

How Well do Bayes Methods Work for On-Line Prediction of $\{\pm 1\}$ values?

D. Haussler

Baskin Center for Computer Engineering and Information Sciences
University of California, Santa Cruz, CA 95064
haussler@cse.ucsc.edu

A. Barron

713 Wright Street
Dept. of Statistics, U. Ill.
Champaign IL 61820
barron@andrew.stat.uiuc.edu

Technical Report UCSC-CRL-92-37
Computer and Information Sciences
University of California at Santa Cruz
July, 1992

Abstract

We look at sequential classification and regression problems in which $\{\pm 1\}$ -labeled instances are given on-line, one at a time, and for each new instance, before seeing the label, the learning system must either predict the label, or estimate the probability that the label is $+1$. We look at the performance of Bayes method for this task, as measured by the total number of mistakes for the classification problem, and by the total log loss (or information gain) for the regression problem. Our results are given by comparing the performance of Bayes method to the performance of a hypothetical “omniscient scientist” who is able to use extra information about the labeling process that would not be available in the standard learning protocol. The results show that Bayes methods perform only slightly worse than the omniscient scientist in many cases. These results generalize previous results of Haussler, Kearns and Schapire, and Opper and Haussler.

1 Introduction

Several recent papers in the area of computational learning theory have studied sequential classification problems in which $\{\pm 1\}$ -labeled instances (examples) are given on-line, one at a time, and for each new instance, the learning system must predict the label before it sees it [HLW90, Lit89, LW89, Vov90b, HKS91, OH91a, SST92, MF92]. Such systems adapt on-line, learning to make better predictions as they see more examples. If n is the total number of examples, then the performance of these on-line learning systems, as a function of n , has been measured both by the total number of mistakes (incorrect predictions) they make during learning, and by the probability of a mistake on the n th example alone. The latter function is often called a *learning curve* (see also [HKLW91]).

Sequential regression problems have also been studied [Daw84, Dawa, Dawb, Vov90a, Vov92, Yam91, Yam92, Ama92, AFS92, SST92, MF92]. In this case, instead of predicting either $+1$ or -1 , the learning system outputs a probability distribution, predicting that the label will be $+1$ with a certain probability, and -1 with one minus that probability. When there is some noise or uncertainty in the labeling process, an output of this type is more informative than a simple prediction of either $+1$ or -1 . The notion that the purpose of statistical inference is to make sequential probability forecasts for future observations, rather than to extract information about parameters, is known as the *prequential approach* in statistics [Daw84]. To measure the performance of a sequential regression system of this type, it is common to use the log loss function. If you predict that the label will be $+1$ with probability p and -1 with probability $1 - p$, then your log loss is $-\log p$ if the label is $+1$ and $-\log(1 - p)$ if the label is -1 , i.e. whatever happens, your loss is the negative logarithm of the probability you assigned to that event. As in sequential classification, performance has been measured both in terms of total log loss over all examples, and in terms of expected loss on the n th example (and in several other ways as well).

In this paper we look at the performance of Bayes methods for both classification and regression, analyzing only the total loss over all examples. Our viewpoint is decision theoretic, so both classification and regression are treated in a common framework. When the examples are generated randomly, the average total loss of a method will be called the *risk*. Bayes methods are optimal in that they have the smallest possible risk, at least when the examples are generated randomly by the same process implicit in the prior distribution used by the method. When the examples are not generated at random by this process, then the performance of Bayes methods degrade. However, we show that they still perform well in many cases.

We do this by introducing the idea of an “omniscient scientist” who is privy to extra information about the nature of the process that generates the examples, and comparing the performance of Bayes method to that of the omniscient scientist. For example, if each label is generated by applying a fixed function to the instance and then adding noise to the result (i.e. flipping the sign of the value with some probability), then the omniscient scientist will already know the fixed “target function” and will only have to deal with the noise, whereas

the Bayes method will have to try to learn the target function and also deal with the noise. We show that Bayes method does not perform much worse than the omniscient scientist in many cases of this type. In particular, the total loss of the omniscient scientist is usually linear in n , whereas the additional loss of Bayes method is only logarithmic in n . We obtain upper bounds on this additional loss that generalize related bounds obtained in [HKS91] and [OH91a] (see also [Ama92, AFS92, SST92]). We also look at the performance of the Bayes method on an arbitrary sequence of examples, as compared with the performance of an omniscient scientist who already knows the best target function to use in predicting the labels of that particular sequence of examples. Again, in many cases Bayes methods do not do much worse. These results extend work in [Vov90b, LW89], and also ties in with the coding/information theory approach in [MF92, FMG92, FM92] (see also [Yu]). Throughout the paper, our emphasis is on obtaining performance bounds that hold for all sample sizes n , rather than asymptotic bounds that hold only for large n .

2 Formal Framework

Here we outline the general decision theoretic framework we use. Let X , Y and A be sets, called the *instance*, *outcome*, and *decision* spaces, respectively, and let $L : Y \times A \rightarrow \mathfrak{R}$ be a *loss function*. In this paper we assume that $Y = \{\pm 1\}$, although the basic formal framework, definition of Bayes method, and the results for log loss hold for more general Y . When $Y = \{\pm 1\}$, elements of Y may be thought of as classification labels. However, because we sometimes consider more general Y , we will use the more general term “outcomes”. The particular kinds of loss functions we consider are the following.

1. *0-1 loss* (used for the classification problem): $A = Y = \{\pm 1\}$, action $\hat{y} \in A$ is interpreted as a prediction of the outcome $y \in Y$, and $L(y, \hat{y}) = 1$ if $\hat{y} \neq y$, $L(y, \hat{y}) = 0$ if $\hat{y} = y$.
2. *log loss* (used for the regression problem): Here instead of predicting a single outcome, an action specifies a probability for each possible outcome $y \in Y$. The decision space A is the family of all probability distributions on Y , and for $y \in Y$ and distribution $P \in A$, $L(y, P) = -\log P(y)$. The base of the logarithm can be arbitrarily chosen. If we need to be specific, we use the notation \ln and \log_2 to denote the natural logarithm and the logarithm base two, respectively.

A pair (x, y) , with $x \in X$ and $y \in Y$ is called an *example*. We assume that we receive a sequence $S^n = (x_1, y_1), \dots, (x_n, y_n)$ of examples on line, one at a time. The number n of examples is called the *sample size*. For each time t , $1 \leq t \leq n$, given only $(x_1, y_1), \dots, (x_{t-1}, y_{t-1})$ and x_t , we must choose an action $a_t \in A$. After taking this action, we observe the outcome y_t , and suffer loss $L(y_t, a_t)$. Our goal is to choose actions a_1, \dots, a_n so as to minimize our total loss $\sum_{t=1}^n L(y_t, a_t)$.

Throughout most of this paper we focus on the case when the sequence x_1, \dots, x_n of instances is fixed arbitrarily and only the outcomes y_1, \dots, y_n vary. If one wants to use these results for the case when both the instances and the outcomes vary, one can either average over possible sequences x_1, \dots, x_n , or take the worst case over such sequences, depending on what type of result one desires. For now let us fix the sequence of instances x_1, \dots, x_n . To simplify the notation in what follows, we will no longer mention the x_t s in our formulae, focusing only on the sequence y_1, \dots, y_n of outcomes. For a particular sequence y_1, \dots, y_n , for further brevity, we define $y^t = (y_1, \dots, y_t)$ for every $0 \leq t \leq n$. Thus y^0 denotes the empty sequence.

Finally, for a random variable X , we denote by $\mathbf{E}(X)$ the expectation of X . We use P, P_θ , etc. to denote probability distributions. For random variables X and Y , we denote a conditional distribution on Y given that $X = x$ by $P_{Y|x}$. If the distribution is not clear from the context when we take an expectation, then we make it explicit by subscripting, as in $\mathbf{E}_{P_\theta}(Y)$ or $\mathbf{E}_{P_{Y|x}}(Y)$. The latter is abbreviated $\mathbf{E}(Y|x)$.

2.1 Priors

In this paper we concentrate on Bayes methods for choosing the actions a_1, \dots, a_n to try to minimize the total loss on the outcomes y_1, \dots, y_n . Bayes methods utilize *prior* information on which sequences of outcomes are more likely than others in order to choose appropriate actions. For fixed instance sequence x_1, \dots, x_n this prior information consists of a *prior* distribution P over an (arbitrary) *index set* Θ , and a class $\{P_\theta : \theta \in \Theta\}$ of probability distributions over the set Y^n of possible outcome sequences. (When Θ is continuous, P is a density.) Each distribution P_θ actually represents a conditional distribution on possible outcome sequences y_1, \dots, y_n , given that the instance sequence is x_1, \dots, x_n . However, since this instance sequence is fixed for now, to avoid cluttering our notation, we omit this implicit conditioning on the x_t s.

The prior information used by a Bayes method can be interpreted as the belief that the sequence of outcomes is generated at random in the following manner. First an index θ is selected at random from Θ according to the prior distribution P . The index θ is viewed as an unknown underlying “state of Nature” that determines the probabilities of various sequences of outcomes via the corresponding distribution P_θ . After θ is chosen, the actual outcome sequence $y^n = y_1, \dots, y_n$ is chosen at random from Y^n according to P_θ . Thus the outcome y_t can be considered to be a realization of the random variable Y_t , $1 \leq t \leq n$, where Y_1, \dots, Y_n are (not usually independent) random variables with joint distribution defined by the above two step process. Note that implicit in this model is the assumption that the action taken at the current time t does not affect the outcome at time t , nor do past actions influence future instances or outcomes. Thus the model studied in this paper is much more appropriate for problems like predicting the weather than for learning to fly an airplane.

