i=1}^{m}-\log\chi_i(f)] \\
&= \mathbf{E}_{f\in\mathcal{P}}[\sum_{i=1}^{m}(\log V_{i-1}(f)-\log V_i(f))] \\
&= \mathbf{E}_{f\in\mathcal{P}}[-\log V_m(f)] \qquad (12)
\end{aligned}$$

since $V_0(f)=1$, and recalling the definition of the volume ratio $\chi_i(f)$.

The final expression obtained in Equation (12) has a natural interpretation. The first m instances $x_1,\ldots,x_m$ of $\mathbf{x}$ induce a partition $\Pi_m^{\mathcal{F}}(\mathbf{x})$ of the concept class $\mathcal{F}$ defined by $\Pi_m^{\mathcal{F}}(\mathbf{x})=\Pi_m^{\mathcal{F}}=\{\mathcal{F}_m(\mathbf{x},f):f\in\mathcal{F}\}$. Note that $|\Pi_m^{\mathcal{F}}|$ is always at most 2^m, but may be considerably smaller, depending on the interaction between $\mathcal{F}$ and $x_1,\ldots,x_m$. It is clear that

$$\mathbf{E}_{f\in\mathcal{P}}[-\log V_m(f)] = -\sum_{\pi\in\Pi_m^{\mathcal{F}}}\mathcal{P}[\pi]\log\mathcal{P}[\pi]$$

Thus the expected cumulative information gained from the labels of $x_1,\ldots,x_m$, is simply the entropy of the partition $\Pi_m^{\mathcal{F}}$ under the distribution $\mathcal{P}$. We shall denote this entropy by

$$\mathcal{H}^{\mathcal{P}}(\Pi_m^{\mathcal{F}}(\mathbf{x}))=\mathcal{H}_m^{\mathcal{P}}(\mathbf{x})=\mathcal{H}_m^{\mathcal{P}}$$

We may now use this simple expression for the cumulative information gain in conjunction with Jensen's inequality and the chain of inequalities ending with Equation (9) to obtain the following bounds on the expected total number of mistakes made by the Gibbs and Bayes algorithms on the first m trials:

$$\begin{aligned}
\frac{c_0\mathcal{H}_m^{\mathcal{P}}}{\log(2m/\mathcal{H}_m^{\mathcal{P}})} &\le m\mathcal{H}^{-1}(\frac{1}{m}\mathcal{H}_m^{\mathcal{P}}) \\
&= m\mathcal{H}^{-1}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{E}_{f\in\mathcal{P}}[\mathcal{H}(\chi_i(f))]\right) \\
&\le \sum_{i=1}^{m}\mathbf{E}_{f\in\mathcal{P}}[\mathcal{H}^{-1}(\mathcal{H}(\chi_i(f)))] \\
&= \mathbf{E}_{f\in\mathcal{P}}[\sum_{i=1}^{m}Bayes_i(f)] \\
&\le \mathbf{E}_{f\in\mathcal{P}}[\sum_{i=1}^{m}Gibbs_i(f)] \\
&\le \frac{1}{2}\mathbf{E}_{f\in\mathcal{P}}[-\log V_m(f)] \\
&= \frac{1}{2}\mathcal{H}_m^{\mathcal{P}} \qquad (13)
\end{aligned}$$

As in the instantaneous case, the upper and lower bounds here depend on properties of the particular $\mathcal{P}$ and $\mathbf{x}$. Also, analogous to the instantaneous case, when the cumulative

information gain is maximum ($\mathcal{H}_m^{\mathcal{P}} = m$), the upper and lower bounds are tight, and the ratio of the bounds grows logarithmically as the entropy becomes small.

These bounds on learning performance in terms of a partition entropy are of special importance to us, since they will form the crucial link between the Bayesian setting and the Vapnik-Chervonenkis dimension theory.

7 Handling incorrect priors

A common criticism of any Bayesian setting is the assumption of the learner's knowledge of an accurate prior $\mathcal{P}$. Taken to its logical extreme, this objection leads us back to worst-case analysis, whose pitfalls and pessimisms we specifically seek to avoid. However, there is a middle ground: namely, we can assume that the learner's perception of the world (formalized as a *perceived prior*) may differ somewhat from the "truth". In this section we present some initial results in this direction that are based on the information-theoretic techniques developed thus far.

Let us use $\mathcal{Q}$ to denote the *true prior* and $\mathcal{P}$ to denote the *perceived prior*. Then when f is chosen randomly according to $\mathcal{Q}$ but the observer uses the prior $\mathcal{P}$, we obtain the following analogues of Equations (1), (2), and (3):

$$\begin{aligned}
\mathbf{E}_{f\in\mathcal{Q}}[\mathcal{I}_{m+1}^{\mathcal{P}}(f)] &= \mathbf{E}_{f\in\mathcal{Q}}[-\log \chi_{m+1}^{\mathcal{P}}(f)] \\
\mathbf{E}_{f\in\mathcal{Q}}[Bayes_{m+1}^{\mathcal{P}}(f)] &= \mathbf{E}_{f\in\mathcal{Q}}\left[\Theta\left(\frac{1}{2} - \chi_{m+1}^{\mathcal{P}}(f)\right)\right] \\
\mathbf{E}_{f\in\mathcal{Q}}[Gibbs_{m+1}^{\mathcal{P}}(f)] &= \mathbf{E}_{f\in\mathcal{Q}}[1 - \chi_{m+1}^{\mathcal{P}}(f)]
\end{aligned}$$

respectively representing the expected information gain from $f(x_{m+1})$, the probability of mistake for the Bayes algorithm on $f(x_{m+1})$, and the probability of mistake for the Gibbs algorithm on $f(x_{m+1})$.

Since for any $0 < p \leq 1$ we have $\Theta(\frac{1}{2} - p) \leq -\log p$, it follows that

$$\mathbf{E}_{f\in\mathcal{Q}}[Bayes_{m+1}^{\mathcal{P}}(f)] \leq \mathbf{E}_{f\in\mathcal{Q}}[\mathcal{I}_{m+1}^{\mathcal{P}}(f)]$$

Since for any $0 < p \leq 1$ we have $1 - p \leq -\ln p = -\ln(2)\log p$, we have

$$\mathbf{E}_{f\in\mathcal{Q}}[Gibbs_{m+1}^{\mathcal{P}}(f)] \leq \ln(2)\mathbf{E}_{f\in\mathcal{Q}}[\mathcal{I}_{m+1}^{\mathcal{P}}(f)]$$

Thus, the probabilities of a mistake on $f(x_{m+1})$ for both algorithms are bounded above by a small constant times the expected information gain. Note that in this general case in which the prior may be incorrect, the upper bound we get for the Gibbs algorithm is actually slightly better than the upper bound we get for the Bayes algorithm.

We now obtain bounds on the cumulative number of mistakes on the first m trials. By analogy with Equation (12), from the above we may derive

$$\mathbf{E}_{f\in\mathcal{Q}}[\sum_{i=1}^{m} Bayes_i^{\mathcal{P}}(f)] \leq \mathbf{E}_{f\in\mathcal{Q}}[-\log V_m^{\mathcal{P}}(f)]$$

and

$$\mathbf{E}_{f\in\mathcal{Q}}[\sum_{i=1}^{m} Gibbs_i^{\mathcal{P}}(f)] \leq \ln(2)\mathbf{E}_{f\in\mathcal{Q}}[-\log V_m^{\mathcal{P}}(f)]$$

Note that we may write

$$\begin{aligned}\mathbf{E}_{f\in\mathcal{Q}}[-\log V_m^{\mathcal{P}}(f)] &= \mathbf{E}_{f\in\mathcal{Q}}[-\log V_m^{\mathcal{Q}}(f)] + \mathbf{E}_{f\in\mathcal{Q}}[\log \frac{V_m^{\mathcal{Q}}(f)}{V_m^{\mathcal{P}}(f)}] \\ &= \mathcal{H}_m^{\mathcal{Q}} + I_m(\mathcal{Q}||\mathcal{P})\end{aligned}$$

where $\mathcal{H}_m^{\mathcal{Q}}$ is the entropy of the partition $\Pi_m^{\mathcal{F}}(\mathbf{x})$ induced on $\mathcal{F}$ by $x_1, \ldots, x_m$ with respect to $\mathcal{Q}$, and $I_m(\mathcal{Q}||\mathcal{P})$ is the Kullback-Leibler divergence between $\mathcal{Q}$ and $\mathcal{P}$ with respect to this partition.

