Suboptimal Behavior of Bayes and MDL in Classification under Misspecification

Peter Grünwald[1] and John Langford[2]

[1] CWI Amsterdam, pdg@cwi.nl, www.grunwald.nl.
[2] TTI-Chicago, jcl@cs.cmu.edu, hunch.net/~jl/.

Abstract. We show that forms of Bayesian and MDL inference that are often applied to classification problems can be *inconsistent*. This means there exists a learning problem such that for all amounts of data the generalization errors of the MDL classifier and the Bayes classifier relative to the Bayesian posterior both remain bounded away from the smallest achievable generalization error.

1 Introduction

Overfitting is a central concern of machine learning and statistics. Two frequently used learning methods that in many cases 'automatically' protect against overfitting are Bayesian inference [5] and the Minimum Description Length (MDL) Principle [21, 2, 11]. We show that, when applied to classification problems, some of the standard variations of these two methods can be *inconsistent* in the sense that they *asymptotically overfit*: there exist scenarios where, no matter how much data is available, the generalization error of a classifier based on MDL or the full Bayesian posterior does not converge to the minimum achievable generalization error within the set of classifiers under consideration.

Some Caveats and Warnings These result must be interpreted carefully. There exist many different versions of MDL and Bayesian inference, only some of which are covered. For the case of MDL, we show our result for a two-part form of MDL that has often been used for classification, see Section **??**. For the case of Bayes, our result may appear to contradict some well-known Bayesian consistency results [6]. Indeed, our result only applies to a 'pragmatic' use of Bayes, where the set of hypotheses under consideration are *classifiers*: functions mapping each input X to a discrete class label Y. To apply Bayes rule, these classifiers must be converted into conditional probability distributions. We do this conversion in a standard manner, crossing a prior on classifiers with a prior on error rates for these classifiers. This may lead to (sometimes subtly) 'misspecified' probability models not containing the 'true' distribution D. Thus, our result may be restated as 'Bayesian methods for classification can be inconsistent under misspecification for common classification probability models'. The result is still interesting, since (1) even under misspecification, Bayesian inference is known to be consistent under fairly broad conditions – we provide an explicit context in which it is not; (2) in practice, Bayesian inference is used frequently for classification under misspecification – see Section 6.

1.1 A Preview

Classification Problems A classification problem is defined on an input (or feature) domain $\mathcal{X}$ and output domain (or class label) $\mathcal{Y} = \{0, 1\}$. The problem is defined by a probability distribution D over $\mathcal{X} \times \mathcal{Y}$. A classifier is a function $c : \mathcal{X} \to \mathcal{Y}$ The error rate of any classifier is quantified as:

$$e_D(c) = E_{(x,y)\sim D} I(c(x) \neq y)$$

where $(x, y) \sim D$ denotes a draw from the distribution D and $I(\cdot)$ is the indicator function which is 1 when its argument is true and 0 otherwise.

The goal is to find a classifier which, as often as possible according to D, correctly predicts the class label given the input feature. Typically, the classification problem is solved by searching for some classifier c in a limited subset $\mathcal{C}$ of all classifiers using a sample $S = (x_1, y_1), \ldots, (x_m, y_m) \sim D^m$ generated by m independent draws from the distribution D. Naturally, this search is guided by the *empirical error rate*. This is the error rate on the subset S defined by:

$$\hat{e}_S(c) := E_{(x,y)\sim S} I(c(x) \neq y) = \frac{1}{|S|} \sum_{(x,y)\in S} I(c(x) \neq y).$$

where $(x, y) \sim S$ denotes a sample drawn from the uniform distribution on S. Note that $\hat{e}_S(c)$ is a random variable dependent on a draw from D^m. In contrast, $e_D(c)$ is a number (an expectation) relative to D.

The Basic Result Our basic result is that certain classifier learning algorithms may not behave well as a function of the information they use, even when given infinitely many samples to learn from. The learning algorithms we analyze are "Bayesian classification" (Bayes), "Maximum a Posteriori classification" (MAP), and "Minimum Description Length classification" (MDL). These algorithms are precisely defined later. Functionally they take as arguments a training sample S and a "prior" P which is a probability distribution over a set of classifiers $\mathcal{C}$. In Section 3 we state our basic result, Theorem 2. The theorem has the following corollary, indicating suboptimal behavior of Bayes and MDL:

Corollary 1. (Classification Inconsistency) *There exists an input domain $\mathcal{X}$, a prior P always nonzero on a countable set of classifiers $\mathcal{C}$, a learning problem D, and a constant $K > 0$ such that the Bayesian classifier $c_{\mathrm{BAYES}(P,S)}$, the MAP classifier $c_{\mathrm{MAP}(P,S)}$, and the MDL classifier $c_{\mathrm{MDL}(P,S)}$ are asymptotically K-suboptimal. That is, for each $\mathbf{e} \in \{e_D(c_{\mathrm{BAYES}(P,S)}), e_D(c_{\mathrm{MAP}(P,S)}), e_D(c_{\mathrm{MDL}(P,S)})\}$, we have*

$$\lim_{m\to\infty} \Pr_{S\sim D^m} \left(\mathbf{e} > K + \inf_{c\in\mathcal{C}} e_D(c) \right) = 1.$$

How dramatic is this result? We may ask (1) are the priors P for which the result holds natural; (2) how large can the constant K become and how small can $\inf_{c\in\mathcal{C}} e_D(c)$ be? (3) perhaps demanding an algorithm which depends on the prior P and the sample S to be consistent (asymptotically optimal) is too strong? The short answer to (1) and (2) is: the priors P have to satisfy several requirements, but they correspond to priors often used in practice. K can be quite large and $\inf_c e_D(c)$ can be quite small - see Section 5.1 and Figure 1.

The answer to (3) is that there do exist simple algorithms which are consistent. An example is the algorithm which minimizes the Occam's Razor bound (ORB) [7], Section 4.2.

Theorem 1. (ORB consistency) *For all priors P nonzero on a set of classifiers $\mathcal{C}$, for all learning problems D, and all constants $K > 0$ the ORB classifier $c_{\mathrm{ORB}(P,S)}$ is asymptotically K-optimal:*

$$\lim_{m\to\infty} \Pr_{S\sim D^m}\left(e_D(c_{\mathrm{ORB}(P,S)}) > K + \inf_{c\in\mathcal{C}} e_D(c)\right) = 0.$$

The remainder of this paper first defines precisely what we mean by the above classifiers. It then states the main inconsistency theorem which implies the above corollary, as well as a theorem that provides an upper-bound on how badly Bayes can behave. In Section 4 we prove our theorems. Variations of the result are discussed in Section 5.1. A discussion of the result from a Bayesian point of view is given in Section 6, and from an MDL point of view in Section **??**.