Even though they implicitly make very specific assumptions about how outcomes are generated, Bayes methods can be applied whether or not outcome sequences are really gen-

erated in the assumed manner or an equivalent manner. We evaluate the performance of Bayes methods both under the optimistic assumption that outcome sequences are generated in the manner described above, and under certain more pessimistic assumptions. In the extreme case, even though the Bayes method uses a prior distribution, we analyze the performance of the method assuming nothing about the way the actual outcome sequence is generated [Daw84, Vov90b, LW89].

We are often interested in certain special types of prior information that may be available to help choose an appropriate action. The type of prior information available determines the kind of learning problem one has. Three special cases of interest are described below, in order of increasing generality. We present results for some of these special cases later.

Case 1: Noise-free functions. In this case, for each state of nature $\theta \in \Theta$ there is a function $f_\theta : X \rightarrow Y$. Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. By assuming that initially a state of nature $\theta \in \Theta$ is chosen at random according to the prior P , we are in effect assuming that a *target function* f_θ is chosen at random from \mathcal{F} according to the induced prior on \mathcal{F} . After the target function is chosen, for any instance $x \in X$, the outcome y is $f_\theta(x)$, independent of any previous instances, outcomes and actions, i.e. a state of nature is deterministic and noise-free. Thus for any θ and any fixed sequence of instances x_1, \dots, x_n , the outcome sequence y_1, \dots, y_n is completely determined: the distribution P_θ assigns probability 1 to the event $(y_1, \dots, y_n) = (f_\theta(x_1), \dots, f_\theta(x_n))$. The performance of Bayes methods for this case was studied in [HKS91, OH91b].

Case 2: Functions corrupted by i.i.d. noise. Here a state of nature is represented by a function, but the observations are altered by an independent noise process. Thus, as in Case 1, for each $\theta \in \Theta$ there is a possible target function $f_\theta : X \rightarrow Y$, and some f_θ is chosen at random according to the prior P . However, for the instance sequence x_1, \dots, x_n , instead of observing the outcome sequence $(y_1, \dots, y_n) = (f_\theta(x_1), \dots, f_\theta(x_n))$, we observe $(y_1, \dots, y_n) = (\eta_1 f_\theta(x_1), \dots, \eta_n f_\theta(x_n))$, where the (unobserved) *noise events* η_1, \dots, η_n are independent and identically distributed, with $\eta_t = -1$ with probability λ and $\eta_t = +1$ with probability $1 - \lambda$, the *noise rate* λ being known. Since for a given θ and instance sequence x_1, \dots, x_n , many different outcome sequences are possible, here P_θ is not a trivial distribution on Y^n like it was in Case 1. The performance of Bayes method for this case was studied in [OH91a] for a particular class of functions.

Case 3: Conditionally independent Y_t s. In this case, the random variables Y_1, \dots, Y_n are conditionally independent given θ and x_1, \dots, x_n (and completely independent of the actions taken). This includes the previous case, as well as the more general cases in which either the noise rate is not known but we have a prior distribution over possible noise rates, or the noise events are independent but not identically distributed. This latter case occurs, for example, if the distribution for the noise event η_t depends on the instance x_t , i.e. observations of the outcome are less reliable for some instances than they are for others. One way to capture this is with the statistical regression setup in which for each $\theta \in \Theta$ there is a distribution \mathcal{D}_θ on $X \times Y$, and after θ is chosen, examples $(x_1, y_1), \dots, (x_n, y_n)$ are chosen independently according to \mathcal{D}_θ . Thus for a given θ and x_1, \dots, x_n , the random variables

Y_1, \dots, Y_n are independent, and the distribution of Y_t is the marginal of D_θ for $X = x_t$. For this case P_θ is the joint distribution of these Y_{i_s} .

2.2 Definition of Bayes Method

We now return to the most general case where Θ indexes an arbitrary set $\{P_\theta : \theta \in \Theta\}$ of distributions on Y^n , i.e. arbitrary joint distributions on the random variables Y_1, \dots, Y_n . In this section we derive the general form of Bayes method for this case. To simplify our formulae, we use the following notation.

- $P(\theta)$ denotes the prior probability of state $\theta \in \Theta$, or the density at θ if the prior is given as a density function.
- For all t , $0 \leq t \leq n$, and $\theta \in \Theta$,

$$P(y^t|\theta) = P_\theta(y^t) = P_\theta\{(\hat{y}_1, \dots, \hat{y}_n) : \hat{y}_i = y_i, 1 \leq i \leq t\}.$$

- For any $y \in Y$,

$$\begin{aligned} P(Y_t = y|y^{t-1}, \theta) &= P_\theta(Y_t = y|y^{t-1}) \\ &= P_\theta\{(\hat{y}_1, \dots, \hat{y}_n) : \hat{y}_i = y_i, 1 \leq i \leq t-1, \hat{y}_t = y\} / P_\theta(y^{t-1}) \end{aligned}$$

(assuming that $P_\theta(y^{t-1}) \neq 0$, else it is undefined.).

- $P(y^t) = \sum_{\theta \in \Theta} P(y^t|\theta)P(\theta)$ if Θ is countable, otherwise $P(y^t) = \int_{\theta \in \Theta} P(y^t|\theta)P(\theta)d\theta$.
- For any $y \in Y$, $P(Y_t = y|y^{t-1}) = \sum_{\theta \in \Theta} P(Y_t = y|y^{t-1}, \theta)P(\theta|y^{t-1})$ if Θ is countable, otherwise $P(Y_t = y|y^{t-1}) = \int_{\theta \in \Theta} P(Y_t = y|y^{t-1}, \theta)P(\theta|y^{t-1})d\theta$.

Note that $P(\theta|y^t)$, used above, is calculated by Bayes rule:

$$P(\theta|y^t) = \frac{P(\theta)P(y^t|\theta)}{P(y^t)}.$$

Given the above notation, Bayes method of choosing actions can be stated quite simply:

At each time t , choose the action $a \in A$ that minimizes $\sum_{y \in Y} P(Y_t = y|y^{t-1})L(y, a)$.

The logic of this is simple. If your belief that outcome sequences are generated at random in the two step process described above is correct, then $P(Y_t = y|y^{t-1})$ is the probability that the t^{th} outcome will be y , given that the previous $t-1$ outcomes were y_1, \dots, y_{t-1} . This is called the *posterior* probability of y (having seen y^{t-1}). Hence $\sum_{y \in Y} P(Y_t = y|y^{t-1})L(y, a)$ is the expected loss you will suffer if you take action a at time t (the *posterior expected loss*). Bayes method is simply to choose the action that minimizes the (posterior) expected loss.

Bayes method leads to familiar strategies for both the 0-1 and log losses. Here we describe the action a_t taken by Bayes method in each of these cases for a general outcome space Y .

1. *0-1 loss*:

$$\begin{aligned} a_t = \hat{y}_t &= \operatorname{argmin}_{\hat{y} \in Y} \sum_{y \in Y, y \neq \hat{y}} P(Y_t = y | y^{t-1}) \\ &= \operatorname{argmax}_{\hat{y} \in Y} P(Y_t = \hat{y} | y^{t-1}) \end{aligned} \quad (1)$$

Hence in this case Bayes method predicts the outcome that has the highest posterior probability.

2. *log loss*:

$$\begin{aligned} a_t = \hat{P}_t &= \operatorname{argmin}_{\hat{P} \in A} \sum_{y \in Y} P(Y_t = y | y^{t-1}) \log \frac{1}{\hat{P}(y)} \\ &= P_{Y^t | y^{t-1}} \quad (\text{whenever this distribution is in } A) \end{aligned} \quad (2)$$

The latter equality follows from the fact that the relative entropy¹ is minimal between a distribution and itself (see e.g. [CT91]). Hence in this case Bayes method simply produces the posterior distribution on Y_t as its action.

Since Bayes method always chooses the action that minimizes the posterior expected loss, it is clear that when the actual outcome sequence is in fact generated by the two step random process implicit in the Bayes prior, then the expected loss at each time t is minimized by this strategy, among all possible prediction strategies. Hence the expected total loss is also minimized by the Bayes method. The expected (total) loss is known as *Bayes risk*, and denoted

$$R(P) = R_{\text{bayes}}(P) = \sum_{y^n \in Y^n} P(y^n) L_{\text{bayes}, P}^T(y^n), \quad (3)$$

where $L_{\text{bayes}, P}^T(y^n)$ is the total loss on the outcome sequence y_1, \dots, y_n when the actions taken are those of the Bayes method using prior P .

2.3 Evaluating Bayes Performance: Omniscient Scientists

Because the Bayes method is optimal in terms of the risk when outcome sequences are drawn according to the given prior, the Bayes risk is a lower bound on the risk of any strategy for choosing actions in this case. In the following sections we give some estimates of the Bayes risk, as a function of the sample size n , in some common cases. However, before proceeding with this, we define a few more pessimistic types of risks we want to look at.

First, let us still assume that the true underlying “state of Nature” is some $\theta^* \in \Theta$, and that the outcome sequence y^n is chosen at random according to the distribution P_{θ^*} .

¹See section 3.2 for a definition of relative entropy.