Our best lower bounds for both the instantaneous mistake probability and the cumulative number of mistakes for the case of an incorrect prior are obtained by observing that the mistake probability is minimized by the Bayes algorithm when $\mathcal{P} = \mathcal{Q}$. Thus $c_0\mathbf{E}_{f\in\mathcal{Q}}[\mathcal{I}_{m+1}^{\mathcal{Q}}(f)]/\log(2/\mathbf{E}_{f\in\mathcal{Q}}[\mathcal{I}_{m+1}^{\mathcal{Q}}(f)])$ is a lower bound on the instantaneous mistake probability, and $c_0\mathcal{H}_m^{\mathcal{Q}}/\log(2m/\mathcal{H}_m^{\mathcal{Q}})$ is a lower bound on the cumulative number of mistakes for both the Bayes and Gibbs algorithms, for any perceived prior $\mathcal{P}$. It would be interesting to obtain lower bounds that incorporate properties of the perceived prior $\mathcal{P}$.

8 The average instantaneous information gain is decreasing

In all of our discussion so far, we have assumed that the instance sequence $\mathbf{x}$ is fixed in advance, but that the target concept f is drawn randomly according to $\mathcal{P}$. We now move to the completely probabilistic model, in which f is drawn according to $\mathcal{P}$, and each instance x_m in the sequence $\mathbf{x}$ is drawn randomly and independently according to a distribution $\mathcal{D}$ over the instance space X.

In this model, we now prove a result that will be used in the next section, but is also of independent interest: namely, that the expected instantaneous information gain $\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\mathcal{I}_m(\mathbf{x}, f)]$ is a non-increasing function of m. (Here we have introduced the notation $\mathbf{x} \in \mathcal{D}^*$ to indicate that each element of the infinite sequence $\mathbf{x}$ is drawn independently according to $\mathcal{D}$.)

We begin by showing that the expected information gain from the first label is at least that of the second. Let us fix the pair of instances $x_1, x_2 \in X$ that are the first two instances seen, but let the *order* of their appearance be chosen uniformly at random. Then we may write

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\{(x_1,x_2),(x_2,x_1)\}}[\mathcal{I}_1(\mathbf{x}, f)] = \frac{1}{2}\mathcal{H}^{\mathcal{P}}(\Pi_1^{\mathcal{F}}(x_1)) + \frac{1}{2}\mathcal{H}^{\mathcal{P}}(\Pi_1^{\mathcal{F}}(x_2))$$

where the subscript $\mathbf{x} \in \{(x_1, x_2), (x_2, x_1)\}$ of the expectation indicates that $\mathbf{x}$ is chosen uniformly at random from these two ordered pairs, and we recall the notation $\mathcal{H}^{\mathcal{P}}(\Pi)$ for the entropy with respect to $\mathcal{P}$ of a partition Π on $\mathcal{F}$.

To obtain an expression for the expected value of $\mathcal{I}_2$ under these same conditions, we use the additivity of information:

$$\begin{aligned}&\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\{(x_1,x_2),(x_2,x_1)\}}[\mathcal{I}_2(\mathbf{x}, f)] \\ &= \mathcal{H}^{\mathcal{P}}(\Pi_2^{\mathcal{F}}(x_1, x_2)) - \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\{(x_1,x_2),(x_2,x_1)\}}[\mathcal{I}_1(\mathbf{x}, f)] \\ &= \mathcal{H}^{\mathcal{P}}(\Pi_2^{\mathcal{F}}(x_1, x_2)) - \frac{1}{2}\mathcal{H}^{\mathcal{P}}(\Pi_1^{\mathcal{F}}(x_1)) - \frac{1}{2}\mathcal{H}^{\mathcal{P}}(\Pi_1^{\mathcal{F}}(x_2))\end{aligned}$$

However, since the partition $\Pi_2^{\mathcal{F}}(x_1, x_2)$ is a refinement of the partitions $\Pi_1^{\mathcal{F}}(x_1)$ and $\Pi_1^{\mathcal{F}}(x_2)$, we have (see e.g. Renyi [25])

$$\mathcal{H}^{\mathcal{P}}(\Pi_2^{\mathcal{F}}(x_1, x_2)) \leq \mathcal{H}^{\mathcal{P}}(\Pi_1^{\mathcal{F}}(x_1)) + \mathcal{H}^{\mathcal{P}}(\Pi_1^{\mathcal{F}}(x_2))$$

Thus

$$\begin{aligned} \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\{(x_1,x_2),(x_2,x_1)\}}[\mathcal{I}_2(\mathbf{x}, f)] &\leq \frac{1}{2}\mathcal{H}^{\mathcal{P}}(\Pi_1^{\mathcal{F}}(x_1)) + \frac{1}{2}\mathcal{H}^{\mathcal{P}}(\Pi_1^{\mathcal{F}}(x_2)) \\ &= \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\{(x_1,x_2),(x_2,x_1)\}}[\mathcal{I}_1(\mathbf{x}, f)] \end{aligned}$$

Since x_1 and x_2 were arbitrary, we may write

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\mathcal{I}_2(\mathbf{x}, f)] \leq \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\mathcal{I}_1(\mathbf{x}, f)]$$

Now for general m, we can compare terms of $\mathcal{I}_m$ and $\mathcal{I}_{m+1}$ by fixing the instances $x_1, \ldots, x_{m-1}$ on this sequence, then applying the above argument to the version space $\mathcal{F}_{m-1}(\langle x_1, \ldots, x_{m-1}\rangle, f)$ and its corresponding posterior $\mathcal{P}_{m-1}$, giving the desired inequality

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\mathcal{I}_{m+1}(\mathbf{x}, f)] \leq \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\mathcal{I}_m(\mathbf{x}, f)]$$

We may apply this result to obtain bounds on the average instantaneous mistake probabilities for the Bayes and Gibbs algorithms on the mth random example in terms of the average entropy of the partition induced by the first m examples. First note that since the total expected information gained by the first m labels is $\mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\mathcal{H}^{\mathcal{P}}(\mathbf{x})]$, with the additivity of information and the above result, we have

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\mathcal{I}_m(\mathbf{x}, f)] \leq \frac{1}{m}\mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\mathcal{H}^{\mathcal{P}}(\mathbf{x})]$$

Thus, using the chain of inequalities ending with Equation (13), we have

$$\begin{aligned} \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Bayes_m(\mathbf{x}, f)] &\leq \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Gibbs_m(\mathbf{x}, f)] \\ &\leq \frac{1}{2m}\mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\mathcal{H}^{\mathcal{P}}(\mathbf{x})] \end{aligned} \tag{14}$$

For the remainder of the paper we shall find it notationally more convenient to discuss the instantaneous mistake probability at trial m (as is done in Equation (14)) rather than at trial $m + 1$.

9 Bayesian learning and the VC dimension: correct priors

Although we have given upper bounds on both the instantaneous probability of mistake and the expected cumulative number of mistakes for the Bayes and Gibbs algorithms in terms of $\mathcal{H}_m^{\mathcal{P}}(\mathbf{x})$, we are still left with the problem of evaluating this entropy, or at least obtaining reasonable upper bounds on it. We can intuitively see that the "worst case" for learning occurs when the partition entropy $\mathcal{H}_m^{\mathcal{P}}(\mathbf{x})$ is as large as possible. In our context, the entropy is qualitatively maximized when two conditions hold:

The instance sequence $\mathbf{x}$ induces a partition of $\mathcal{F}$ that is the largest possible.