2 Some Classification Algorithms

The basic inconsistency result is about particular classifier learning algorithms which we define next.

The Bayesian Classification algorithm The Bayesian approach to inference starts with a prior probability distribution P over a set of distributions $\mathcal{P}$ which typically represents a measure of "belief" that some $p \in \mathcal{P}$ is the process generating data. Bayes' rule states that, given sample data S, the posterior probability $P(\cdot \mid S)$ that some p is the process generating the data is:

$$P(p \mid S) = \frac{p(S)P(p)}{P(S)}.$$

where $P(S) := E_{p\sim P}\, p(S)$. In classification problems with sample size $m = |S|$, each $p \in \mathcal{P}$ is a distribution on $(\mathcal{X}\times\mathcal{Y})^m$ and the outcome $S = (x_1, y_1), \ldots, (x_m, y_m)$ is the sequence of labeled examples.

If we intend to perform classification based on a set of classifiers $\mathcal{C}$ rather than distributions $\mathcal{P}$, it is natural to introduce a "prior" $P(c)$ that a particular classifier $c : \mathcal{X} \to \{0, 1\}$ is the best classifier for solving some learning problem. This, of course, is *not* a Bayesian prior in the conventional sense because classifiers do not induce a measure over the training data. It is the standard method of converting a "prior" over classifiers into a Bayesian prior over distributions on the observations which our inconsistency result applies to.

One common conversion [14, 22, 12] transforms the set of classifiers $\mathcal{C}$ into a simple logistic regression model – the precise relationship to logistic regression is discussed in Section 5.2. In our case $c(x) \in \{0, 1\}$ is binary valued, and then (but only then) the conversion amounts to assuming that the error rate θ of the optimal classifier is independent of the feature value x. This is known as "homoskedasticity" in statistics and "label noise" in learning theory. More precisely, it is assumed that, for the optimal classifier $c \in C$, there exists some θ such that $\forall x \; P(c(x) \neq y) = \theta$. Given this assumption, we can construct a conditional probability distribution $p_{c,\theta}$ over the labels given the unlabeled data:

$$p_{c,\theta}(y^m \mid x^m) = \theta^{m\hat{e}_S(c)}(1-\theta)^{m-m\hat{e}_S(c)}. \tag{1}$$

For each fixed $\theta < 0.5$, the log likelihood $\log p_{c,\theta}(y^m \mid x^m)$ is linearly decreasing in the empirical error that c makes on S. By differentiating with respect to θ, we see that for fixed c, the likelihood (1) is maximized by setting $\theta := \hat{e}_S(c)$, giving

$$\log \frac{1}{p_{c,\hat{e}_S(c)}(y^m \mid x^m)} = mH(\hat{e}_S(c)). \tag{2}$$

where H is the binary entropy $H(\mu) = -\mu \log \mu - (1-\mu)\log(1-\mu)$, which is strictly increasing for $\hat{e}_S(c) \in [0, 0.5)$. We further assume that *some* distribution p_x on $\mathcal{X}^m$ generates the x-values[3]. We can apply Bayes rule to get a posterior on $p_{c,\theta}$, denoted as $P(c, \theta \mid S)$, without knowing p_x, since the $p_x(x^m)$-factors cancel:

$$P(c, \theta \mid S) = \frac{p_{c,\theta}(y^m|x^m)p_x(x^m)P(c,\theta)}{P(y^m \mid x^m)p_x(x^m)} = \frac{p_{c,\theta}(y^m|x^m)P(c,\theta)}{E_{c,\theta \sim P}\, p_{c,\theta}(y^m \mid x^m)}. \tag{3}$$

To make (3) applicable, we need to incorporate a prior measure on the joint space $\mathcal{C} \times [0,1]$ of classifiers and θ-parameters. In the next section we discuss the priors under which our theorems hold.

Bayes rule (3) is formed into a classifier learning algorithm by choosing the most likely label given the input x and the posterior $P(\cdot|S)$:

$$c_{\text{BAYES}(P,S)}(x) := \begin{cases} 1 & \text{if } E_{c,\theta \sim P(\cdot|S)}\, p_{c,\theta}(Y = 1|X = x) \geq \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The MAP classification algorithm The integrations of the full Bayesian classifier can be too computationally intensive, so we sometimes predict using the Bayesian Maximum A Posteriori (MAP) classifier. This classifier is given by:

$$c_{\text{MAP}(P,S)} = \arg\max_{c \in \mathcal{C}} \max_{\theta \in [0,1]} P(c, \theta \mid S) = \arg\max_{c \in \mathcal{C}} \max_{\theta \in [0,1]} p_{c,\theta}(y^m \mid x^m) P(c, \theta)$$

with ties broken arbitrarily. Integration over $\theta \in [0,1]$ being much less problematic than summation over $c \in \mathcal{C}$, one sometimes uses a learning algorithm which integrates over θ (like full Bayes) but maximizes over c (like MAP):

$$c_{\text{SMAP}(P,S)} = \arg\max_{c \in \mathcal{C}} P(c \mid S) = \arg\max_{c \in \mathcal{C}} E_{\theta \sim P(\theta)}\, p_{c,\theta}(y^m \mid x^m) P(c \mid \theta).$$

The MDL Classification algorithm The MDL approach to classification is transplanted from the MDL approach to density estimation. There is no such thing as a 'definition' of MDL for classification because the transplant has been performed in various ways by various authors. Nonetheless, as we discuss in Section **??**, most implementations are essentially equivalent to the following algorithm [20, 21, 15, 12]:

$$c_{\text{MDL}(P,S)} = \arg\min_{c \in \mathcal{C}} \log \frac{1}{P(c)} + \log \binom{m}{m\hat{e}_S(c)}. \tag{5}$$

The quantity minimized has a coding interpretation: it is the number of bits required to describe the classifier plus the number of bits required to describe the labels on S given the classifier and the unlabeled data. We call $-\log P(c) + \log \binom{m}{m\hat{e}_S(c)}$ the *two-part MDL codelength* for encoding data S with classifier c.

[3] And, in particular that this distribution is independent of c and θ.