However, let us not assume that θ^* itself is actually chosen at random. Rather, for each possible $\theta \in \Theta$, we define the risk *when θ is true* by

$$r(\theta) = r_{\text{bayes}}(\theta) = \sum_{y^n \in Y^n} P(y^n|\theta)L_{\text{bayes},P}^T(y^n). \quad (4)$$

Thus $r(\theta)$ is the average total loss of Bayes method using an (implicit) prior P over possible states of Nature, when the outcome sequences are in fact generated randomly according to the particular state of nature θ . Of course, Bayes method does not minimize the risk in this case (i.e. for a particular $\theta = \theta^*$). To minimize the risk for a particular θ^* , we would require an *omniscient scientist (OS)* who somehow knew at the outset, before any examples were given, that θ^* was the true state of nature. To obtain optimal risk for this particular θ^* , the omniscient scientist would then use a Bayes method in which the prior distribution P over the index set Θ assigns probability 1 to θ^* and probability 0 to everything else. Hence the omniscient scientist is also a Bayesian, but a better informed one. We denote the total loss on the outcome sequence y_1, \dots, y_n when the actions taken are those of the omniscient scientist using true state θ^* by $L_{\text{OS},\theta^*}^T(y^n)$. Similarly, we denote the risk of the omniscient scientist when θ^* is true by

$$r_{\text{OS}}(\theta^*) = \sum_{y^n \in Y^n} P(y^n|\theta^*)L_{\text{OS},\theta^*}^T(y^n). \quad (5)$$

Note that for noise-free functions, the risk $r_{\text{OS}}(\theta^*)$ of the omniscient scientist is zero for any reasonable loss function, since in this case knowledge of the true state of nature θ^* allows one to predict the outcomes perfectly. In the case that Θ indexes a set of functions corrupted by i.i.d. noise, the risk $r_{\text{OS}}(\theta^*)$ is simply n times the average loss of predicting one noisy outcome, knowing the distribution of that outcome. Surprisingly, we show below that in many cases the risk of the original Bayes method, which does not know θ^* , is not much worse than that of the omniscient scientist, no matter what θ^* is the true state of Nature.

Finally, we might be much more pessimistic, and assume nothing whatsoever about the actual outcome sequence y^n . We can simply look directly at the total loss $L_{\text{bayes},P}^T(y^n)$ of the Bayes method (using prior P) for each fixed outcome sequence y^n . Of course, it is usually the case that for any outcome sequence y^n there is a strategy for choosing actions that does extremely well on that particular sequence. So it is uninteresting (and perhaps unfair) to compare the performance of Bayes algorithm on each particular sequence to the performance of the best strategy for that sequence. However, if we again restrict ourselves to the type of omniscient scientist introduced above, then we do get some interesting results. In this case the omniscient scientist is not as omniscient as she could be, i.e. she doesn't know beforehand what the outcome sequence y^n will be; rather, she knows which state of nature $\theta \in \Theta$ is the best one to assume true for the particular sequence of outcomes y^n that is about to happen, i.e. she knows $\hat{\theta} = \hat{\theta}(y^n) = \operatorname{argmin}_{\theta \in \Theta} L_{\text{OS},\theta}^T(y^n)$. The risk for the OS in this case is

$$L_{\text{OS},\hat{\theta}}^T(y^n) = \min_{\theta \in \Theta} L_{\text{OS},\theta}^T(y^n).$$

Again surprisingly, we show below that in many cases the risk of the original Bayes method is not much worse than that of the omniscient scientist, no matter what the actual outcome sequence y^n is.

One final note: In defining $r(\theta)$, $R(P)$, etc. in this and the previous section, we have assumed that the particular loss function being used is clear from the context. If this is not the case, then a subscript will be used to denote the loss function, as in R_{\log} or R_{0-1} . Other subscripts may be dropped if they are clear from the context.

3 Results for Log Loss

Throughout this section the loss function is assumed to be the log loss. All of the results in this section, except for the specific applications mentioned in the last subsection, hold when Y is an arbitrary countable set. By changing to densities in appropriate places, they hold also for continuous Y .

3.1 Performance on Arbitrary Outcome Sequences

We begin with the most pessimistic case, assuming nothing about the outcome sequence y^n . As we have noted above, for the log loss scenario, Bayes method simply returns the posterior distribution as its action. Thus the action taken at time t is

$$a_t = P_{Y^t|y^{t-1}}.$$

Hence the total loss of Bayes method is

$$\begin{aligned} L_{\text{bayes},P}^T(y^n) &= \sum_{t=1}^n L(y_t, a_t) \\ &= -\sum_{t=1}^n \log P(y_t|y^{t-1}) \\ &= -\log \prod_{t=1}^n P(y_t|y^{t-1}) \\ &= -\log P(y^n). \end{aligned} \tag{6}$$

Hence the total loss is the information gained by seeing the outcome sequence y^n . This simple information theoretic interpretation of the total loss is what makes the log loss so useful.

We want to compare the total loss of Bayes method to that of an omniscient scientist who already knows the best state of nature $\hat{\theta} \in \Theta$ to assume for predicting the outcome sequence y^n , even before the outcomes are observed.

For each time t , the omniscient scientist returns the distribution

$$a_t = \hat{P}_t = P_{Y^t|y^{t-1},\hat{\theta}}.$$

Hence the total loss of the omniscient scientist is

$$\begin{aligned}
L_{\text{OS},\hat{\theta}}^T(y^n) &= \sum_{t=1}^n L(y_t, a_t) \\
&= -\sum_{t=1}^n \log P(y_t|y^{t-1}, \hat{\theta}) \\
&= -\log \prod_{t=1}^n P(y_t|y^{t-1}, \hat{\theta}) \\
&= -\log P(y^n|\hat{\theta}).
\end{aligned} \tag{7}$$

The state $\hat{\theta}$ used by the omniscient scientist is the best possible for the sequence y^n , i.e. $\hat{\theta} = \hat{\theta}(y^n) = \operatorname{argmin}_{\theta \in \Theta} \{L_{\text{OS},\theta}^T(y^n)\} = \operatorname{argmin}_{\theta \in \Theta} \{-\log P(y^n|\theta)\} = \operatorname{argmax}_{\theta \in \Theta} P(y^n|\theta)$.

Hence in the case of the log loss, $\hat{\theta}$ is the *maximum likelihood estimate (MLE)* of the “true” state of nature, based on the (as yet unseen) outcome sequence y^n . (Even though in this section we do not assume there really is a “true” state of nature.)

We focus now on the difference between the Bayes loss and the loss of the omniscient scientist. This difference is given by

$$\begin{aligned}
L^\Delta(y^n) &\doteq L_{\text{bayes},P}^T(y^n) - L_{\text{OS},\hat{\theta}}^T(y^n) \\
&= \log \frac{P(y^n|\hat{\theta})}{P(y^n)}.
\end{aligned} \tag{8}$$

Let us first assume that Θ is countable. Then from the above, we have

$$\begin{aligned}
L^\Delta(y^n) &= \log \frac{P(y^n|\hat{\theta})}{P(y^n)} \\
&= \log \frac{P(y^n|\hat{\theta})}{\sum_{\theta \in \Theta} P(y^n|\theta)P(\theta)} \\
&\leq \log \frac{P(y^n|\hat{\theta})}{P(y^n|\hat{\theta})P(\hat{\theta})} \\
&= \log \frac{1}{P(\hat{\theta})}
\end{aligned} \tag{9}$$

Thus the additional loss suffered by the Bayes method is at most the extra information provided to the omniscient scientist, namely the number of bits needed to describe the MLE $\hat{\theta}$ with respect to the prior P . This observation was made in [DMW88]. However, it may be something of a “folk theorem” in the statistics/information theory community.

The above argument cannot be applied if Θ is uncountable and P is a distribution on Θ unless P puts positive mass at the point $\hat{\theta} \in \Theta$ (and hence cannot be represented as a

density), or P assigns positive probability to the set of all θ that are MLEs, in the case that the MLE $\hat{\theta}$ is not unique. Furthermore, for both countable and uncountable Θ , even if the prior distribution does put positive mass on the MLE $\hat{\theta}$, the above estimate ignores the beneficial effect of other $\theta \in \Theta$ that may also give the outcome sequence y^n relatively high probability, and thereby help the Bayes method to perform better on y^n . We now derive some better upper bounds that overcome these shortcomings. These upper bounds can also be applied in the case of countable Θ , although we state them only in the continuous form here. We begin with the following observation.

$$\begin{aligned}
L^\Delta(y^n) &= \log \frac{P(y^n|\hat{\theta})}{P(y^n)} \\
&= \log \frac{P(y^n|\hat{\theta})}{\int_{\theta \in \Theta} P(y^n|\theta)P(\theta)d\theta} \\
&= -\log \int_{\theta \in \Theta} \frac{P(y^n|\theta)}{P(y^n|\hat{\theta})} P(\theta)d\theta \\
&= -\log \int_0^1 P\{\theta : \frac{P(y^n|\theta)}{P(y^n|\hat{\theta})} \geq z\} dz \tag{10}
\end{aligned}$$

$$\begin{aligned}
&= -\log \int_0^1 P\{\theta : \frac{1}{n} \ln \frac{P(y^n|\hat{\theta})}{P(y^n|\theta)} \leq \frac{-\ln z}{n}\} dz \\
&= -\log \int_0^\infty P\{\theta : \frac{1}{n} \ln \frac{P(y^n|\hat{\theta})}{P(y^n|\theta)} \leq x\} n e^{-nx} dx \tag{11}
\end{aligned}$$

Step (10) follows from the fact that $\mathbf{E}(Z) = \int_0^M P\{Z \geq z\} dz$ for any random variable Z taking values in the interval $[0, M]$, and the last step follows by a simple change of variable.