The prior $\mathcal{P}$ gives equal weight to each element of this partition.

In this section, we move from our Bayesian average-case setting to obtain worst-case bounds by formalizing these two conditions in terms of combinatorial parameters depending only on the concept class $\mathcal{F}$. In doing so, we form the link between the theory developed so far and the VC dimension theory.

The second of the two conditions above is easily quantified. Since the entropy of a partition is at most the logarithm of the number of classes in it, a trivial upper bound on the entropy which holds for all priors $\mathcal{P}$ is

$$\mathcal{H}_m^{\mathcal{P}}(\mathbf{x}) \quad \leq \quad \log |\Pi_m^{\mathcal{F}}(\mathbf{x})|$$

Now let $\mathcal{D}$ be a distribution on the instance space X and assume that instances in $\mathbf{x}$ are drawn independently at random according to $\mathcal{D}$ as in the previous section. Then using Equation (14) we have that for all $\mathcal{P}$,

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Bayes_m(\mathbf{x},f)] \leq \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Gibbs_m(\mathbf{x},f)] \leq \frac{1}{2m}\mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\log|\Pi_m^{\mathcal{F}}(\mathbf{x})|] \qquad (15)$$

and using Equation (13) that

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\sum_{i=1}^{m} Bayes_i(\mathbf{x},f)] \leq \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\sum_{i=1}^{m} Gibbs_i(\mathbf{x},f)] \leq \frac{1}{2}\mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\log|\Pi_m^{\mathcal{F}}(\mathbf{x})|] \qquad (16)$$

The expectation

$$\mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\log|\Pi_m^{\mathcal{F}}(\mathbf{x})|]$$

is the *VC entropy* defined by Vapnik and Chervonenkis in their seminal paper on uniform convergence [35], and plays a central role in their characterization of the uniform convergence of empirical frequencies to probabilities in a class of events. Here we see how simple information-theoretic arguments can be used to relate the VC entropy to the learning curves of the Bayes and Gibbs algorithms.

In the remainder of this section we will show how the other combinatorial parameter introduced in the paper of Vapnik and Chervonenkis, known in the computational learning theory literature as the Vapnik-Chervonenkis (VC) dimension of the concept class $\mathcal{F}$, can provide useful bounds on the size of $\Pi_m^{\mathcal{F}}(\mathbf{x})$, and how it can be used directly to give bounds on the instantaneous probability of mistake that are independent of the prior $\mathcal{P}$ and the distribution $\mathcal{D}$ on the instance space X.

We say that the instances $x_1,\ldots,x_d \in X$ *shatter* the concept class $\mathcal{F}$ if $|\Pi_m^{\mathcal{F}}(\langle x_1,\ldots,x_d\rangle)| = 2^d$, that is, for every possible labeling of $x_1,\ldots,x_d$ there is some target concept in $\mathcal{F}$ that gives this labeling. For any set $S \subseteq X$, the *Vapnik-Chervonenkis (VC) dimension* of $\mathcal{F}$ on S, denoted $\dim(\mathcal{F},S)$, is the largest d such that there exist instances $x_1,\ldots,x_d \in S$ that shatter $\mathcal{F}$. If arbitrarily long sequences of instances from S shatter $\mathcal{F}$ then $\dim(\mathcal{F},S) = \infty$. Often $S = X$, so we abbreviate $\dim(\mathcal{F},X)$ by $\dim(\mathcal{F})$. Further, if $\mathbf{x} = x_1,x_2,\ldots$ is an infinite sequence of instances from X, for each $m \geq 1$ we use $\dim_m(\mathcal{F},\mathbf{x})$ to denote $\dim(\mathcal{F},\{x_1,\ldots,x_m\})$. Clearly $\dim_m(\mathcal{F},\mathbf{x}) \leq \dim(\mathcal{F})$ for all $\mathbf{x}$ and all m.

The VC dimension has been calculated for many of the fundamental concept classes. For example, if the instance space $X = \Re^n$ and $\mathcal{F}$ is the set of all linear threshold functions on

X then $\dim(\mathcal{F}) = n+1$; if the threshold functions are homogeneous (i.e., the threshold is 0) then $\dim(\mathcal{F}) = n$. If $\mathcal{F}$ is the set of all indicator functions for axis-parallel rectangles in $\Re^n$ then $\dim(\mathcal{F}) = 2n$; also if $\mathcal{F}$ is the set of all indicator functions for n-fold unions of intervals on $X = \Re$ then $\dim(\mathcal{F}) = 2n$. These and many other examples are given in the papers of Dudley [13] and Blumer et al. [4] and elsewhere.

The following important combinatorial result relating $\dim_m(\mathcal{F}, \mathbf{x})$ and $|\Pi_m^{\mathcal{F}}(\mathbf{x})|$ has been proven independently by Sauer [27], Vapnik and Chervonenkis [34], and others (see Assouad [1]): for all $\mathbf{x}$,

$$\log |\Pi_m^{\mathcal{F}}(\mathbf{x})| \leq \log \sum_{i=0}^{\dim_m(\mathcal{F},\mathbf{x})} \binom{m}{i} \leq (1+o(1)) \dim_m(\mathcal{F}, \mathbf{x}) \log \frac{m}{\dim_m(\mathcal{F}, \mathbf{x})} \tag{17}$$

where $o(1)$ is a quantity that goes to zero as $\alpha = m/\dim_m(\mathcal{F}, \mathbf{x})$ goes to infinity. This result can be used directly in conjunction with Equations (15) and (16) to get instantaneous and cumulative mistake bounds. Thus we have that for all $\mathcal{P}$,

$$\begin{aligned} \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Bayes_m(\mathbf{x}, f)] &\leq \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Gibbs_m(\mathbf{x}, f)] \\ &\leq (1+o(1))\mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\frac{\dim_m(\mathcal{F},\mathbf{x})}{2m} \log \frac{m}{\dim_m(\mathcal{F},\mathbf{x})}] \\ &\leq (1+o(1))\frac{\dim(\mathcal{F})}{2m} \log \frac{m}{\dim(\mathcal{F})} \end{aligned} \tag{18}$$

and

$$\begin{aligned} \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\sum_{i=1}^{m} Bayes_i(\mathbf{x}, f)] &\leq \mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\sum_{i=1}^{m} Gibbs_i(\mathbf{x}, f)] \\ &\leq (1+o(1))\mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\frac{\dim_m(\mathcal{F},\mathbf{x})}{2} \log \frac{m}{\dim_m(\mathcal{F},\mathbf{x})}] \\ &\leq (1+o(1))\frac{\dim(\mathcal{F})}{2} \log \frac{m}{\dim(\mathcal{F})} \end{aligned} \tag{19}$$

Haussler, Littlestone and Warmuth [17] (Section 3, latter part) show that specific distributions $\mathcal{D}$ and priors $\mathcal{P}$ can be constructed for each of the classes $\mathcal{F}$ listed above (i.e., (homogeneous) linear threshold functions, indicator functions for axis-parallel rectangles and unions of intervals) for which

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[\sum_{i=1}^{m} Bayes_i(\mathbf{x}, f)] \geq (1-o(1))\frac{\dim(\mathcal{F})}{2} \ln \frac{m}{\dim(\mathcal{F})} \tag{20}$$

This shows that the bound given by Equation (19) is tight to within a factor of $1/\ln(2) \approx 1.44$ in each of these cases and hence cannot be improved by more than this factor in general. It also follows that the expected total number of mistakes of the Bayes and the Gibbs algorithms differ by a factor of at most about 1.44 in each of these cases; this was not previously known. Opper and Haussler [24] give a similar comparison between the instantaneous mistake bounds for the Bayes and Gibbs algorithms for homogeneous linear threshold functions using different priors and instance space distributions. Finally, note that the simplicity of the derivation of the bound in Equation (19) makes this a very appealing way to obtain useful average-case cumulative mistake bounds.