3 Main Theorems

In this section we prove the basic inconsistency theorem. We prove inconsistency for some countable set of classifiers $\mathcal{C} = \{c_0, c_1, \ldots\}$ which we define later. The inconsistency is attained for priors with 'heavy tails', satisfying

$$\log \frac{1}{P(c_k)} \leq \log k + o(\log k). \tag{6}$$

This condition is satisfied, by, for example, Rissanen's *universal prior for the integers*, [21]. The sensitivity of our result to the choice of prior is analyzed further in Section 5.1. The prior on θ can be any distribution on $[0,1]$ with a continuously differentiable density P bounded away from 0, i.e. for some $\gamma > 0$,

$$\text{for all } \theta \in [0,1], P(\theta) > \gamma. \tag{7}$$

For example, we may take the uniform distribution with $P(\theta) \equiv 1$. We assume that the priors $P(\theta)$ on $[0,1]$ and the prior $P(c)$ on $\mathcal{C}$ are independent, so that $P(c,\theta) = P(c)P(\theta)$. In the theorem, $H(\mu) = -\mu \log \mu - (1-\mu)\log(1-\mu)$ stands for the binary entropy of a coin with bias μ.

Theorem 2. (Classification Inconsistency) *There exists an input space $\mathcal{X}$ and a countable set of classifiers $\mathcal{C}$ such that the following holds: let P be any prior satisfying (6) and (7). For all $\mu \in (0, 0.5)$ and all $\mu' \in [\mu, H(\mu)/2)$, there exists a D with $\min_{c\in\mathcal{C}} e_D(c) = \mu$ such that, for all large m, all $\delta > 0$,*

$$\begin{aligned}
\Pr_{S\sim D^m}\left(e_D(c_{\text{MAP}(P,S)}) = \mu'\right) &\geq 1 - a_m \\
\Pr_{S\sim D^m}\left(e_D(c_{\text{SMAP}(P,S)}) = \mu'\right) &\geq 1 - a_m \\
\Pr_{S\sim D^m}\left(e_D(c_{\text{MDL}(P,S)}) = \mu'\right) &\geq 1 - a_m, \\
\Pr_{S\sim D^m}\left(e_D(c_{\text{BAYES}(P,S)}) \geq \mu' - \delta\right) &\geq 1 - a_m, \qquad \textit{where } a_m = 3\exp(-2\sqrt{m}).
\end{aligned}$$

The theorem states that Bayes is inconsistent for *all large m* on a fixed distribution D. This is a significantly more difficult statement than "for all (large) m, there exists a learning problem where Bayes is inconsistent"[4]. Differentiation of $0.5H(\mu) - \mu$ shows that the maximum discrepancy between $e_D(c_{\text{MAP}(P,S)})$ and μ is achieved for $\mu = 1/5$. With this choice of μ, $0.5H(\mu) - \mu = 0.1609\ldots$ so that, by choosing μ' arbitrarily close to $H(\mu)$, the discrepancy $\mu' - \mu$ comes arbitrarily close to $0.1609\ldots$. These findings are summarized in Figure 1.

How large can the discrepancy between $\mu = \inf_c e_D(c)$ and $\mu' = e_D(c_{\text{BAYES}(P,S)})$ be in the large m limit, for general learning problems? Our next theorem, again summarized in Figure 1, gives an upperbound, namely, $\mu' < H(\mu)$:

Theorem 3. (Maximal Inconsistency of Bayes) *Let S^i be the sequence consisting of the first i examples $(x_1, y_1), \ldots, (x_i, y_i)$. For all priors P nonzero on a set of classifiers $\mathcal{C}$, for all learning problems D with $\inf_{c\in\mathcal{C}} e_D(c) = \mu$, for all $\delta > 0$, for all large m, with D^m-probability $\geq 1 - \exp(-2\sqrt{m})$,*

$$\frac{1}{m}\sum_{i=1}^{m}\left|y_i - c_{\text{BAYES}(P,S^{i-1})}(x_i)\right| \leq H(\mu) + \delta.$$

---

[4] In fact, a meta-argument can be made that *any* nontrivial learning algorithm is 'inconsistent' in this sense for finite m.

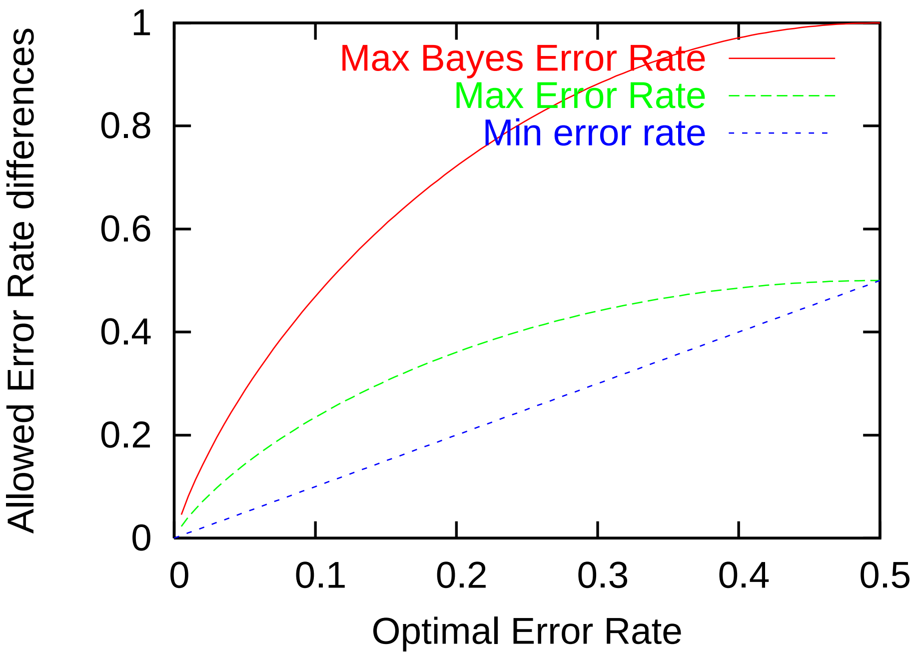

Fig. 1. A graph depicting the set of asymptotically allowed error rates for different classification algorithms. The x-axis depicts the optimal classifier's error rate μ (also shown as the straight line). The lower curve is just $0.5H(\mu)$ and the upper curve is $H(\mu)$. Theorem 2 says that any (μ, μ') between the straight line and the lower curve can be achieved for some learning problem D and prior P. Theorem 3 shows that the Bayesian learner can never have asymptotic error rate μ' above the upper curve.

The theorem says that for large m, the total number of mistakes when successively classifying y_i given x_i made by the Bayesian algorithm based on S^{i-1}, divided by m, is not larger than $H(\mu)$. By the law of large numbers, it follows that for large m, $e_D(c_{\text{BAYES}(P,S^{i-1})}(x_i))$, *averaged* over all i, is no larger than $H(\mu)$. Thus, it is not ruled out that sporadically, for some i, $e_D(c_{\text{BAYES}(P,S^{i-1})}(x_i)) > H(\mu)$; but this must be 'compensated' for by most other i. We did not find a proof that $e_D(c_{\text{BAYES}(P,S^{i-1})}(x_i)) < H(\mu)$ for *all* large i.