Let

$$\mathcal{N}_r(\hat{\theta}, y^n) = \{\theta : \frac{1}{n} \ln \frac{P(y^n|\hat{\theta})}{P(y^n|\theta)} \leq r\}$$

and

$$g(r) = P(\mathcal{N}_r(\hat{\theta}, y^n))$$

for each $r \geq 0$. Intuitively, for small r we may think of $\mathcal{N}_r(\hat{\theta}, y^n)$ as a kind of “neighborhood” around the MLE $\hat{\theta}$ in which other $\theta \in \Theta$ live who assign similar probabilities to the outcome sequence y^n . When this neighborhood has large enough prior probability $g(r)$ for small enough “radius” r , then Bayes method will work well for that y^n . For larger r , the negative exponential term in the integral (11) dominates, and the contribution is negligible. What radius is small enough depends on the rate at which $g(r)$ grows. Since $g(r)$ is nondecreasing, we can use the estimate

$$\int_0^\infty g(x) n e^{-nx} dx \geq \sup_{r \geq 0} \{g(r) \int_r^\infty n e^{-nx} dx\}$$

$$= \sup_{r \geq 0} \{g(r)e^{-nr}\} \quad (12)$$

to obtain (from (11) above)

$$\begin{aligned} L^\Delta(y^n) &\leq -\log \sup_{r \geq 0} \{g(r)e^{-nr}\} \\ &= \inf_{r \geq 0} \{-\log(g(r)e^{-nr})\} \\ &= \inf_{r \geq 0} \{nr - \log(g(r))\} \\ &= \inf_{r \geq 0} \{nr - \log P(\mathcal{N}_r(\hat{\theta}, y^n))\} \end{aligned} \quad (13)$$

Note that when $g(0) > 0$, i.e. when the prior puts positive mass on the set $\mathcal{N}_0(\hat{\theta}, y^n)$ of all θ that are MLEs for y^n , then we can use the estimate

$$L^\Delta(y^n) \leq -\log P(\mathcal{N}_0(\hat{\theta}, y^n)) \quad (14)$$

This gives a slightly more general version of the “folk theorem” (9).

Another interesting case is when $g(0) = 0$ but the prior probability $g(r) = P(\mathcal{N}_r(\hat{\theta}, y^n))$ grows with the radius r at least as fast as a volume of radius r in k -dimensional space for some k . This can happen when Y is continuous and the index set Θ gives a smooth enough finite dimensional real vector-valued parameterization of the class $\{P_\theta : \theta \in \Theta\}$ (see e.g. [Ris86, CB90, Dawa, Yam92]). In particular, let us assume that $g(0) = 0$ but there exist $c, k, \epsilon_0 > 0$ such that $g(r) \geq cr^k$ for all $0 \leq r \leq \epsilon_0$. In this case from (13) we have

$$L^\Delta(y^n) \leq \inf_{0 \leq r \leq \epsilon_0} \{nr - k \log r - \log c\} \quad (15)$$

Differentiating with respect to r , we find that the infimum is obtained for $r = k/n$. When $k/n \leq \epsilon_0$, this gives

$$L^\Delta(y^n) \leq k + k \log \frac{n}{k} - \log c = (1 + o(1))k \log n, \quad (16)$$

where $o(1)$ is a quantity that goes to zero as the sample size $n \rightarrow \infty$.

3.2 Performance on Random Sequences Assuming True State of Nature

We now look at the risk (i.e. average total loss) of Bayes method when the outcome sequence is generated at random according to an unknown true state of nature $\theta^* \in \Theta$, as compared

to the risk for an omniscient scientist who knows the true state θ^* . By the same argument used in the previous section, the difference in these two risks is

$$\begin{aligned}
r^\Delta(\theta^*) &\doteq r_{\text{bayes}}(\theta^*) - r_{OS}(\theta^*) \\
&= \sum_{y^n \in Y^n} P(y^n | \theta^*) L_{\text{bayes}, P}^T(y^n) - \sum_{y^n \in Y^n} P(y^n | \theta^*) L_{OS, \theta^*}^T(y^n) \\
&= \sum_{y^n \in Y^n} P(y^n | \theta^*) (L_{\text{bayes}, P}^T(y^n) - L_{OS, \theta^*}^T(y^n)) \\
&= \sum_{y^n \in Y^n} P(y^n | \theta^*) \log \frac{P(y^n | \theta^*)}{P(y^n)} \\
&\doteq I(P_{Y^n | \theta^*} \parallel P_{Y^n}), \tag{17}
\end{aligned}$$

where $I(P \parallel Q) = \mathbf{E}_P \log \frac{P(X)}{Q(X)}$ denotes the *relative entropy* or *Kullback-Leibler divergence* between distribution P and distribution Q [Kul59].

Continuing the analogy with the previous section, for each $r \geq 0$ let

$$\mathcal{N}_r(\theta^*) = \{\theta : I(P_{Y^n | \theta^*} \parallel P_{Y^n | \theta}) \leq rn\}.$$

Hence now we define a neighborhood $\mathcal{N}_r(\theta^*)$ around θ^* instead of a neighborhood around $\hat{\theta}(y^n)$, and this neighborhood includes all $\theta \in \Theta$ that assign probabilities to outcomes y^n that are similar to the probabilities assigned by θ^* , at least on y^n that are likely under θ^* . Now for countable Θ and any $r \geq 0$

$$\begin{aligned}
I(P_{Y^n | \theta^*} \parallel P_{Y^n}) &= \sum_{y^n \in Y^n} P(y^n | \theta^*) \log \frac{P(y^n | \theta^*)}{P(y^n)} \\
&= \sum_{y^n \in Y^n} P(y^n | \theta^*) \log \frac{P(y^n | \theta^*)}{\sum_{\theta \in \Theta} P(y^n | \theta) P(\theta)} \\
&\leq \sum_{y^n \in Y^n} P(y^n | \theta^*) \log \frac{P(y^n | \theta^*)}{\sum_{\theta \in \mathcal{N}_r(\theta^*)} P(y^n | \theta) P(\theta)} \\
&= \sum_{y^n \in Y^n} P(y^n | \theta^*) \log \frac{P(y^n | \theta^*)}{\sum_{\theta \in \mathcal{N}_r(\theta^*)} \frac{P(\theta)}{P(\mathcal{N}_r(\theta^*))} P(y^n | \theta)} - \log P(\mathcal{N}_r(\theta^*)) \\
&\leq \sum_{y^n \in Y^n} P(y^n | \theta^*) \sum_{\theta \in \mathcal{N}_r(\theta^*)} \frac{P(\theta)}{P(\mathcal{N}_r(\theta^*))} \log \frac{P(y^n | \theta^*)}{P(y^n | \theta)} - \log P(\mathcal{N}_r(\theta^*)) \tag{18} \\
&= \sum_{\theta \in \mathcal{N}_r(\theta^*)} \frac{P(\theta)}{P(\mathcal{N}_r(\theta^*))} I(P_{Y^n | \theta^*} \parallel P_{Y^n | \theta}) - \log P(\mathcal{N}_r(\theta^*)) \\
&\leq rn - \log P(\mathcal{N}_r(\theta^*)), \tag{19}
\end{aligned}$$

where (18) follows from Jensen's inequality, using the convexity of $\log \frac{1}{x}$, and (19) follows from the definition of $\mathcal{N}_r(\theta^*)$. The same result follows similarly for continuous Θ (see [Bar87]).

From (17) and (19), in analogy with (13), we have

$$r^\Delta(\theta^*) \leq \inf_{r \geq 0} \{rn - \log P(\mathcal{N}_r(\theta^*))\}. \quad (20)$$

3.3 Bayes Risk

Finally we look at the Bayes risk. This is the expected cumulative loss of Bayes method when the outcome sequence is generated in the manner specified by the prior P , namely, a true state of nature $\theta^* \in \Theta$ is selected at random according to the prior P , and then a sequence y^n of observations is generated at random according to P_{θ^*} . In keeping with the philosophy of the previous sections, we compare the Bayes risk with the risk of an omniscient scientist who knows θ^* before seeing y^n . Thus we define

$$R_{OS}(P) \doteq \sum_{\theta^* \in \Theta} P(\theta^*) r_{OS}(\theta^*) \quad (21)$$

and

$$\begin{aligned} R^\Delta(P) &\doteq R_{\text{bayes}}(P) - R_{OS}(P) \\ &= \sum_{\theta^* \in \Theta} P(\theta^*) (r_{\text{bayes}}(\theta^*) - r_{OS}(\theta^*)) \\ &= \sum_{\theta^* \in \Theta} P(\theta^*) I(P_{Y^n|\theta^*} \| P_{Y^n}) \\ &\doteq \mathcal{I}(\Theta; Y^n), \end{aligned} \quad (22)$$

where $\mathcal{I}(X; Y) = \sum_{x \in X} P_X(x) I(P_{Y|x} \| P_Y)$ denotes the *mutual information* between the random variables X and Y . Here we view the state of nature $\theta \in \Theta$ and the outcome sequence $y^n \in Y^n$ as dependent random variables with the joint distribution defined (implicitly) by the prior P and the set $\{P_\theta : \theta \in \Theta\}$.

Let Π be any partition of Θ into a countable sequence $\Theta_1, \Theta_2, \dots$ of pairwise disjoint subsets of Θ with $\bigcup_i \Theta_i = \Theta$. Each Θ_i is called an *equivalence class*. For each $\theta \in \Theta$, let $\Pi(\theta)$ denote the equivalence class containing θ .