Unfortunately the instantaneous mistake bound given in Equation (18) is not as tight as possible. However, using the results of Haussler, Littlestone and Warmuth [17], we can show that[2] for all $\mathcal{P}$,

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Bayes_m(\mathbf{x},f)] \leq \mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\frac{\dim_m(\mathcal{F},\mathbf{x})}{m}] \leq \frac{\dim(\mathcal{F})}{m} \tag{21}$$

Ignoring the middle bound for the moment, the proof of this fact is straightforward, given the results of Haussler, Littlestone and Warmuth [17] (which are not straightforward to prove, as far as we know). In particular, Theorem 2.3 of that paper shows that for any instance space X and any class $\mathcal{F}$ of concepts on X, there exists a randomized learning algorithm A (the *1-inclusion graph algorithm*) such that for any distribution $\mathcal{D}$ on X and any target concept f in $\mathcal{F}$, when instances $x_1,\ldots,x_m$ are drawn randomly from X according to $\mathcal{D}$ and A is given $(x_1,f(x_1)),\ldots,(x_{m-1},f(x_{m-1}))$ and x_m, the probability that A makes a mistake predicting $f(x_m)$ is at most $\dim(\mathcal{F})/m$. It follows that for any prior $\mathcal{P}$ on $\mathcal{F}$, when f is selected at random according to $\mathcal{P}$, the probability that A makes a mistake predicting $f(x_m)$ is at most $\dim(\mathcal{F})/m$. Thus the probability of a mistake for Bayes algorithm is also at most $\dim(\mathcal{F})/m$, by the optimality of Bayes algorithm. (From a statistical viewpoint, here we are just using the fact that the Bayes risk is always less than the maximum risk of any statistical procedure.)

To prove the middle bound of Equation (21), we can generalize the proof of Haussler, Littlestone and Warmuth's [17] Theorem 2.3 to obtain this sharper, instance space distribution dependent form of the bound for the 1-inclusion graph algorithm for all target concepts, and then apply the argument described in the previous paragraph to obtain the desired result. Alternately, we can also derive the result directly from the lemmas used in establishing their Theorem 2.3. This latter approach is outlined in the discussion section of [16].

From Equation (21) we can also obtain similar upper bounds for the Gibbs algorithm. In particular, using Equation (11) and Equation (21) we have for all $\mathcal{P}$,

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Gibbs_m(\mathbf{x},f)] \leq \mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\frac{2\dim_m(\mathcal{F},\mathbf{x})}{m}] \leq \frac{2\dim(\mathcal{F})}{m} \tag{22}$$

Note that in each of Equations (21) and (22) the second inequality gives a bound that is independent of the distribution $\mathcal{D}$ on the instance space, and of the prior $\mathcal{P}$ on the concept class $\mathcal{F}$.

The same specific distributions and priors constructed by Haussler, Littlestone and Warmuth [17] that we mentioned above also show that for each of the classes $\mathcal{F}$ of (homogeneous) linear threshold functions, indicator functions for axis-parallel rectangles and unions of intervals, there is an instance space distribution $\mathcal{D}$ and a prior $\mathcal{P}$ such that

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Bayes_m(\mathbf{x},f)] \quad \geq \quad (1-o(1))\frac{\dim(\mathcal{F})}{2m}$$

This shows that the bound given by Equation (21) is tight to within a factor of 1/2 in each of these cases and hence cannot be improved by more than this factor in general. We conjecture that in fact the lower bound is correct, and thus the upper bounds in Equations (21) and

[2] Vapnik had obtained the special case of this result for homogeneous linear threshold functions [33]. Also, see [30] for further interesting properties of $\mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[\dim_m(\mathcal{F},\mathbf{x})]$.

(22) can each be improved by a factor of 1/2. It should be noted that if this conjecture holds, then using standard inequalities for partial sums of the harmonic series, the bounds in Equation (21) could be summed to give bounds similar to those in Equation (19), but using ln in place of log. As mentioned above, this bound would be best possible as far as multiplicative constants are concerned.

It is both a strength and a weakness of these bounds that they are given in a form that is independent of the prior $\mathcal{P}$, and possibly also of the distribution $\mathcal{D}$ on the instance space: a strength because the same upper bounds hold for all $\mathcal{P}$ and $\mathcal{D}$, and a weakness because they may not be tight for specific $\mathcal{P}$ and $\mathcal{D}$. While it is always possible to construct degenerate $\mathcal{P}$ and $\mathcal{D}$ for which these upper bounds are far too high, the real question is how far off they are for "typical" or "natural" prior and instance space distributions, as might arise in practice. The distributions used in the lower bounds from the latter part of Section 3 of Haussler, Littlestone and Warmuth [17] mentioned above are unfortunately not very natural. However, in a recent paper [24] the natural case in which $\mathcal{F}$ is the set of homogeneous linear threshold functions on $\Re^d$ and both the distribution $\mathcal{D}$ and the prior $\mathcal{P}$ on possible target concepts (represented also by vectors in $\Re^d$) are uniform on the unit sphere in $\Re^d$ is examined. (For homogeneous linear threshold functions only the directions of the target concept and the instance matter, so the specific choice of the unit sphere is actually immaterial.) In this case, under certain reasonable assumptions used in statistical mechanics, it is shown that for $m >> d >> 1$,

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Bayes_m(\mathbf{x},f)] \approx \frac{0.44d}{m}$$

(compared with the $0.5d/m$ conjectured general upper bound and the d/m proven general upper bound given for any class of VC dimension d above) and, as was previously shown in [15],

$$\mathbf{E}_{f\in\mathcal{P},\mathbf{x}\in\mathcal{D}^*}[Gibbs_m(\mathbf{x},f)] \approx \frac{0.62d}{m}$$

(compared with the $2d/m$ general upper bound proven above). Thus at least in this case, the bounds are still accurate to within a constant factor.

10 Bayesian learning and the VC dimension: incorrect priors

We now look at how the notion of VC dimension can be used to get better bounds on the performance of the Bayes and Gibbs algorithms when the prior $\mathcal{P}$ is incorrect — that is, the target concept is actually chosen at random from some different distribution $\mathcal{Q}$ on $\mathcal{F}$, as in Section 7. Let us say that the prior $\mathcal{P}$ is *nondegenerate* for $\mathcal{F}$ if for any instances $x_1,\ldots,x_m \in X$ and any $f \in \mathcal{F}$, we have $V_m^{\mathcal{P}}(\langle x_1,\ldots,x_m\rangle, f) > 0$, that is, $\mathcal{P}$ never assigns zero probability to any legitimate version space from $\mathcal{F}$. Note that by assigning arbitrarily small probabilities to certain version spaces, the upper bounds given on cumulative mistakes in Section 7 can be made arbitrarily high, even for a nondegenerate prior $\mathcal{P}$. The same holds for the instantaneous mistake bounds. However, the actual probability of mistake, and

expected total number of mistakes in m trials, are trivially bounded by 1 and m respectively, so these bounds cannot be very tight in these extreme cases.