4 Proofs

In this section we present the proofs of our three theorems. Theorem 2 and 3 both make use of the following lemma:

Lemma 1. *There exists $\gamma > 0$ such that for all classifiers c, $\alpha > 0$, $m > 0$, all $S \sim D^m$ satisfying $\alpha + 1/\sqrt{m} < \hat{e}_S(c) \leq 0.5$, all priors satisfying (7):*

$$\log \frac{1}{P(y^m \mid x^m, c, \hat{e}_S(c))} \leq \log \frac{1}{P(y^m \mid x^m, c)} \leq \log \frac{1}{P(y^m \mid x^m, c, \hat{e}_S(c))} + \frac{1}{2}\log m + \frac{1}{2}\frac{1}{\alpha(1-\alpha)} - \log \gamma. \quad (8)$$

Proof. **(sketch)** For the first inequality, note

$$\log \frac{1}{P(y^m \mid x^m, c)} = \log \frac{1}{\int P(y^m \mid x^m, c, \theta) P(\theta) d\theta} \geq \log \frac{1}{P(y^m \mid x^m, c, \hat{e}_S(c))},$$

since the likelihood $P(y^m \mid x^m, c, \theta)$ is maximized at $\theta = \hat{e}_S(c)$. For the second inequality, note that

$$\int_0^1 P(y^m \mid x^m, c, \theta) P(\theta) d\theta \geq \int_{\hat{e}_S(c)-1/\sqrt{m}}^{\hat{e}_S(c)+1/\sqrt{m}} \exp\bigl(\log P(y^m \mid x^m, c, \theta) + \log P(\theta)\bigr) d\theta.$$

We obtain (8) by expanding $\log P(y^m \mid x^m, c, \theta)$ around the maximum $\theta = \hat{e}_S(c)$ using a second-order Taylor approximation. See, [2] for further details.

4.1 Inconsistent Learning Algorithms: Proof of Theorem 2

Below we first define the particular learning problem that causes inconsistency. We then analyze the performance of the algorithms on this learning problem.

The Learning Problem For given μ and $\mu' \geq \mu$, we construct a learning problem and a set of classifiers $\mathcal{C} = \{c_0, c_1, \ldots\}$ such that c_0 is the 'good' classifier with $e_D(c_0) = \mu$ and $c_1, c_2, \ldots$ are all 'bad' classifiers with $e_D(c_j) = \mu' \geq \mu$. $\mathcal{X}$ consists of one binary feature per classifier[5], and the classifiers simply output the value of their special feature. The underlying distribution D is constructed in terms of μ and μ' and a proof parameter $\mu_{\text{hard}} \geq \frac{1}{2}$ (the error rate for "hard" examples). To construct an example (x, y), we first flip a fair coin to determine y, so $y = 1$ with probability $1/2$. We then flip a coin with bias $p_{\text{hard}} := \frac{\mu'}{\mu_{\text{hard}}}$ which determines if this is a "hard" example or an "easy" example. Based upon these two coin flips, each x_j is independently generated based on the following 3 cases.

1. For a "hard" example, and for each classifier c_j with $j \geq 1$, set $x_j = |1 - y|$ with probability μ_{hard} and $x_j = y$ otherwise.
2. For an "easy" example, and every $j \geq 1$ set $x_j = y$.
3. For the "good" classifier c_0 (with true error rate μ), set $x_0 = |1 - y|$ with probability μ and $x_0 = y$ otherwise.

The error rates of each classifier are $e_D(c_0) = \mu$ and $e_D(c_j) = \mu'$ for all $j \geq 1$.

[5] This input space has a countably infinite size. The Bayesian posterior is still computable for any finite m if we order the features according to the prior of the associated classifier. We need only consider features which have an associated prior greater than $\frac{1}{2^m}$ since the minus log-likelihood of the data is always less than m bits. Alternatively, we could use stochastic classifiers and a very small input space.

Bayes and MDL are inconsistent We now prove Theorem 2. In Stage 1 we show that there exists a k_m such that for every value of m, with probability converging to 1, there exists some 'bad' classifier c_j with $0 < j \leq k_m$ that has 0 empirical error. In Stage 2 we show that the prior of this classifier is large enough so that its posterior is exponentially larger than that of the good classifier c_0, showing the convergence $e_D(c_{\text{MAP}(P,S)}) \to \mu'$. In Stage 3 we sketch the convergences $e_D(c_{\text{SMAP}(P,S)}) \to \mu', e_D(c_{\text{MDL}(P,S)}) \to \mu', e_D(c_{\text{BAYES}(P,S)}) \to \mu'$.

Stage 1 Let m_{hard} denote the number of hard examples generated within a sample S of size m. Let k be a positive integer and $\mathcal{C}_k = \{c_j \in \mathcal{C} : 1 \leq j \leq k\}$. For all $\epsilon > 0$ and $m \geq 0$, we have:

$$
\begin{aligned}
&\Pr_{S\sim D^m}(\forall c \in \mathcal{C}_k : \ \hat{e}_S(c) > 0) \\
&\overset{(a)}{=} \Pr_{S\sim D^m}\left(\forall c \in \mathcal{C}_k : \ \hat{e}_S(c) > 0 \mid \frac{m_{\text{hard}}}{m} > p_{\text{hard}} + \epsilon\right) \Pr_{S\sim D^m}\left(\frac{m_{\text{hard}}}{m} > p_{\text{hard}} + \epsilon\right) \\
&+ \Pr_{S\sim D^m}\left(\forall c \in \mathcal{C}_k : \ \hat{e}_S(c) > 0 \mid \frac{m_{\text{hard}}}{m} \leq p_{\text{hard}} + \epsilon\right) \Pr_{S\sim D^m}\left(\frac{m_{\text{hard}}}{m} \leq p_{\text{hard}} + \epsilon\right) \\
&\overset{(b)}{\leq} e^{-2m\epsilon^2} + \Pr_{S\sim D^m}\left(\forall c \in \mathcal{C}_k : \ \hat{e}_S(c) > 0 \mid \frac{m_{\text{hard}}}{m} \leq p_{\text{hard}} + \epsilon\right) \\
&\overset{(c)}{\leq} e^{-2m\epsilon^2} + (1 - (1-\mu_{\text{hard}})^{m(p_{\text{hard}}+\epsilon)})^k \overset{(d)}{\leq} e^{-2m\epsilon^2} + e^{-k(1-\mu_{\text{hard}})^{m(p_{\text{hard}}+\epsilon)}}. \quad (9)
\end{aligned}
$$