The *entropy* of Π is given by

$$\mathcal{H}(\Pi) = -\mathbf{E}_P \log P(\Pi(\theta)) = -\sum_{i=1}^{\infty} P(\Theta_i) \log P(\Theta_i).$$

(The entropy can be infinite.) The *average diameter* of Π is defined by

$$\mathcal{D}(\Pi) = \sum_{i=1}^{\infty} P(\Theta_i) \sup_{\theta^*, \theta \in \Theta_i} I(P_{Y^n|\theta^*} \| P_{Y^n|\theta}).$$

Using tricks like those used in the previous section, for any countable Θ and any partition Π we get

$$\begin{aligned}
\mathcal{I}(\Theta; Y^n) &= \sum_{\theta^* \in \Theta} P(\theta^*) \sum_{y^n \in Y^n} P(y^n | \theta^*) \log \frac{P(y^n | \theta^*)}{P(y^n)} \\
&= \sum_{\theta^* \in \Theta} P(\theta^*) \sum_{y^n \in Y^n} P(y^n | \theta^*) \log \frac{P(y^n | \theta^*)}{\sum_{\theta \in \Theta} P(y^n | \theta) P(\theta)} \\
&\leq \sum_{\theta^* \in \Theta} P(\theta^*) \sum_{y^n \in Y^n} P(y^n | \theta^*) \log \frac{P(y^n | \theta^*)}{\sum_{\theta \in \Pi(\theta^*)} P(y^n | \theta) P(\theta)} \\
&= \sum_{\theta^* \in \Theta} P(\theta^*) \left(\sum_{y^n \in Y^n} P(y^n | \theta^*) \log \frac{P(y^n | \theta^*)}{\sum_{\theta \in \Pi(\theta^*)} \frac{P(\theta)}{P(\Pi(\theta^*))} P(y^n | \theta)} - \log P(\Pi(\theta^*)) \right) \\
&\leq \sum_{\theta^* \in \Theta} P(\theta^*) \left(\sum_{\theta \in \Pi(\theta^*)} \frac{P(\theta)}{P(\Pi(\theta^*))} \sum_{y^n \in Y^n} P(y^n | \theta^*) \log \frac{P(y^n | \theta^*)}{P(y^n | \theta)} - \log P(\Pi(\theta^*)) \right) \\
&= \left(\sum_{\theta^* \in \Theta} P(\theta^*) \sum_{\theta \in \Pi(\theta^*)} \frac{P(\theta)}{P(\Pi(\theta^*))} I(P_{Y^n | \theta^*} \| P_{Y^n | \theta}) \right) + \mathcal{H}(\Pi) \\
&\leq \mathcal{D}(\Pi) + \mathcal{H}(\Pi).
\end{aligned} \tag{23}$$

The same result also holds for continuous Θ .

From (22) and (23) we have

$$R^\Delta(P) \leq \inf_{\text{partitions } \Pi \text{ of } \Theta} \{\mathcal{D}(\Pi) + \mathcal{H}(\Pi)\}. \tag{24}$$

3.4 Applications

To upper bound the total risk of Bayes method under the log loss, we can simply calculate the risk of the omniscient scientist, and add to it the upper bounds on the difference between the risk of the Bayes method and that of the omniscient scientist. This method is usually easy to use, because the risk for the omniscient scientist is usually easy to calculate. In particular, whenever the Y_t s are conditionally independent given θ , then the risk of the omniscient scientist in each of the cases treated in the previous three subsections is:

- for arbitrary y^n : $L_{\text{OS}, \hat{\theta}}^T(y^n) = -\sum_{t=1}^n \log P(y_t | \hat{\theta})$,
- for particular true state of nature θ^* : $r_{\text{OS}}(\theta^*) = \sum_{t=1}^n \mathcal{H}(Y_t | \theta^*)$, and
- for random state of nature drawn by P : $R_{\text{OS}}(P) = \sum_{\theta^* \in \Theta} P(\theta^*) \sum_{t=1}^n \mathcal{H}(Y_t | \theta^*)$,

where $\mathcal{H}(Y_t|\theta^*) = -\sum_{y \in Y} P(Y_t = y|\theta^*) \log P(Y_t = y|\theta^*)$. Thus, for example, for learning $\{\pm 1\}$ -valued functions with i.i.d. sign noise with rate λ ,

$$R_{\text{os}}(P) = r_{\text{os}}(\theta^*) = h_\lambda n, \quad (25)$$

where $h_\lambda = -\lambda \log \lambda - (1 - \lambda) \log(1 - \lambda)$, and

$$L_{\text{os}, \hat{\theta}}^T(y^n) = N \log \frac{1}{\lambda} + (n - N) \log \frac{1}{1 - \lambda}, \quad (26)$$

where N is the minimum number of noisy outcomes in any interpretation of y^n , i.e. $N = \min_{\theta \in \Theta} |\{t : y_t \neq f_\theta(x_t)\}|$.

3.4.1 Finite Θ

Let us illustrate this method with a simple application of the “folk theorem” (9). Suppose that Θ is finite and the prior distribution P is uniform over Θ . Then by (9), for any outcome sequence y^n ,

$$L_{\text{bayes}}^\Delta(y^n) \leq \log \frac{1}{P(\hat{\theta})} = \log |\Theta|. \quad (27)$$

Hence for Θ indexing a class of $\{\pm 1\}$ -valued functions with i.i.d. sign noise with rate λ ,

$$L_{\text{bayes}}^T(y^n) \leq N \log \frac{1}{\lambda} + (n - N) \log \frac{1}{1 - \lambda} + \log |\Theta|. \quad (28)$$

For large enough n , the additional loss $\log |\Theta|$ of the Bayes method over and above that of the omniscient scientist will be negligible. By averaging over y^n , we obtain the related upper bound of $h_\lambda n + \log |\Theta|$ for the risk of Bayes method when the outcome sequence is generated at random, either according to a fixed state of nature θ^* or a random θ^* .

3.4.2 Infinite Θ , noisy $\{\pm 1\}$ -valued functions

For infinite (possibly uncountable) Θ indexing a class of $\{\pm 1\}$ -valued functions with i.i.d. sign noise with rate λ , an only slightly more sophisticated method gives useful bounds in many cases. These are the cases when the class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ has finite *VC dimension*. This notion can be defined briefly as follows:

For any class \mathcal{F} of functions from $X \rightarrow \{\pm 1\}$ and sequence $S = (x_1, \dots, x_m)$ of points in X , we say that S is *shattered* by \mathcal{F} if for any sequence (y_1, \dots, y_m) of values in $\{\pm 1\}$, there is a function $f \in \mathcal{F}$ with $y_i = f(x_i)$ for all i , $1 \leq i \leq m$. The VC Dimension of \mathcal{F} , denoted $\mathbf{dim}(\mathcal{F})$, is defined as the length m of the longest such shattered sequence, over all possible finite sequences of points in X . Further discussion and examples of this concept can be found in [Vap82, BEHW89].

Now let us consider the case where $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ and x_1, \dots, x_n is a fixed instance sequence. Let us define the partition Π of Θ by letting

$$\Pi(\theta) = \Pi(\theta^*) \leftrightarrow f_\theta(x_t) = f_{\theta^*}^*(x_t), 1 \leq t \leq n.$$

By the most basic theorem connected with the VC dimension, known as the Sauer/VC lemma [VC71, Sau72], the number $|\Pi|$ of distinct equivalence classes in Π is bounded by

$$|\Pi| \leq \sum_{i=0}^d \binom{n}{i} \leq \left(\frac{en}{d}\right)^d, \quad (29)$$

where $d = \mathbf{dim}(\mathcal{F})$ and e is the base of the natural logarithm. (The latter inequality holds for $n \geq d$.) It follows that for any prior P on Θ ,

$$\mathcal{H}(\Pi) \leq d \log \frac{en}{d}, \quad (30)$$

since the entropy is maximized for the uniform distribution on Π , and takes the value $\log |\Pi|$ in this case. Furthermore, by our definition of Π it is clear that if $\Pi(\theta) = \Pi(\theta^*)$, then θ and θ^* induce the same conditional distribution on Y^n . Hence $I(P_{Y^n|\theta^*} \parallel P_{Y^n|\theta}) = 0$. It follows that $\mathcal{D}(\Pi) = 0$. Putting this all together, and applying (24), we get the following bound for the Bayes risk:

For all $n \geq d = \mathbf{dim}(\mathcal{F})$, all instance sequences x_1, \dots, x_n , and all priors P on Θ ,

$$R^\Delta(P) \leq \mathcal{D}(\Pi) + \mathcal{H}(\Pi) \leq d \log \frac{en}{d} = (1 + o(1))d \log n \quad (31)$$

and hence

$$R(P) \leq h_\lambda n + d \log \frac{en}{d}, \quad (32)$$

when the noise rate is λ , where $h_\lambda = -\lambda \log \lambda - (1 - \lambda) \log(1 - \lambda)$.

Using the other bounds given in the previous sections, upper bounds can also be obtained for the risk of Bayes method when the outcome sequence y^n is arbitrary, and when it is generated at random from a particular θ^* . However, the bounds in these cases depend on the probability $P(\Pi(\hat{\theta}))$ (resp. $P(\Pi(\theta^*))$) of the relevant equivalence class in the partition Π . If all equivalence classes have roughly the same prior probability, then the result is much the same as given above for the Bayes risk. Otherwise more careful analysis is required.