Better-behaved bounds can be obtained using the VC dimension. In particular, in terms of instantaneous mistake bounds, it can be shown that for any nondegenerate prior $\mathcal{P}$, any actual distribution $\mathcal{Q}$ on $\mathcal{F}$, and any distribution $\mathcal{D}$ on the instance space

$$\begin{aligned}\mathbf{E}_{f\in\mathcal{Q},\mathbf{x}\in\mathcal{D}^*}[Gibbs^{\mathcal{P}}_{m+1}(\mathbf{x},f)] &\leq \inf_{k\geq 1}\left(\frac{\ln \mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[|\Pi^{\mathcal{F}}_{m+k}(x_1,\ldots,x_{m+k})|]+1}{k\ln(1+m/k)}+\frac{1}{k}\right)\\ &\leq (1+o(1))\frac{\dim(\mathcal{F})}{m}\ln\frac{m}{\dim(\mathcal{F})} \qquad (23)\end{aligned}$$

where $o(1)$ represents a quantity that goes to zero as $\alpha = m/\dim(\mathcal{F})$ goes to infinity. A similar result holds for the Bayes algorithm, but with an additional factor of 2, giving

$$\begin{aligned}\mathbf{E}_{f\in\mathcal{Q},\mathbf{x}\in\mathcal{D}^*}[Bayes^{\mathcal{P}}_{m+1}(\mathbf{x},f)] &\leq 2\inf_{k\geq 1}\left(\frac{\ln \mathbf{E}_{\mathbf{x}\in\mathcal{D}^*}[|\Pi^{\mathcal{F}}_{m+k}(x_1,\ldots,x_{m+k})|]+1}{k\ln(1+m/k)}+\frac{1}{k}\right)\\ &\leq (1+o(1))\frac{2\dim(\mathcal{F})}{m}\ln\frac{m}{\dim(\mathcal{F})} \qquad (24)\end{aligned}$$

The argument required to establish these bounds is fairly lengthy, and hence is given in the appendix.

Because these bounds do not depend on the distribution $\mathcal{Q}$ used to choose the target concept, they are essentially worst case bounds on the performance of the Bayes and Gibbs algorithms over all possible target concepts in $\mathcal{F}$. Furthermore, the bounds in the second inequalities do not depend on the distribution $\mathcal{D}$ on the instance space X either. If tighter versions of these bounds are desired, the distribution-specific forms given in the middle inequalities may be used.

The middle inequalities also have an interesting consequences when $\mathcal{F}$ if finite. In this case we note that $|\Pi^{\mathcal{F}}_{m+k}(x_1,\ldots,x_{m+k})| \leq |\mathcal{F}|$ for all $x_1,\ldots,x_{m+k}$. Hence

$$\mathbf{E}_{f\in\mathcal{Q},\mathbf{x}\in\mathcal{D}^*}[Gibbs^{\mathcal{P}}_{m+1}(\mathbf{x},f)] \leq \inf_{k\geq 1}\left(\frac{\ln|\mathcal{F}|+1}{k\ln(1+m/k)}+\frac{1}{k}\right) = \frac{\ln|\mathcal{F}|+1}{m} \qquad (25)$$

since $\alpha\ln(1+1/\alpha)\leq 1$ for $\alpha>0$, and $\lim_{\alpha\to\infty}\alpha\ln(1+1/\alpha)=1$. A similar result holds for the Bayes algorithm with an additional factor of two.

A bound similar to that given in Equation (23) is given by Haussler, Littlestone and Warmuth [17], but with a slightly higher constant. As in that paper, it can be shown that the bound given in Equation (23) holds not only for the Gibbs algorithm but for any algorithm that always predicts by finding a hypothesis in $\mathcal{F}$ that is consistent with all the labels of examples it has seen so far (see the appendix). This includes the maximum a posteriori algorithm, which returns the hypothesis with the maximum posterior probability, mentioned in Section 4. Furthermore, a result given in that paper (Theorem 4.2) shows that the leading asymptotic constant of 1 in our bound cannot be improved below $1-1/e$, indicating that for bounds of this generality, this is about the best that can be done. It is unclear how information-theoretic tools, or other VC dimension tools such as those used in obtaining the results of the previous sections, could be used to give stronger versions of this result that depend explicitly on the distributions $\mathcal{P}$ and $\mathcal{Q}$.

11 Learning classes of infinite VC dimension

One limitation of the basic VC dimension analysis given thus far is the assumption that the target concept is drawn from a class of finite VC dimension. Vapnik has extended the theory to include the case when $\mathcal{F}$ has infinite VC dimension, but can be decomposed into a sequence $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ of subclasses with nonzero, finite VC dimensions $d_1, d_2, \ldots$, respectively [34]. A typical decomposition might let $\mathcal{F}_i$ be all neural networks of a given type with at most i weights, in which case $d_i = O(i \log i)$ [3].

We can also look at this from a Bayesian point of view by letting the prior $\mathcal{P}$ be over all concepts in $\mathcal{F}$, and decomposing it as a linear sum $\mathcal{P} = \sum_{i=1}^{\infty} \alpha_i \mathcal{P}_i$, where $\mathcal{P}_i$ is an arbitrary prior over $\mathcal{F}_i$ and $\sum_{i=1}^{\infty} \alpha_i = 1$. We now derive upper bounds on the cumulative number of mistakes and the instantaneous mistake probabilities for the Bayes and Gibbs algorithms by bounding the information gain.

Fix the instance sequence $\mathbf{x}$. As in the analysis of Section 5, we find it convenient to replace the random selection of the target concept $f \in \mathcal{F}$ with a sequence $\mathbf{y} \in \{0,1\}^m$, representing the boolean labels for the first m instances of $\mathbf{x}$. We define $\mathcal{P}_i(\mathbf{y}) = \mathbf{Pr}_{f \in \mathcal{P}_i}[f(x_1) = y_1, \ldots, f(x_m) = y_m]$. This immediately gives $\mathcal{P}(\mathbf{y}) = \sum_{i=1}^{\infty} \alpha_i \mathcal{P}_i(\mathbf{y})$. Letting $\mathcal{H}_m^{\mathcal{P}}$ denote the entropy with respect to $\mathcal{P}$ of the partition induced on $\mathcal{F}$ by $x_1, \ldots, x_m$ (as was done in Section 6), we may write

$$\begin{aligned}
\mathcal{H}_m^{\mathcal{P}} &= -\sum_{\mathbf{y} \in \{0,1\}^m} \mathcal{P}(\mathbf{y}) \log \mathcal{P}(\mathbf{y}) \\
&= -\sum_{\mathbf{y} \in \{0,1\}^m} \left(\sum_{i=1}^{\infty} \alpha_i \mathcal{P}_i(\mathbf{y}) \right) \log \left(\sum_{i=1}^{\infty} \alpha_i \mathcal{P}_i(\mathbf{y}) \right) \\
&\leq -\sum_{\mathbf{y} \in \{0,1\}^m} \sum_{i=1}^{\infty} \alpha_i \mathcal{P}_i(\mathbf{y}) \log \alpha_i \mathcal{P}_i(\mathbf{y}) \\
&= -\sum_{\mathbf{y} \in \{0,1\}^m} \left(\sum_{i=1}^{\infty} \alpha_i \mathcal{P}_i(\mathbf{y}) \log \alpha_i + \sum_{i=1}^{\infty} \alpha_i \mathcal{P}_i(\mathbf{y}) \log \mathcal{P}_i(\mathbf{y}) \right) \\
&= -\left(\sum_{i=1}^{\infty} \alpha_i \log \alpha_i \sum_{\mathbf{y} \in \{0,1\}^m} \mathcal{P}_i(\mathbf{y}) \right) - \left(\sum_{i=1}^{\infty} \alpha_i \sum_{\mathbf{y} \in \{0,1\}^m} \mathcal{P}_i(\mathbf{y}) \log \mathcal{P}_i(\mathbf{y}) \right) \\
&= -\sum_{i=1}^{\infty} \alpha_i \log \alpha_i + \sum_{i=1}^{\infty} \alpha_i \mathcal{H}_m^{\mathcal{P}_i}
\end{aligned}$$