Here (a) follows because $P(a) = \sum_b P(a|b)P(b)$. (b) follows by $\forall a, P : \ P(a) \leq 1$ and the Chernoff bound. (c) holds since $(1-(1-\mu_{\text{hard}})^{m(p_{\text{hard}}+\epsilon)})^k$ is monotonic in ϵ, and (d) by $\forall x \in [0,1], k > 0 : \ (1-x)^k \leq e^{-kx}$. We now set $\epsilon_m := m^{-0.25}$ and $k(m) = \frac{2m\epsilon_m^2}{(1-\mu_{\text{hard}})^{m(p_{\text{hard}}+\epsilon_m)}}$. Then (9) becomes

$$
\Pr_{S\sim D^m}(\forall c \in \mathcal{C}_{k(m)} : \ \hat{e}_S(c) > 0) \leq 2e^{-2\sqrt{m}} \quad (10)
$$

On the other hand, by the Chernoff bound we have $\Pr_{S\sim D^m}(\hat{e}_S(c_0) < e_D(c_0) - \epsilon_m) \leq e^{-2\sqrt{m}}$ for the optimal classifier c_0. Combining this with (10) using the union bound, we get that, with D^m-probability larger than $1 - 3e^{-2\sqrt{m}}$, the following event holds:

$$
\exists c \in \mathcal{C}_{k(m)} : \ \hat{e}_S(c) = 0 \text{ and } \ \hat{e}_S(c_0) \geq e_D(c_0) - \epsilon_m. \quad (11)
$$

Stage 2 In the following derivation, we assume that the large probability event (11) holds. We show that this implies that for large m, the posterior on some $c^* \in \mathcal{C}_{k(m)}$ with $\hat{e}_S(c^*) = 0$ is greater than the posterior on c_0, which implies that the MAP algorithm is inconsistent. Taking the log of the posterior ratios, we get:

$$
\begin{aligned}
&\log \frac{\max_\theta P(c_0, \theta \mid x^m, y^m)}{\max_\theta P(c^*, \theta \mid x^m, y^m)} = \log \frac{\max_\theta P(c_0)P(\theta)P(y^m \mid x^m, c_0, \theta)}{\max_\theta P(c^*)P(\theta)P(y^m \mid x^m, c^*, \theta)} = \\
&\log \max_\theta P(c_0)P(\theta)P(y^m \mid x^m, c_0, \theta) - \log \max_\theta P(c^*)P(\theta)P(y^m \mid x^m, c^*, \theta). \quad (12)
\end{aligned}
$$

Using (2) we see that the leftmost term is no larger than

$$\log\left(\max_{\theta} P(c_0)P(\theta)\right)\cdot\left(\max_{\theta'} P(y^m \mid x^m, c_0, \theta')\right) = -mH(\hat{e}_S(c_0)) + O(1) \leq$$
$$-mH(e_D(c_0)) - Km\epsilon_m + O(1) = -mH(\mu) - m^{0.75}K + O(1) \quad (13)$$

where K is some constant. The last line follows because $H(\mu)$ is continuously differentiable in a small enough neighborhood around μ.

For the rightmost term in (12), by the condition on prior $p(\theta)$, (7),

$$-\log \max_{\theta} P(c^*)P(\theta)P(y^m \mid x^m, c^*, \theta) \leq -\log P(c^*) + \log\gamma. \quad (14)$$

Using condition (6) on prior $P(c^*)$ and using $c^* \in \mathcal{C}_{k(m)}$, we find:

$$\log\frac{1}{P(c^*)} \leq \log k(m) + o(\log k(m)), \quad (15)$$

where $\log k(m) = \log 2\sqrt{m} - (mp_{\text{hard}} + m^{0.25})\log(1-\mu_{\text{hard}})$. Choosing $\mu_{\text{hard}} = 1/2$, this becomes $\log k(m) = \frac{1}{2}\log m + 2m\mu' + m^{0.25} + O(1)$. Combining this with (15), we find that

$$\log\frac{1}{P(c^*)} \leq 2m\mu' + o(m) \quad (16)$$

which implies that (14), is no larger than $2m\mu' + o(m)$. Since $\mu' < H(\mu)/2$, the difference between the leftmost term (13) and the rightmost term (14) in (12) is less than 0 for large m, implying that then $e_D(c_{\text{MAP}(P,S)}) = \mu'$. We derived all this from (11) which holds with probability $\geq 1 - 3\exp(-2\sqrt{m})$. Thus, for all large m, $\Pr_{S\sim D^m}\left(c_{\text{MAP}(P,S)} = \mu'\right) \geq 1 - 3\exp(-2\sqrt{m})$, and the result follows.

Stage 3 **(sketch)** The proof that the integrated MAP classifier $c_{\text{SMAP}(P,S)}$ is inconsistent is similar to the proof for $c_{\text{MAP}(P,S)}$ that we just gave, except that (12) now becomes

$$\log P(c_0)P(y^m \mid x^m, c_0) - \log P(c^*)P(y^m \mid x^m, c^*). \quad (17)$$

By Lemma 1 we see that, if (11) holds, the difference between (12) and (17) is of order $O(\log m)$. The proof then proceeds exactly as for the MAP case.

To prove inconsistency of $c_{\text{MDL}(P,S)}$, note that the MDL code length of y^m given x^m according to c_0 is given by $\log\binom{m}{m\hat{e}_S(c_0)}$. If (11) holds, then a simple Stirling's approximation as in [12] or [15] shows that $\log\binom{m}{m\hat{e}_S(c_0)} = mH(\hat{e}_S(c_0)) - O(\log m)$. Thus, the difference between two-part codelengths achieved by c_0 and c^* is given by

$$-mH(\hat{e}_S(c_0)) + O(\log m) - \log P(c^*). \quad (18)$$

The proof then proceeds as for the MAP case, with (12) replaced by (18) and a few immediate adjustments.

To prove inconsistency of $c_{\text{BAYES}(P,S)}$, we take μ_{hard} not equal to $1/2$ but to $1/2 + \delta$ for some small $\delta > 0$. By taking δ small enough, the proof for $c_{\text{MAP}(P,S)}$ above goes through unchanged so that, with probability $\geq 1 - 3\exp(-2\sqrt{m})$, the Bayesian posterior puts all its weight, except for an exponentially small part, on a mixture of distributions $p_{c_j,\theta}$ whose Bayes classifier has error rate μ' and error rate on hard examples $> 1/2$. It can be shown that this implies that for large m, the classification error $c_{\text{BAYES}(P,S)}$ converges to μ'; we omit details.