Sometimes we can also obtain interesting bounds even in the case that the VC dimension of \mathcal{F} is infinite. Here we also average over random choices of the instances $x^n = (x_1, \dots, x_n)$, and instead of using the VC dimension, we use the VC entropy [VC71] (see also [HKS91]). Let Q be a distribution on the instance space X , and assume that each x_t is chosen independently at random according to Q . For each x^n define the partition Π_{x^n} of Θ as above by letting

$$\Pi_{x^n}(\theta) = \Pi_{x^n}(\theta^*) \leftrightarrow f_\theta(x_t) = f_{\theta^*}^*(x_t), 1 \leq t \leq n.$$

The *VC entropy* of \mathcal{F} (for sample size n) is $\mathbf{E}_{Q^n}(\log |\Pi_{x^n}|)$. It is clear that the above derivation of (32) can be generalized to obtain

$$\mathbf{E}_{Q^n}(R(P)) \leq h_\lambda n + \mathbf{E}_{Q^n}(\log |\Pi_{x^n}|). \quad (33)$$

4 0-1 Loss

We now derive upper bounds on the risk of Bayes method for the 0-1 loss. As shown in (1), for this loss function Bayes method predicts the outcome with the highest posterior probability. Thus for outcome sequence $y^n = (y_1, \dots, y_n) \in Y^n$, where $Y = \{\pm 1\}$, and $1 \leq t \leq n$, the action a_t of Bayes method at time t is the prediction

$$a_t = \hat{y}_t = \operatorname{argmax}_{y \in \{\pm 1\}} P(y|y^{t-1}). \quad (34)$$

The total 0-1 loss for an outcome sequence $y^n \in \{\pm 1\}^n$ is the total number of times this prediction is incorrect, i.e.

$$L_{\text{bayes}, P, 0-1}^T(y^n) = |\{t : y_t \neq \hat{y}_t, 1 \leq t \leq n\}|. \quad (35)$$

This is often called the *number of mistakes*.

4.1 Performance on Arbitrary Outcome Sequences

Littlestone and Warmuth [LW89, Lit89] and Vovk [Vov90b] have obtained bounds on the number of mistakes made by Bayes method for arbitrary $\{\pm 1\}$ -valued outcome sequences for the case that Θ is a countable class $\{f_\theta : \theta \in \Theta\}$ of functions and the prior information assumes that outcomes are generated by applying sign noise with known rate λ (even though the actual outcome sequence is arbitrary). The bound from [LW89] is particularly easy to derive, and generalizes to continuous Θ easily as well. We give a variant of this derivation now.

First, let us define the *Heavyside* function Θ by letting $\Theta(x) = 1$ if $x \geq 0$ and $\Theta(x) = 0$ if $x < 0$. We use the fact that for any b and x , where $0 < b, x \leq 1$,

$$\Theta(b - x) \leq \frac{\log \frac{1}{x}}{\log \frac{1}{b}}, \quad (36)$$

which is easily verified (assuming $0/0 = 1$). We note also that since $0 \leq \lambda \leq 1/2$,

$$\frac{1}{2(1-\lambda)} \leq 1$$

and

$$\begin{aligned} P(y^t) &= \sum_{\theta \in \Theta} P(y^t|\theta)P(\theta) \\ &= \sum_{\theta \in \Theta} P(y_t|\theta)P(y^{t-1}|\theta)P(\theta) \\ &\leq \sum_{\theta \in \Theta} (1-\lambda)P(y^{t-1}|\theta)P(\theta) \\ &= (1-\lambda)P(y^{t-1}), \end{aligned}$$

thus

$$\frac{P(y^t)}{(1-\lambda)P(y^{t-1})} \leq 1.$$

(Of course the same result holds for continuous Θ .)

It is clear that Bayes method makes a mistake only when the posterior probability of the outcome y_t , given the previous outcomes y^{t-1} , is less than or equal to one half. Hence the total number of mistakes is bounded by

$$\begin{aligned} L_{\text{bayes},P,0-1}^T(y^n) &\leq |\{t : P(y_t|y^{t-1}) \leq 1/2, 1 \leq t \leq n\}| \\ &= |\{t : \frac{P(y^t)}{P(y^{t-1})} \leq 1/2, 1 \leq t \leq n\}| \\ &= \sum_{t=1}^n \Theta\left(\frac{1}{2} - \frac{P(y^t)}{P(y^{t-1})}\right) \\ &= \sum_{t=1}^n \Theta\left(\frac{1}{2(1-\lambda)} - \frac{P(y^t)}{(1-\lambda)P(y^{t-1})}\right) \\ &\leq \sum_{t=1}^n \frac{\log_2 \frac{(1-\lambda)P(y^{t-1})}{P(y^t)}}{\log_2(2(1-\lambda))} \\ &= \frac{n \log_2(1-\lambda) - \log_2 P(y^n)}{1 + \log_2(1-\lambda)} \\ &= \frac{n \log_2(1-\lambda) + L_{\text{bayes},P,\log_2}^T(y^n)}{1 + \log_2(1-\lambda)}, \end{aligned} \tag{37}$$

where the last equality follows from (6).

It is very useful to have a bound for the risk of Bayes method under the 0-1 loss in terms of the risk under the log loss, because this allows us to apply the results of the previous section. In particular, from (26), we have

$$\begin{aligned} L_{\text{bayes},P,\log}^T(y^n) &= L_{\text{OS},\hat{\theta},\log}^T(y^n) + L_{\text{bayes},\log}^\Delta(y^n) \\ &= N \log \frac{1}{\lambda} + (n - N) \log \frac{1}{1-\lambda} + L_{\text{bayes},P,\log}^\Delta(y^n), \end{aligned} \tag{38}$$

where $N = \min_{\theta \in \Theta} |\{t : y_t \neq f_\theta(x_t)\}|$ is the minimum number of noisy outcomes in any interpretation on y^n . Note that this is the same as the 0-1 loss of the omniscient scientist, i.e. $N = L_{\text{OS},P,0-1}^T(y^n)$. Hence, combining (37) and (38),

$$L_{\text{bayes},P,0-1}^T(y^n) \leq \frac{\left(\log_2 \frac{1-\lambda}{\lambda}\right) L_{\text{OS},P,0-1}^T(y^n) + L_{\text{bayes},P,\log_2}^\Delta(y^n)}{1 + \log_2(1-\lambda)} \tag{39}$$

In particular, if Θ is finite and the prior P is uniform on Θ then from (27) we have

$$L_{\text{bayes},P,0-1}^T(y^n) \leq \frac{\left(\log_2 \frac{1-\lambda}{\lambda}\right) L_{OS,P,0-1}^T(y^n) + \log_2 |\Theta|}{1 + \log_2(1 - \lambda)}. \quad (40)$$

Other bounds on $L_{\text{Bayes},P,\log}^\Delta(y^n)$ can be plugged in in other circumstances.

4.2 Performance for Random Sequences Assuming True State of Nature

Using the same prior as in the previous section, we now assume that the outcome sequence is actually generated at random according to the distribution P_{θ^*} for some true state of nature θ^* . In the case of $\{\pm 1\}$ -valued functions with independent sign noise, which we examine here, this means that $y_t = f_{\theta^*}(x_t)$ with probability $1 - \lambda$ and $y_t = -f_{\theta^*}(x_t)$ with probability λ . The risk of Bayes method under 0-1 loss for this case is the average total number of mistakes made by Bayes method.

We can obtain bounds on the average total number of mistakes by averaging the formulae from the previous section (e.g. (39)), using the fact that $\min_{\theta \in \Theta} |\{t : y_t \neq f_\theta(x_t)\}| \leq |\{t : y_t \neq f_{\theta^*}(x_t)\}|$, and hence the average performance of the omniscient scientist who uses knowledge of the MLE for each individual outcome sequence is at least as good as the risk (= average performance) of the omniscient scientist that uses knowledge of θ^* . Since the expectation of $|\{t : y_t \neq f_{\theta^*}(x_t)\}|$ is λn , from (39), this gives

$$r_{\text{bayes},P,0-1}(\theta^*) \leq \frac{\log_2 \frac{1-\lambda}{\lambda}}{1 + \log_2(1 - \lambda)} \lambda n + \frac{r_{\text{Bayes},P,\log_2}^\Delta(\theta^*)}{1 + \log_2(1 - \lambda)}.$$

Unfortunately, this bound is not tight. We show how to get rid of the extra $\frac{\log_2 \frac{1-\lambda}{\lambda}}{1 + \log_2(1 - \lambda)}$ factor in front of the λn . This is especially important, since in most cases the remaining term is $O(\log n)$.