Here we have used the fact $-\log(x+y) \leq \min(-\log x, -\log y)$. The final expression obtained shows an interesting decomposition: the sum $-\sum_{i=1}^{\infty} \alpha_i \log \alpha_i$ is simply the entropy of the infinite sequence of $\alpha = \alpha_1, \alpha_2, \ldots$, which we shall denote $\mathcal{H}(\alpha)$. The sum $\sum_{i=1}^{\infty} \alpha_i \mathcal{H}_m^{\mathcal{P}_i}$ is a sum of the entropies of the component distributions $\mathcal{P}_i$, weighted by the contribution of each component to $\mathcal{P}$. Now from Equation (13) we may immediately write

$$\begin{aligned}
\mathbf{E}_{f \in \mathcal{P}}[\sum_{i=1}^{m} Bayes_i(f)] &\leq \mathbf{E}_{f \in \mathcal{P}}[\sum_{i=1}^{m} Gibbs_i(f)] \\
&\leq \frac{1}{2} \left(\mathcal{H}(\alpha) + \sum_{i=1}^{\infty} \alpha_i \mathcal{H}_m^{\mathcal{P}_i} \right)
\end{aligned}$$

Recall from Section 9 that $\mathcal{H}_m^{\mathcal{P}}(\mathbf{x}) \leq \log |\Pi_m^{\mathcal{F}}(\mathbf{x})|$ for any $\mathcal{F}$ and any prior $\mathcal{P}$ on $\mathcal{F}$. By using a variant of Sauer's lemma (Equation 17), it can be shown that if the VC dimension of $\mathcal{F}$ is $d \geq 1$, then

$$\log |\Pi_m^{\mathcal{F}}(\mathbf{x})| \leq \sum_{i=0}^{d} \binom{m}{i} \leq \log(m^d + 1) \leq d \log(m+1)$$

for all $m \geq 1$. Hence

$$\mathcal{H}_m^{\mathcal{P}_i} \leq d_i \log(m+1) \text{ for all } i.$$

Combined with the above, this yields[3]

$$\begin{aligned}
\mathbf{E}_{f \in \mathcal{P}}[\sum_{i=1}^{m} Bayes_i(f)] &\leq \mathbf{E}_{f \in \mathcal{P}}[\sum_{i=1}^{m} Gibbs_i(f)] \\
&\leq \frac{1}{2}\left(\mathcal{H}(\alpha) + \sum_{i=1}^{\infty} \alpha_i d_i \log(m+1)\right) \\
&= \frac{1}{2}\left(\mathcal{H}(\alpha) + \log(m+1) \sum_{i=1}^{\infty} \alpha_i d_i\right)
\end{aligned}$$

We may interpret this final bound as follows: the term $\mathcal{H}(\alpha)$ can be regarded as a "penalty" for our uncertainty as to which $\mathcal{F}_i$ the target will be drawn from. Provided the sequence of α_i decreases more rapidly than $\frac{1}{i \log i}$ (roughly), this penalty will be only a constant number of mistakes. The term $\log(m+1)\sum_{i=1}^{\infty} \alpha_i d_i$ is the usual logarithmic bound times a kind of VC dimension, only now this dimension is actually a kind of "effective VC dimension" $\sum_{i=1}^{\infty} \alpha_i d_i$, where the contribution of each d_i is proportional to the weight α_i of $\mathcal{F}_i$. This is the dominant term in the final bound, and will result in a cumulative mistake bound that is logarithmic in m provided that $\sum_{i=1}^{\infty} \alpha_i d_i$ is finite.

Finally, we may obtain bounds on the instantaneous mistake probabilities in the setting where each instance in $\mathbf{x}$ is drawn randomly according to $\mathcal{D}$ by applying Equation (14), giving

$$\begin{aligned}
\mathbf{E}_{f \in \mathcal{P}, \mathbf{x} \in \mathcal{D}^*}[Bayes_m(\mathbf{x}, f)] &\leq \mathbf{E}_{f \in \mathcal{P}, \mathbf{x} \in \mathcal{D}^*}[Gibbs_m(\mathbf{x}, f)] \\
&\leq \frac{\mathcal{H}(\alpha)}{2m} + \frac{1}{2m}\mathbf{E}_{\mathbf{x} \in \mathcal{D}^*}\left[\sum_{i=1}^{\infty} \alpha_i \mathcal{H}_m^{\mathcal{P}_i}(\mathbf{x})\right] \\
&\leq \frac{\mathcal{H}(\alpha)}{2m} + \frac{\log(m+1)}{2m} \sum_{i=1}^{\infty} \alpha_i d_i
\end{aligned}$$

12 Conclusions and future research

Perhaps the most important general conclusion to be drawn from the work presented here is that the various theories of learning curves based on diverse ideas from information theory, statistical physics and the VC dimension are all in fact closely related, and can be naturally and beneficially placed in a common Bayesian framework.

[3]Somewhat stronger, but more complex upper bounds can be obtained by using more refined upper bounds on $\sum_{i=0}^{d} \binom{m}{i}$.

The focus of our ongoing research is that of making the basic theory presented here more applicable to the situations encountered by practitioners of machine learning in neural networks, artificial intelligence, and other areas. Below we briefly mention some extensions of our model for which we have partial results:

Learning with noise. Here we extend many of our general results relying on information-theoretic notions to handle the case where the classification labels may be corrupted by noise.

Learning multi-valued functions. Here we relax the restriction that the target function have $\{0,1\}$-valued output to allow multiple possible output values. These results can be used to study the learning of real-valued functions, which is often the situation in empirical neural network research.

Learning with other loss functions. In conjunction with the above extension, here we seek to generalize the theory by studying measures of a learning algorithm's performance other than the $\{0,1\}$-loss function studied here. A typical choice is the *quadratic loss*, often used to obtain the standard sum-of-squared-errors measure for real-valued or vector-valued functions.

Acknowledgements

We are greatly indebted to Manfred Opper and Ron Rivest for their valuable suggestions and guidance, and to Sara Solla and Naftali Tishby for insightful ideas in the early stages of this investigation. We also thank Andrew Barron, Andy Kahn, Nick Littlestone, Phil Long, Terry Sejnowski and Haim Sompolinsky for stimulating discussions on these topics. This research was supported by ONR grant N00014-91-J-1162, AFOSR grant AFOSR-89-0506, ARO grant DAAL03-86-K-0171, DARPA contract N00014-89-J-1988, and a grant from the Siemens Corporation.

Appendix

Here we give the derivation of Equations (23) and (24). First we will need to establish a few lemmas.

Lemma 1 *Let M be an arbitrary n by m matrix of 0s and 1s. Suppose that t of the m columns of M are selected at random without replacement and eliminated, along with all rows of M that have a 1 in any of these columns. Let M' be the remaining matrix. Let the random variable ψ denote the maximum number of 1s in any row of M', or 0 if M' is empty. Then*

$$\mathbf{E}(\psi) \leq \frac{\ln n + 1}{\ln(m/(m-t))} \; + \; 1$$

where the expectation is over the random choice of the t columns.

Proof: Let $k = m - t$. Clearly $0 \leq \psi \leq k$. For each j, $1 \leq j \leq k$, let p_j be the probability that $\psi \geq j$. Then

$$\mathbf{E}(\psi) = \sum_{j=1}^{k} p_j$$

Now fix j, and fix a particular row of M that has $r \geq j$ 1s. If we choose t of the m columns at random and eliminate all rows that contain a 1 in any of these columns, then the probability that this row is not eliminated is

$$\frac{\binom{m-r}{t}}{\binom{m}{t}} = \frac{(m-r)!k!}{(k-r)!m!} = \frac{k(k-1)\cdots(k-r+1)}{m(m-1)\cdots(m-r+1)} \leq \left(\frac{k}{m}\right)^r \leq \left(\frac{k}{m}\right)^j$$

Hence the probability that there is any row of M with j or more 1s that is not eliminated is at most $\min(1, n(k/m)^j)$. Since $\psi \geq j$ only if there is a row of M with j or more 1s that is not eliminated, it follows that

$$p_j \leq \min(1, n(k/m)^j)$$

Thus

$$\begin{aligned}
\mathbf{E}(\psi) &\leq \sum_{j=1}^{k} \min(1, n(k/m)^j) \\
&\leq s + n \sum_{j=s+1}^{\infty} (k/m)^j \quad \text{for any } s \geq 0 \\
&= s + n\frac{(k/m)^{s+1}}{1-(k/m)} \\
&= s + \frac{nk}{m-k} e^{-s\ln(m/k)}
\end{aligned}$$