4.2 A Consistent Algorithm: Proof of Theorem 1

In order to prove the theorem, we first state the Occam's Razor Bound classification algorithm, based on minimizing the bound given by the following theorem.

Theorem 4. *(Occam's Razor Bound)* [7] *For all priors P on a countable set of classifiers $\mathcal{C}$, for all distributions D, with probability $1-\delta$:*

$$\forall c: \quad e_D(c) \leq \hat{e}_S(c) + \sqrt{\frac{\ln \frac{1}{P(c)} + \ln \frac{1}{\delta}}{2m}}.$$

We state the algorithm here in a suboptimal form, which good enough for our purposes (see [18] for more sophisticated versions):

$$c_{\mathrm{ORB}(P,S)} := \arg\min_{c \in \mathcal{C}} \hat{e}_S(c) + \sqrt{\frac{\ln \frac{1}{P(c)} + \ln m}{2m}}.$$

Proof of Theorem 1 Set $\delta_m := 1/m$. It is easy to see that

$$\min_{c \in \mathcal{C}} e_D(c) + \sqrt{\frac{\ln \frac{1}{P(c)} + \ln m}{2m}}$$

is achieved for at least one $c \in \mathcal{C} = \{c_0, c_1, \ldots\}$. Among all $c_j \in \mathcal{C}$ achieving the minimum, let $\tilde{c}_m$ be the one with smallest index j. By the Chernoff bound, we have with probability at least $1 - \delta_m = 1 - 1/m$,

$$e_D(\tilde{c}_m) \geq \hat{e}_S(\tilde{c}_m) - \sqrt{\frac{\ln(1/\delta_m)}{2m}} = \hat{e}_S(\tilde{c}_m) - \sqrt{\frac{\ln m}{2m}}, \tag{19}$$

whereas by Theorem 4, with probability at least $1 - \delta_m = 1 - 1/m$,

$$e_D(c_{\mathrm{ORB}(P,S)}) \leq \min_{c \in \mathcal{C}} \hat{e}_S(c) + \sqrt{\frac{-\ln P(c) + \ln m}{2m}} \leq \hat{e}_S(\tilde{c}_m) + \sqrt{\frac{-\ln P(\tilde{c}_m) + \ln m}{2m}}.$$

Combining this with (19) using the union bound, we find that

$$e_D(c_{\mathrm{ORB}(P,S)}) \leq e_D(\tilde{c}_m) + \sqrt{\frac{-\ln P(\tilde{c}_m) + \ln m}{2m}} + \sqrt{\frac{\ln m}{2m}},$$

with probability at least $1 - 2/m$. The theorem follows upon noting that the right-hand side of this expression converges to $\inf_{c \in \mathcal{C}} e_D(c)$ with increasing m.

4.3 Proof of Theorem 3

Without loss of generality assume that c_0 achieves $\min_{c \in \mathcal{C}} e_D(c)$. Consider both the 0/1-loss and the log loss of sequentially predicting with the Bayes predictive distribution $P(Y_i = \cdot \mid X_i = \cdot, S^{i-1})$ given by $P(y_i \mid x_i, S^{i-1}) = E_{c,\theta \sim P(\cdot \mid S^{i-1})} p_{c,\theta}(y_i | x_i)$. Every time $i \in \{1, \ldots, m\}$ that the Bayes classifier based on S^{i-1} classifies y_i incorrectly, $P(y_i \mid x_i, S^{i-1})$ must be $\leq 1/2$ so that $-\log P(y_i \mid x_i, S^{i-1}) \geq 1$. Therefore,

$$\sum_{i=1}^{m} -\log P(y_i \mid x_i, S^{i-1}) \geq \sum_{i=1}^{m} |y_i - c_{\mathrm{BAYES}(P,S^{i-1})}(x_i)|. \tag{20}$$

On the other hand we have

$$\sum_{i=1}^{m} - \log P(y_i \mid x_i, S^{i-1}) = - \log \prod_{i=1}^{m} P(y_i \mid x_i, x^{i-1}, y^{i-1}) =$$
$$- \log \prod_{i=1}^{m} P(y_i \mid x^m, y^{i-1}) = - \log \prod_{i=1}^{m} \frac{P(y^i|x^m)}{P(y^{i-1}|x^m)} = - \log P(y^m \mid x^m) =$$
$$- \log \sum_{j=0,1,2\ldots} P(y^m \mid x^m, c_j) P(c_j) \leq - \log P(y^m \mid x^m, c_0) - \log P(c_0), \quad (21)$$

where the inequality follows because a sum is larger than each of its terms. By the Chernoff bound, for all small enough $\epsilon > 0$, with probability larger than $1 - 2\exp(-2m\epsilon^2)$, we have $|\hat{e}_S(c_0) - e_D(c_0)| < \epsilon$. We now set $\epsilon_m = m^{-0.25}$. Then, using Lemma 1, with probability larger than $1 - 2\exp(-2\sqrt{m})$, for all large m (21) is less than or equal to

$$- \log P(y^m \mid x^m, c_0, \hat{e}(c_0)) + \frac{1}{2} \log m + C_m \overset{(a)}{=} mH(\hat{e}_S(c_0)) + \frac{1}{2} \log m + C_m \leq$$
$$mH(e_D(c_0)) + Km^{0.75} + \frac{1}{2} \log m + C_m, \quad (22)$$

where $C_m = (e_D(c_0) - \epsilon_m - m^{-0.5})^{-1}(1 - e_D(c_0) + \epsilon_m + m^{-0.5})^{-1}$ and K is a constant not depending on $S = S^m$. Here (a) follows from Equation 2 and (b) follows because $H(\mu)$ is continuously differentiable in a neighborhood of μ.

Combining (22) with (20) and using $C_m = O(1)$ we find that with probability $\geq 1 - \exp(-2\sqrt{m})$, $\sum_{i=1}^{m} |y_i - c_{\text{BAYES}(P, S^{i-1})}(x_i)| \leq mH(e_D(c_0)) + o(m)$, QED.