The main idea is to bound the difference in 0-1 loss at time t between the Bayes method and the omniscient scientist in terms of the difference at time t of the corresponding log loss. Fix the outcome sequence y^n . The loss of Bayes method at time t is $-\log P(y_t|y^{t-1})$, and the loss at time t of the omniscient scientist who knows the true state of nature θ^* is $-\log P(y_t|y^{t-1}, \theta^*) = -\log P(y_t|\theta^*)$, since Y_t is conditionally independent of Y_1, \dots, Y_{t-1} given θ^* in the case of $\{\pm 1\}$ -functions with independent sign noise. Thus the difference in these losses is

$$L_{t,\log}^\Delta(y^n) \doteq \log \frac{P(y_t|\theta^*)}{P(y_t|y^{t-1})}. \quad (41)$$

Let $K_\lambda(x) = \lambda \log \frac{\lambda}{x} + (1 - \lambda) \log \frac{1-\lambda}{1-x}$ for $\lambda \leq x \leq 1 - \lambda$. We can rewrite the difference between the risk of Bayes method and the risk of the omniscient scientist under the log loss

as follows.

$$\begin{aligned}
r_{\log}^{\Delta}(\theta^*) &= \sum_{y^n \in Y^n} P(y^n | \theta^*) \sum_{t=1}^n L_{t,\log}^{\Delta}(y^n) \\
&= \sum_{y^n \in Y^n} P(y^n | \theta^*) \sum_{t=1}^n \log \frac{P(y_t | \theta^*)}{P(y_t | y^{t-1})} \\
&= \sum_{y^n \in Y^n} P(y^n | \theta^*) \sum_{t=1}^n \sum_{y \in Y} P(Y_t = y | \theta^*) \log \frac{P(Y_t = y | \theta^*)}{P(Y_t = y | y^{t-1})} \tag{42}
\end{aligned}$$

$$= \sum_{y^n \in Y^n} P(y^n | \theta^*) \sum_{t=1}^n K_{\lambda}(p_t), \tag{43}$$

where $p_t = P(Y_t = -f_{\theta^*}(x_t) | y^{t-1})$. Equation (42) looks a bit strange at first, but all that is really being introduced here is superfluous averaging over possibilities that we are already averaging over in the outermost sum. Thus the overall expectation is unchanged. This method is analogous to that used in [HKS91]. The last equation, (43), follows directly from the definition of the noise model. It is easily verified that we always have $\lambda \leq p_t \leq 1 - \lambda$.

We now derive analogous equations for the 0-1 loss. Let h be the modified Heavyside function defined by $h(x) = 1$ if $x > 0$, $h(x) = 0$ if $x < 0$ and $h(0) = 1/2$. We assume that for the 0-1 loss, Bayes method tosses a fair coin to make its prediction when the posterior probabilities of -1 and $+1$ are each $1/2$. This will make our analysis easier, and does not affect previous results.

Under these assumptions, for fixed y^n the 0-1 loss of Bayes method at time t (averaging over coin tosses when necessary) is $h(\frac{1}{2} - P(y_t | y^{t-1}))$. The 0-1 loss of the omniscient scientist at time t is $h(\frac{1}{2} - P(y_t | \theta^*))$. Thus the difference in these losses is

$$L_{t,0-1}^{\Delta}(y^n) \doteq h(\frac{1}{2} - P(y_t | y^{t-1})) - h(\frac{1}{2} - P(y_t | \theta^*)). \tag{44}$$

In analogy with the function K_{λ} , let us define $J_{\lambda}(x) = (1 - 2\lambda)h(x - \frac{1}{2})$. We can then express the difference between the risk of Bayes method and the risk of the omniscient scientist under the 0-1 loss as follows.

$$\begin{aligned}
r_{0-1}^{\Delta}(\theta^*) &= \sum_{y^n \in Y^n} P(y^n | \theta^*) \sum_{t=1}^n L_{t,0-1}^{\Delta}(y^n) \\
&= \sum_{y^n \in Y^n} P(y^n | \theta^*) \sum_{t=1}^n [h(\frac{1}{2} - P(y_t | y^{t-1})) - h(\frac{1}{2} - P(y_t | \theta^*))] \\
&= \sum_{y^n \in Y^n} P(y^n | \theta^*) \sum_{t=1}^n \sum_{y \in Y} P(Y_t = y | \theta^*) [h(\frac{1}{2} - P(Y_t = y | y^{t-1})) - h(\frac{1}{2} - P(Y_t = y | \theta^*))] \\
&= \sum_{y^n \in Y^n} P(y^n | \theta^*) \sum_{t=1}^n [\lambda(h(\frac{1}{2} - p_t) - h(\frac{1}{2} - \lambda)) + (1 - \lambda)(h(p_t - \frac{1}{2}) - h(\lambda - \frac{1}{2}))]
\end{aligned}$$

$$\leq \sum_{y^n \in Y^n} P(y^n | \theta^*) \sum_{t=1}^n J_\lambda(p_t), \tag{45}$$

where $p_t = P(Y_t = -f_{\theta^*}(x_t) | y^{t-1})$ as above. The last inequality is readily verified by case analysis.

Note that K_λ increases monotonically as x goes from λ to $1 - \lambda$, and $J_\lambda(x)$ is a step function that is 0 for $x < 1/2$ and $1 - 2\lambda$ for $x > 1/2$. Hence it is clear that

$$J_\lambda(x) \leq \frac{1 - 2\lambda}{K_\lambda(1/2)} K_\lambda(x) = \frac{C_\lambda}{1 - 2\lambda} K_\lambda(x) \tag{46}$$

for all $\lambda \leq x \leq 1 - \lambda$, where

$$C_\lambda = \frac{(1 - 2\lambda)^2}{\lambda \log(2\lambda) + (1 - \lambda) \log(2(1 - \lambda))}.$$

It can be shown that if \log is the natural log, then $\frac{1}{\log 2} \leq C_\lambda \leq 2$ for $0 \leq \lambda \leq 1/2$. The lower bound is tight as $\lambda \rightarrow 0$, and the upper bound is tight as $\lambda \rightarrow 1/2$.

Hence it follows from (43), (45) and (46) that

$$r_{0-1}^\Delta(\theta^*) \leq \frac{C_\lambda}{1 - 2\lambda} r_{ln}^\Delta(\theta^*), \tag{47}$$

where $\frac{1}{\ln 2} \leq C_\lambda \leq 2$.

4.3 Bayes Risk For 0-1 Loss

In this final section we look at the 0-1 Bayes risk, that is the risk of Bayes procedure for 0-1 loss under the assumption that the outcome sequence y^n is generated by choosing θ^* according to the prior distribution and then generating y^n according to P_{θ^*} . As above, we compare the risk of the Bayes method to the risk of an omniscient scientist who knows θ^* . Again, we can simply average the results of the previous section to obtain bounds on the excess risk of the Bayes method. However, a direct analysis yields a bound that is better by a factor of 2. As above we consider only the case of $\{\pm 1\}$ -valued functions times independent sign noise with rate λ . Our analysis is similar to that given in [HKS91] for the noise-free case, and generalizes the corresponding results given there. It also parallels the analysis of the previous section.

The Bayes risk for the log loss is

$$R_{log}(P) = \sum_{y^n \in Y^n} P(y^n) \sum_{t=1}^n \log \frac{1}{P(y_t | y^{t-1})}$$

$$\begin{aligned}
&= \sum_{y^n \in Y^n} P(y^n) \sum_{t=1}^n \sum_{y \in Y} P(Y_t = y | y^{t-1}) \log \frac{1}{P(Y_t = y | y^{t-1})} \\
&= \sum_{y^n \in Y^n} P(y^n) \sum_{t=1}^n \mathcal{H}(Y_t | y^{t-1})
\end{aligned} \tag{48}$$

The corresponding risk for the omniscient scientist under the log loss is $h_\lambda n$, where $h_\lambda = \lambda \log \frac{1}{\lambda} + (1 - \lambda) \log \frac{1}{1 - \lambda}$, since the omniscient scientist knows everything except which values are changed by the noise. Hence

$$R_{\log}^\Delta(P) = \sum_{y^n \in Y^n} P(y^n) \sum_{t=1}^n [\mathcal{H}(Y_t | y^{t-1}) - h_\lambda] \tag{49}$$

Now turning to the 0-1 loss, the Bayes risk is

$$\begin{aligned}
R_{0-1}(P) &= \sum_{y^n \in Y^n} P(y^n) \sum_{t=1}^n h\left(\frac{1}{2} - P(y_t | y^{t-1})\right) \\
&= \sum_{y^n \in Y^n} P(y^n) \sum_{t=1}^n \sum_{y \in Y} P(Y_t = y | y^{t-1}) h\left(\frac{1}{2} - P(Y_t = y | y^{t-1})\right) \\
&= \sum_{y^n \in Y^n} P(y^n) \sum_{t=1}^n \tilde{\mathcal{H}}(Y_t | y^{t-1}),
\end{aligned} \tag{50}$$

where $\tilde{\mathcal{H}}(Y_t | y^{t-1}) = \min(P(Y_t = +1 | y^{t-1}), P(Y_t = -1 | y^{t-1}))$.

The corresponding risk for the omniscient scientist under the 0-1 loss is λn . Hence

$$R_{0-1}^\Delta(P) = \sum_{y^n \in Y^n} P(y^n) \sum_{t=1}^n [\tilde{\mathcal{H}}(Y_t | y^{t-1}) - \lambda] \tag{51}$$

Both $\mathcal{H}(Y_t | y^{t-1}) - h_\lambda$ and $\tilde{\mathcal{H}}(Y_t | y^{t-1}) - \lambda$ are functions of $x = P(Y_t = +1 | y^{t-1})$ that are symmetric about $x = 1/2$, equal 0 for $x = \lambda$ and $x = 1 - \lambda$, increasing as x goes from λ to $1/2$ and (by symmetry) decreasing as x goes from $1/2$ to $1 - \lambda$. In fact, $\tilde{\mathcal{H}}(x) - \lambda$ is linear for $\lambda \leq x \leq 1/2$ and $\mathcal{H}(x) - h_\lambda$ is concave. It follows that $\tilde{\mathcal{H}}(x) - \lambda \leq \frac{\tilde{\mathcal{H}}(1/2) - \lambda}{\mathcal{H}(1/2) - h_\lambda} (\mathcal{H}(x) - h_\lambda)$ for all $\lambda \leq x \leq 1 - \lambda$, using symmetry again to verify the inequality for the second half of the range. Furthermore,

$$\frac{\tilde{\mathcal{H}}(1/2) - \lambda}{\mathcal{H}(1/2) - h_\lambda} = \frac{\frac{1}{2}(1 - 2\lambda)}{\lambda \log(2\lambda) + (1 - \lambda) \log(2(1 - \lambda))} = \frac{C_\lambda}{2(1 - 2\lambda)},$$

where C_λ is as defined in the previous section, and we assume natural logs. Since x always lies between λ and $1 - \lambda$ when $x = P(Y_t = +1 | y^{t-1})$, it follows from (49) and (51) that

$$R_{0-1}^\Delta(P) \leq \frac{C_\lambda}{2(1 - 2\lambda)} R_{\log}^\Delta(P), \tag{52}$$

where $\frac{1}{\ln 2} \leq C_\lambda \leq 2$.