Let s be the least integer greater than

$$\frac{\ln(\frac{nk}{m-k}\ln(m/k))}{\ln(m/k)}$$

Making this substitution and simplifying, we obtain

$$\mathbf{E}(\psi) \leq \frac{\ln n + \ln((k/(m-k))\ln(m/k)) + 1}{\ln(m/k)} + 1$$

Since $\ln(x) \leq x - 1$ for all $x > 0$, we have

$$\ln(m/k) \leq (m/k) - 1 = (m-k)/k,$$

and thus $\ln((k/(m-k))\ln(m/k)) < 0$. It follows that

$$\mathbf{E}(\psi) \leq \frac{\ln n + 1}{\ln(m/k)} + 1$$

giving the result. □

Lemma 2 *Let $\mathcal{P}$ be a nondegenerate prior distribution on $\mathcal{F}$. Let $x_1, \ldots, x_m$ be any sequence of instances in the instance space X and f be any (unknown) target concept in $\mathcal{F}$. Suppose that $t+1$ of the m instances $x_1, \ldots, x_m$ are selected uniformly at random without replacement, we are given the values of f on the first t of these instances, and we are asked to predict the value of f on the last instance. Then if we use the Gibbs learning algorithm with prior $\mathcal{P}$, or indeed any learning algorithm that always predicts by selecting a hypothesis in $\mathcal{F}$ that is consistent with all the examples it has seen so far, the probability that we predict incorrectly is at most*

$$\frac{\ln n + 1}{(m-t)\ln(m/(m-t))} + \frac{1}{m-t}$$

where $n = |\Pi_m^{\mathcal{F}}(x_1, \ldots, x_m)|$. Furthermore, if we use the Bayes algorithm with prior $\mathcal{P}$, the probability that we predict incorrectly is at most twice this value.

Proof: Choose a representative $f_i \in \mathcal{F}$ for each equivalence class of $\Pi_m^{\mathcal{F}}(x_1, \ldots, x_m)$, for $1 \leq i \leq n$. Define the n by m matrix M by letting $M_{i,j} = 1$ if $f_i(x_j) = f(x_j)$ and $M_{i,j} = 0$ otherwise. Thus each row in M indicates for which instances in $x_1, \ldots, x_m$ the functions in the ith equivalence class will predict the wrong label. In particular, the row representing the equivalence class of f itself is all 0s.

Let us assume that the instances $x_{j_1}, \ldots, x_{j_{t+1}}$ are chosen at random without replacement from $x_1, \ldots, x_m$ and that we are given the value of f on the first t of these chosen instances. Consider the problem of predicting the value of f on $x_{j_{t+1}}$. Suppose we are using a learning algorithm that predicts by choosing a hypothesis $\hat{f}$ from $\mathcal{F}$ that is consistent with the labels it has seen so far, that is, $\hat{f}(x_{j_1}) = f(x_{j_1}), \ldots, \hat{f}(x_{j_t}) = f(x_{j_t})$. The Gibbs algorithm is one such algorithm. Since all that matters as far as mistakes in prediction on points in $x_1, \ldots, x_m$ is concerned is the equivalence class of the hypothesis chosen, any such learning algorithm corresponds to choosing a row i in M with a 0 in each of the t columns $j_1, \ldots, j_t$. Now since the $t+1$st instance $x_{j_{t+1}}$ is randomly chosen from among the $m-t$ instances left after the first t instances are chosen, the probability (with respect to the choice of this $t+1$st random instance but fixing the choice of the first t instances) that the label of the $t+1$st instance is predicted incorrectly is $r_i/(m-t)$, where r_i is the number of 1s in the row i of M chosen by the algorithm.

Let M' be the matrix obtained from M by eliminating the t columns $j_1, \ldots, j_t$, and eliminating any row that has a 1 in any of these columns. Note that M' is nonempty since M has an all 0 row. Then for any row i chosen by a consistent learning algorithm we have $r_i/(m-t) \leq \psi/(m-t)$, where ψ is the maximum number of 1s in any row of M'. It follows that the probability (with respect to the random choice of all $t+1$ instances) that the label of this $t+1$st instance is predicted incorrectly is at most $\mathbf{E}(\psi)/(m-t)$, where the expectation is with respect to the random choice of the first t instances. By the previous lemma,

$$\frac{\mathbf{E}(\psi)}{m-t} \leq \frac{\ln n + 1}{(m-t)\ln(m/(m-t))} + \frac{1}{m-t}$$

This gives the first result.

For the second result, again assume that the instances $x_{j_1}, \ldots, x_{j_t}$ are the first t instances selected at random (without replacement) from $x_1, \ldots, x_m$ and define the matrix M' as above. Given the labels $f(x_{j_1}), \ldots, f(x_{j_t})$, let $\mathcal{P}_t$ be the posterior distribution induced on $\mathcal{F}$ as defined in Section 4. For each i let p_i denote the probability, with respect to $\mathcal{P}_t$, of the

equivalence class represented by row i of the matrix M'. Since $\mathcal{P}$ is nondegenerate, $p_i > 0$ for each row i of M'.

Let us define the *mistake weight* $\rho(j)$ of column j of M' by letting

$$\rho(j) = \sum_i p_i M_{i,j}$$

Thus $\rho(j)$ is the total posterior probability of all rows that have a 1 in column j. Note that a mistake is made by the Bayes algorithm in predicting the label of the $t+1$st random instance $x_{j_{t+1}}$ with probability 1 if the mistake weight $\rho(j_{t+1}) > 1/2$, and with probability $1/2$ if $\rho(j_{t+1}) = 1/2$. Thus this probability of a mistake on the $t+1$st random instance is at most $\gamma/(m-t)$, where γ is the number of columns in M' with mistake weight at least $1/2$.

Let us define the *total mistake weight* ρ of M' by $\rho = \sum_j \rho(j)$. Since the number of columns with mistake weight at least $1/2$ is at most twice the total mistake weight of all columns, we have $\gamma \leq 2\rho$. However, since $\rho = \sum_{i,j} p_i M_{i,j} = \sum_i p_i r_i$, where r_i is the number of 1s in row i of M', it is also clear that $\rho \leq \psi$, where ψ is the maximum number of 1s in any row of M'. Hence, the probability of a mistake for the Bayes algorithm on the $t+1$st random instance is at most $2\psi/(m-t)$. The remainder of the proof is as above. □

Theorem 3 *Let $\mathcal{P}$ be a nondegenerate prior on $\mathcal{F}$ and $\mathcal{Q}$ be any distribution on $\mathcal{F}$. Let $\mathcal{D}$ be a distribution on X. Assume the target function f is drawn at random from $\mathcal{F}$ according to $\mathcal{Q}$. Suppose that $t+1$ instances are selected independently at random with replacement from X according to $\mathcal{D}$. Assume we are given the values of f on the first t of these instances, and we are asked to predict the value of f on the last instance. Then if we use the Gibbs learning algorithm, or indeed any learning algorithm that always predicts by selecting a hypothesis in $\mathcal{F}$ that is consistent with all the examples it has seen so far, the probability that we predict incorrectly is at most*

$$\inf_{k \geq 1} \left(\frac{\ln \mathbf{E}_{\mathbf{x} \in \mathcal{D}^{t+k}}(\Pi^{\mathcal{F}}_{t+k}(\mathbf{x})) + 1}{k \ln(1 + t/k)} + \frac{1}{k} \right)$$

If we use the Bayes algorithm, the probability that we predict incorrectly is at most twice this value. Further, if $d = \dim(\mathcal{F}) < \infty$, then this value is at most

$$(1 + o(1)) \frac{d}{t} \ln \frac{t}{d}$$

where $o(1)$ represents a quantity that goes to zero as t/d goes to infinity.