5 Technical Discussion

5.1 Variations of Theorem 2 and dependency on the prior

Prior on classifiers The requirement (6) that $-\log P(c_k) \geq \log k + o(\log k)$ is needed to obtain (16), which is the key inequality in the proof of Theorem 2. If $P(c_k)$ decreases at polynomial rate, but at a degree d larger than one, i.e. if

$$- \log P(c_k) = d \log k + o(\log k), \quad (23)$$

then a variation of Theorem 2 still applies but the maximum possible discrepancies between μ and μ' become much smaller: essentially, if we require $\mu \leq \mu' < \frac{1}{2d} H(\mu)$ rather than $\mu \leq \mu' < \frac{1}{2} H(\mu)$ as in Theorem 2, then the argument works for all priors satisfying (23). Since the derivative $dH(\mu)/d\mu \to \infty$ as $\mu \downarrow 0$, by setting μ close enough to 0 it is possible to obtain inconsistency for any fixed polynomial degree of decrease d. However, the higher d, the smaller $\mu = \inf_{c \in \mathcal{C}} e_D(c)$ must be to get any inconsistency with our argument.

Prior on error rates Condition (7) on the prior on the error rates is satisfied for most reasonable priors. Some approaches to applying MDL to classification problems amount to assuming priors of the form $p(\theta^*) = 1$ for a single $\theta^* \in [0, 1]$ (Section **??**). In that case, we can still prove a version of Theorem 2, but the maximum discrepancy between μ and μ' may now be either larger or smaller than $H(\mu)/2 - \mu$, depending on the choice of θ^*.

5.2 Properties of the transformation from classifiers to distributions

Optimality and Reliability Assume that the conditional distribution of y given x according to the 'true' underlying distribution D is defined for all $x \in \mathcal{X}$, and let $p_D(y|x)$ denote its mass function. Define $\Delta(p_{c,\theta})$ as the Kullback-Leibler (KL) divergence [9] between $p_{c,\theta}$ and the 'true' conditional distribution p_D:

$$\Delta(p_{c,\theta}) := \mathrm{KL}(p_D \| p_{c,\theta}) = E_{(x,y)\sim D}[-\log p_{c,\theta}(y|x) + \log p_D(y|x)].$$

Proposition 1. *Let $\mathcal{C}$ be any set of classifiers, and let $c^* \in \mathcal{C}$ achieve* $\min_{c\in\mathcal{C}} e_D(c) = e_D(c^*)$.

1. *If $e_D(c^*) < 1/2$, then*

$$\min_{c,\theta} \Delta(p_{c,\theta}) \text{ is uniquely achieved for } (c,\theta) = (c^*, e_D(c^*)).$$

2. $\min_{c,\theta} \Delta(p_{c,\theta}) = 0$ *iff $p_{c^*,e_D(c^*)}$ is 'true', i.e. if $\forall x, y : p_{c^*,e_D(c^*)}(y|x) = p_D(y|x)$.*

Property 1 follows since for each fixed c, $\min_{\theta\in[0,1]} \Delta(p_{c,\theta})$ is uniquely achieved for $\theta = e_D(c)$ (this follows by differentiation) and satisfies $\min_\theta \Delta(p_{c,\theta}) = \Delta(p_{c,e_D(c)}) = H(e_D(c)) - K_D$, where $K_D = E[\log p_D(y|x)]$ does not depend on c or θ, and $H(\mu)$ is monotonically increasing for $\mu < 1/2$. Property 2 follows from the information inequality [9].

Proposition 1 implies that our transformation is a good candidate for turning classifiers into probability distributions.

Namely, let $\mathcal{P} = \{p_\alpha : \alpha \in A\}$ be a set of i.i.d. distributions indexed by parameter set A and let $P(\alpha)$ be a prior on A. By the law of large numbers, for each $\alpha \in A$, $m^{-1} \log p_\alpha(y^m \mid x^m)P(\alpha) \to \mathrm{KL}(p_D \| p_\alpha)$. By Bayes rule, this implies that if the class $\mathcal{P}$ is 'small' enough so that the law of large numbers holds *uniformly* for all $p_\alpha \in \mathcal{P}$, then for all $\epsilon > 0$, the Bayesian posterior will concentrate, with probability 1, on the set of distributions in $\mathcal{P}$ within ϵ of the $p^* \in \mathcal{P}$ minimizing KL-divergence to D. In our case, if $\mathcal{C}$ is 'simple' enough so that the corresponding $\mathcal{P} = \{p_{c,\theta} : c \in \mathcal{C}, \theta \in [0,1]\}$ admits uniform convergence [12], then the Bayesian posterior asymptotically concentrates on the $p_{c^*,\theta^*} \in \mathcal{P} = \{p_{c,\theta}\}$ closest to D in KL-divergence. By Proposition 1, this p_{c^*,θ^*} corresponds to the $c^* \in \mathcal{C}$ with smallest generalization error rate $e_D(c^*)$ (p_{c^*,θ^*} is *optimal* for 0/1-loss), and for the $\theta^* \in [0,1]$ with $\theta^* = e_D(c^*)$ (p_{c^*,θ^*} gives a *reliable* impression of its prediction quality). This convergence to an optimal and reliable p_{c^*,θ^*} will happen if, for example, $\mathcal{C}$ has finite VC-dimension [12]. We can only get trouble as in Theorem 2 if we allow $\mathcal{C}$ to be of infinite VC-dimension.

Logistic regression interpretation let $\mathcal{C}$ be a set of functions $\mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}$ ($\mathcal{Y}$ does not need to be binary-valued). The corresponding logistic regression model is the set of conditional distributions $\{p_{c,\beta} : c \in \mathcal{C}; \beta \in \mathbb{R}\}$ of the form

$$p_{c,\beta}(1 \mid x) := \frac{e^{-\beta c(x)}}{1 + e^{-\beta c(x)}} \quad ; \quad p_{c,\beta}(0 \mid x) := \frac{1}{1 + e^{-\beta c(x)}}. \tag{24}$$

This is the standard construction used to convert classifiers with real-valued output such as support vector machines and neural networks into conditional distributions [14, 22], so that Bayesian inference can be applied. By setting $\mathcal{C}$

to be a set of $\{0,1\}$-valued classifiers, and substituting $\beta = \ln(1-\theta) - \ln\theta$ as in (**??**), we see that our construction is a special case of the logistic regression transformation (24). It may seem that (24) does not treat $y = 1$ and $y = 0$ on equal footing, but this is not so: we can alternatively define a symmetric version of (24) by defining, for each $c \in \mathcal{C}$, a corresponding $c' : \mathcal{X} \to \{-1, 1\}$, $c'(x) := 2c(x) - 1$. Then we can set

$$p_{c,\beta}(1 \mid x) := \frac{e^{-\beta c(x)}}{e^{\beta c(x)} + e^{-\beta c(x)}} \quad ; \quad p_{c,\beta}(-1 \mid x) := \frac{e^{\beta c(x)}}{e^{\beta c(x)} + e^{-\beta c(x)}}. \tag{25}$$

By setting $\beta' = 2\beta$ we see that $p_{c,\beta}$ as in (24) is identical to $p_{c,\beta'}$ as in (25), so that the two models really coincide.