For $\lambda = 0$, $C_\lambda = \frac{1}{\ln 2}$ and from (52) we get $R_{0-1}^\Delta(P) \leq \frac{1}{2\ln 2} R_{l_n}^\Delta(P) = \frac{1}{2} R_{\log_2}^\Delta(P)$, which was the result from [HKS91]. As λ approaches $1/2$, C_λ approaches 2, and we get $R_{0-1}^\Delta(P) \leq \frac{1}{1-2\lambda} R_{l_n}^\Delta(P)$. Combining this with (31) from section 3.4.2, if the class of functions has VC dimension d then this gives

$$R_{0-1}^\Delta(P) \leq \frac{d}{1-2\lambda} \ln \frac{en}{d}.$$

Hence

$$R_{0-1}(P) \leq \lambda n + \frac{d}{1-2\lambda} \ln \frac{en}{d}. \quad (53)$$

This holds for any class \mathcal{F} of VC dimension d , any sample size $n \geq d$, and any prior P .

5 Conclusion

We have derived a number of upper bounds on the risk of Bayes method for sequential classification and regression problems on the outcome space $Y = \{\pm 1\}$. Many of our techniques should generalize easily to other kinds of outcome spaces, as well as other decision spaces and loss functions. Some of these techniques may also help in analyzing other learning methods, such as the ‘‘Gibbs’’ method [GT90, HKS91, OH91b, OH91a, SST92]. However, a major problem remaining is to develop equally simple and general techniques to obtain lower bounds on the risk, so that we can see how tight these upper bounds are.

Very tight upper and lower bounds on the risk of Bayes methods under log loss are available for the case when Θ is a compact subset of R^d and the relative entropy $I(P_\theta \parallel P_{\theta^*})$ is twice continuously differentiable at $\theta = \theta^*$ for almost all $\theta^* \in \Theta$, so that the Fisher information is well-defined [Sch78, Efr79, Ris86, CB90, Dawa, Yam91, Yam92]. This is very often not the case for a discrete outcome space such as $\{\pm 1\}$, so we have concentrated on more general bounds here, which involve much weaker assumptions, such as finiteness of the VC dimension or bounds on the VC entropy. It remains to see exactly how the results given here relate to the more standard statistical approaches using Fisher information.

It should also be noted that whereas the bounds obtained for the log loss are quite general, those we have obtained for the 0-1 loss are restricted to the case of noisy functions with known noise rate. It would be nice to have more general bounds for 0-1 loss. There also many cases of interest in which the class of functions indexed by Θ can be decomposed into classes $\mathcal{F}_1, \mathcal{F}_2, \dots$ of increasing VC dimension [Vap82], including the case where Θ consists of a sequence of smooth real parameterizations as above of increasing dimension [Yam92]. Some general results on the performance of Bayes methods in this case for noise-free outcomes are reported in [HKS]. These results should be (and can be) extended to the noisy case.

Finally, it would be nice to have good bounds on the average loss for the n th example, in addition to bounds on the average total loss for the first n examples. Yamanishi gives bounds of this type for some special cases [Yam91, Yam92] (see also [Ama92, AFS92]). The former quantity is what is usually displayed as a ‘‘learning curve’’. Of course, if we had matching

upper and lower bounds on the average total loss for the first n examples for each n , then we could obtain a learning curve by simply subtracting the average loss on the first $n - 1$ examples from the average loss on the first n . When the average loss grows logarithmically, as it appears to in many of the cases we study here (after subtracting off the loss of the omniscient scientist), and as it provably does in some other cases, then this gives a learning curve of the form c/n for some constant c that we can estimate. However, this subtraction is not valid without extremely tight upper and lower bounds on the average total loss. Thus our current results are merely suggestive regarding the shape of the learning curve for Bayes method.

Acknowledgements

The first author would like to acknowledge the many insights gained from numerous discussions with Michael Kearns and Manfred Opper about the performance of Bayes methods, especially discussions with Michael Kearns about the noisy functions case, in which some interesting preliminary results were obtained, and discussions with Manfred Opper about the critical role played by the scaling behavior of the prior probability of a neighborhood around θ^* as a function of its radius. Discussions with Nick Littlestone about concepts related to that of the omniscient scientist were also very helpful. In many ways these discussions, along with our previous collaborative work, layed the foundation for this paper. He would also like to thank David Helmbold, Phil Long and Anselm Blumer for helpful conversations/comments. Both authors would like to acknowledge helpful discussions with Bin Yu, and, indirectly through Bin, with L. LeCam.

In addition, the first author gratefully acknowledges the support of ONR grant N00014-91-J-1162 and NSF grants CDA-9115268 and IRI-9123692. Both authors acknowledge the support of the Mathematical Sciences Research Institute at Berkeley where much of this work was done, and NSF grant NSF_DMS 8505550, which supported the work at MSRI.

References

- [AFS92] S. Amari, N. Fujita, and S. Shinomoto. Four types of learning curves. *Neural Computation*, 4:605–619, 1992.
- [Ama92] S. Amari. A universal theorem on learning curves. Unpublished manuscript, 1992.
- [Bar87] Andrew 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.

- [BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.
- [CB90] Bertrand Clarke and Andrew Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.
- [CT91] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.
- [Dawa] A.P. Dawid. Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian Statistics 4*. To appear.
- [Dawb] A.P. Dawid. Prequential data analysis. *Current Issues in Statistical Inference*. To appear.
- [Daw84] A.P. Dawid. Statistical theory: The prequential approach. *J. Roy. Statist. Soc. A*, pages 278–292, 1984.
- [DMW88] Alfredo DeSantis, George Markowski, and Mark N. Wegman. Learning probabilistic prediction functions. In *Proceedings of the 1988 Workshop on Computational Learning Theory*, pages 312–328. Morgan Kaufmann, 1988.
- [Efr79] S. Yu. Efroimovich. Information contained in a sequence of observations. *Problems in Information Transmission*, 15:178–189, 1979.
- [FM92] Meir Feder and Neri Merhav. Relations between entropy and error probability. unpublished manuscript, 1992.
- [FMG92] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Trans. Info. Th.*, 38:1258–1270, 1992.
- [GT90] G. Gyorgyi and N. Tishby. Statistical theory of learning a rule. In K. Thueemann and R. Koeberle, editors, *Neural Networks and Spin Glasses*. World Scientific, 1990.
- [HKLW91] David Haussler, Michael Kearns, Nick Littlestone, and Manfred K. Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*, 95:129–161, 1991.
- [HKS] D. Haussler, M. Kearns, and R. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning*. to appear.
- [HKS91] D. Haussler, M. Kearns, and R. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. In *Proceedings of the Fourth Workshop on Computational Learning Theory*, pages 61–74, 1991.

- [HLW90] David Haussler, Nick Littlestone, and Manfred Warmuth. Predicting $\{0, 1\}$ -functions on randomly drawn points. Technical Report UCSC-CRL-90-54, University of California Santa Cruz, Computer Research Laboratory, December 1990. To appear in *Information and Computation*.
- [Kul59] Solomon Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.
- [Lit89] Nick Littlestone. *Mistake Bounds and Logarithmic Linear-threshold Learning Algorithms*. PhD thesis, University of California Santa Cruz, 1989.
- [LW89] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. In *30th Annual IEEE Symposium on Foundations of Computer Science*, pages 256–261, 1989.
- [MF92] Neri Merhav and Meir Feder. Universal sequential learning and decision from individual data sequences. *The 1992 Workshop on Computational Learning Theory*, 1992.
- [OH91a] 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.
- [OH91b] M. Opper and D. Haussler. Generalization performance of Bayes optimal classification algorithm for learning a perceptron. *Physical Review Letters*, 66(20):2677–2680, May 1991.
- [Ris86] Jorma Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100, 1986.
- [Sau72] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (Series A)*, 13:145–147, 1972.
- [Sch78] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [SST92] H.S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, April 15, 1992.
- [Vap82] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.
- [VC71] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–80, 1971.

- [Vov90a] V. G. Vovk. Prequential probability theory. unpublished manuscript, 1990.
- [Vov90b] Volodimir Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383. Morgan Kaufmann, 1990.
- [Vov92] V. G. Vovk. Universal forecasting algorithms. *Information and Computation*, 96(2):245–277, Feb. 1992.
- [Yam91] Kenji Yamanishi. A loss bound model for on-line stochastic prediction strategies. In L. G. Valiant and M. Warmuth, editors, *Proceedings of the 1991 Workshop on Computational Learning Theory*, pages 290–302, San Mateo, CA, 1991. Morgan Kaufmann.
- [Yam92] Kenji Yamanishi. *A Statistical Approach to Computational Learning Theory*. PhD thesis, University of Tokyo, March 1992.
- [Yu] Bin Yu. On optimal rate universal d-semifair coding. *IEEE Transactions on Information Theory*. To appear.