Proof: Fix $k \geq 1$ and let $m = t + k$. Fix the target concept $f \in \mathcal{F}$. The previous lemma shows that for any fixed sequence $\mathbf{x} = (x_1, \ldots, x_{t+k})$ of instances from X, if we randomly select $t+1$ of these, and use the labels of the first t to predict the label of the $t+1$st, then using any consistent learning algorithm, the probability we predict incorrectly is at most

$$\frac{\ln n + 1}{(m-t)\ln(m/(m-t))} + \frac{1}{m-t} = \frac{\ln n + 1}{k \ln(1 - t/k)} + \frac{1}{k} \tag{26}$$

where $n = |\Pi^{\mathcal{F}}_{t+k}(\mathbf{x})|$. Since this bound holds for any fixed sequence $\mathbf{x} \in X^{t+k}$, it also holds if the x_is in $\mathbf{x}$ are drawn independently with replacement from any distribution on X, when n is replaced with $\mathbf{E}_{\mathbf{x} \in \mathcal{D}^{t+k}}(|\Pi^{\mathcal{F}}_{t+k}(\mathbf{x})|)$. However, when $x_1, \ldots, x_{t+k}$ are drawn independently

with replacement from some fixed distribution $\mathcal{D}$ and then $t+1$ of these $t+k$ instances are selected at random (without replacement), the overall distribution on the set of all possible sequences of the resulting $t+1$ instances is the same if they were directly selected from $\mathcal{D}$ independently with replacement. Hence, for each $k \geq 1$, the value (26) above is a bound on the probability of a mistake in predicting the label of the last instance in a sequence of $t+1$, drawn independently with replacement, given the labels of the first t variables. Finally, since this bound holds for any target $f \in \mathcal{F}$, it also holds in expectation when the target f is selected at random according to any distribution $\mathcal{Q}$ on $\mathcal{F}$. This gives the first result of the theorem. The argument is similar for the result about the Bayes algorithm, using the second part of the previous lemma.

To establish the last result, note that by Sauer's lemma (Equation (17)),

$$\ln \mathbf{E}_{\mathbf{x} \in \mathcal{D}^{t+k}}(\Pi^{\mathcal{F}}_{t+k}(\mathbf{x})) \leq (1+o(1))d \ln \frac{t+k}{d}$$

Let $k = \lceil t \ln(t/d) \rceil$. Then

$$\begin{aligned}
\frac{\ln \mathbf{E}_{\mathbf{x} \in \mathcal{D}^{t+k}}(\Pi^{\mathcal{F}}_{t+k}(\mathbf{x})) + 1}{k \ln(1+t/k)} + \frac{1}{k} &\leq \frac{(1+o(1))d \ln \frac{t+t\ln(t/d)}{d}}{t \ln(t/d) \ln(1+1/\ln(t/d))} \\
&= \frac{(1+o(1))d \ln(t/d)}{t \ln(t/d) \ln(1+1/\ln(t/d))} \\
&= \frac{(1+o(1))d}{t \ln(1+1/\ln(t/d))} \\
&= (1+o(1))\frac{d}{t} \ln \frac{t}{d}
\end{aligned}$$

This gives the result. □

Note that the trick employed in the proof above of varying the additional number of instances k to get better averages has also been used in [28] and [22] to get other bounds on related measures based on the VC dimension.

References

[1] P. Assouad. Densité et dimension. *Annales de l'Institut Fourier*, 33(3):233–282, 1983.

[2] J. M. Barzdin and R. V. Freivald. On the prediction of general recursive functions. *Soviet Mathematics-Doklady*, 13:1224–1228, 1972.

[3] E. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1(1):151–160, 1989.

[4] 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.

[5] W. Buntine. *A Theory of Learning Classification Rules.* PhD thesis, University of Technology, Sydney, 1990.

[6] W. Buntine and A. Weigend. Bayesian back propagation. Unpublished manuscript, 1991.

[7] B. Clarke and A. Barron. Entropy, risk and the Bayesian central limit theorem. manuscript.

[8] B. Clarke and A. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

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

[10] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel, and J. Hopfield. Automatic learning, rule extraction and generalization. *Complex Syst.*, 1:877–922, 1987.

[11] A. DeSantis, G. Markowski, and M. N. Wegman. Learning probabilistic prediction functions. In *Proceedings of the 1988 Workshop on Computational Learning Theory*, pages 312–328. Morgan Kaufmann, 1988.

[12] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[13] R. M. Dudley. A course on empirical processes. *Lecture Notes in Mathematics*, 1097:2–142, 1984.

[14] R. Fano. Class notes for course 6.574. Technical report, Massachusetts Institute of Technology, 1952.

[15] G. Gyorgyi and N. Tishby. Statistical theory of learning a rule. In K. Thuemann and R. Koeberle, editors, *Neural Networks and Spin Glasses*. World Scientific, 1990.

[16] D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. Technical Report UCSC-CRL-91-41, University of Calif. Computer Research Laboratory, Santa Cruz, CA, 1991.

[17] D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. Technical Report UCSC-CRL-90-54, University of California Santa Cruz, Computer Research Laboratory, Dec. 1990. To appear in Information and Computation.

[18] N. Littlestone. *Mistake Bounds and Logarithmic Linear-threshold Learning Algorithms*. PhD thesis, University of California Santa Cruz, 1989.

[19] N. Littlestone, P. M. Long, and M. K. Warmuth. On-line learning of linear functions. In *Proceedings of the Twenty Third Annual ACM Symposium on Theory of Computing*, pages 465–475, 1991.

[20] N. Littlestone and M. Warmuth. The weighted majority algorithm. Technical Report UCSC-CRL-89-16, Computer Research Laboratory, University of Santa Cruz, July 1989.

[21] D. MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1992.

[22] P. Massart. Rates of convergence in the central limit theorem for empirical processes. *Annales de l'Institut H. Poincaré Probab. Statist.*, 22:381–423, 1986.

[23] B. K. Natarajan. Learning over classes of distributions. In *Proceedings of the 1988 Workshop on Computational Learning Theory*, pages 408–409, San Mateo, CA, 1988. published by Morgan Kaufmann.

[24] M. Opper and D. Haussler. Calculation of the learning curve of Bayes optimal classification algorithm for learning a perceptron with noise. In *Computational Learning Theory: Proceedings of the Fourth Annual Workshop*, pages 75–87. Morgan Kaufmann, 1991.

[25] A. Renyi. *Probability Theory*. North Holland, Amsterdam, 1970.

[26] W. Sarrett and M. Pazzani. Average case analysis of empirical and explanation-based learning algorithms. Technical Report 89-35, UC Irvine, 1989. To appear in Machine Learning.

[27] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (Series A)*, 13:145–147, 1972.

[28] J. Shawe-Taylor, M. Anthony, and N. Biggs. Bounding sample size with the Vapnik-Chervonenkis dimension. Technical Report CSD-TR-618, University of London, Surrey, England, 1989.

[29] H. Sompolinsky, N. Tishby, and H. Seung. Learning from examples in large neural networks. *Phys.Rev.Lett.*, 65:1683–1686, 1990.

[30] M. Talagrand. Donsker classes of sets. *Probability Theory and Related Fields*, 78:169–191, 1988.

[31] N. Tishby, E. Levin, and S. Solla. Consistent inference of probabilities in layered networks: predictions and generalizations. In *IJCNN International Joint Conference on Neural Networks*, volume II, pages 403–409. IEEE, 1989.

[32] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–42, 1984.

[33] V. N. Vapnik. *Theorie der Zeichenerkennung*. Akademie-Verlag, 1979.

[34] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

[35] 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.

[36] V. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383. Morgan Kaufmann, 1990.