6 Interpretation from a Bayesian perspective

Bayesian Consistency It is well-known that Bayesian inference is strongly consistent under very broad conditions. For example, when applied to our setting, the celebrated Blackwell-Dubins consistency theorem [6] says the following. Let $\mathcal{C}$ be countable and suppose D is such that, for some $c^* \in \mathcal{C}$ and $\theta^* \in [0,1]$, p_{c^*,θ^*} is equal to p_D, the true distribution/ mass function of y given x. Then with D-probability 1, the Bayesian posterior concentrates on c^*: $\lim_{m\to\infty} P(c^* \mid S^m) = 1$.

Consider now the learning problem underlying Theorem 2 as described in Section 4.1. Since c_0 achieves $\min_{c\in\mathcal{C}} e_D(c)$, it follows by part 1 of Proposition 1 that $\min_{c,\theta} \Delta(p_{c,\theta}) = \Delta(p_{c_0,e_D(c_0)})$. *If* $\Delta(p_{c_0,e_D(c_0)})$ were 0, then by part 2 of Proposition 1, Blackwell-Dubins would apply, and we would have $P(c_0 \mid S^m) \to 1$. Theorem 2 states that this does *not* happen. It follows that the premisse $\Delta(p_{c_0,e_D(c_0)}) = 0$ must be false. But since $\Delta(p_{c,\theta})$ is minimized for $(c_0, e_D(c_0))$, the Proposition implies that for *no* $c \in \mathcal{C}$ and *no* $\theta \in [0,1]$, $p_{c,\theta}$ is equal to $p_D(\cdot|\cdot)$ - in statistical terms, the model $\mathcal{P} = \{p_{c,\theta} : c \in \mathcal{C}, \theta \in [0,1]\}$ is *misspecified.*

Why is the result interesting for a Bayesian? Here we answer several objections that a Bayesian might have to our work.

Bayesian inference has never been designed to work under misspecification. So why is the result relevant?
We would maintain that in *practice*, Bayesian inference is applied *all the time* under misspecification in classification problems [12]. It is very hard to avoid misspecification with Bayesian classification, since the modeler often has no idea about the noise-generating process. Even though it may be known that noise is not homoskedastic, it may be practically impossible to incorporate all ways in which the noise may depend on x into the prior.

It is already well-known that Bayesian inference can be inconsistent even if $\mathcal{P}$ is well-specified, *i.e. if it contains D [10]. So why is our result interesting?*
The (in)famous inconsistency results by Diaconis and Freedman [10] are based on nonparametric inference with uncountable sets $\mathcal{P}$. Their theorems require that the true p has small prior density, and in fact prior *mass* 0 (see also [1]). In contrast, Theorem 2 still holds if we assign $p_{c_0,e_D(c_0)}$ arbitrarily large prior mass < 1, which, by the Blackwell-Dubins theorem, guarantees consistency if $\mathcal{P}$ is well-specified. We show that consistency may *still* fail dramatically if $\mathcal{P}$ is

misspecified. This is interesting because even under misspecification, Bayes is consistent under fairly broad conditions [8, 16], in the sense that the posterior concentrates on a neighborhood of the distribution that minimizes KL-divergence to the true D. Thus, we feel our result is relevant at least from the *inconsistency under misspecification* interpretation.

So how can our result co-exist with theorems establishing Bayesian consistency under misspecification?
Such results are typically proved under either one of the following two assumptions:

1. The set of distributions $\mathcal{P}$ is 'simple', for example, finite-dimensional parametric. In such cases, ML estimation is usually also consistent - thus, for large m the role of the prior becomes negligible. In case $\mathcal{P}$ corresponds to a classification model $\mathcal{C}$, this would obtain, for example, if $\mathcal{C}$ were finite or had finite VC-dimension.
2. $\mathcal{P}$ may be arbitrarily large or complex, but it is *convex*: any finite mixture of elements of $\mathcal{P}$ is an element of $\mathcal{P}$. An example is the family of Gaussian mixtures with an arbitrary but finite number of components [17].

Our setup violates both conditions: $\mathcal{C}$ has infinite VC-dimension, and the corresponding $\mathcal{P}$ is not closed under taking mixtures. This suggests that we could make Bayes consistent again if, instead of $\mathcal{P}$, we would base inferences on its convex closure $\overline{\mathcal{P}}$. Computational difficulties aside,this approach will not work, since we now use the crucial part (1) of Proposition 1 will not hold any more: the conditional distribution in $\overline{\mathcal{P}}$ closest in KL-divergence to the true $p_D(y|x)$, when used for classification, may end up having larger generalization error (expected 0/1-loss) than the optimal classifier c^* in the set $\mathcal{C}$ on which $\mathcal{P}$ was based. We will give an explicit example of this in the journal version of this paper. Thus, with a prior on $\overline{\mathcal{P}}$, the Bayesian posterior will converge, but potentially it converges to a distribution that is suboptimal in the performance measure we are interested in.

How 'standard' is the conversion from classifiers to probability distributions on which our results are based?
One may argue that our notion of 'converting' classifiers into probability distributions is not always what Bayesians do in practice. For classifiers which produce *real-valued* output, such as neural networks and support vector machines, our transformation coincides with the logistic regression transformation, which is a standard Bayesian tool; see for example [14, 22]. But our theorems are based on classifiers with 0/1-output. With the exception of decision trees, such classifiers have not been addresses frequently in the Bayesian literature. Decision trees have usually been converted to conditional distributions differently, by assuming a different noise rate *in each leaf* of the decision tree [13]. This makes the set of all decision trees on a given input space $\mathcal{X}$ coincide with the set of all conditional distributions on $\mathcal{X}$, and thus avoids the misspecification problem, at the cost of using a much larger model space.

Thus, this a weak point in our analysis: we use a transformation that has mostly been applied to real-valued classifiers, whereas our classifiers are 0/1-valued. Whether our inconsistency results can be extended in a natural way to classifiers with real-valued output remains to be seen. The fact that the Bayesian model corresponding to such neural networks will still typically be misspecified suggests (but does not prove) that similar scenarios may be constructed.

7 Acknowledgments

The ideas in this paper were developed in part during the workshop *Complexity and Inference*, held at the DIMACS center, Rutgers University, June 2003. We would like to thank Mark Hansen, Paul Vitányi and Bin Yu for organizing this workshop, and Dean Foster and Abraham Wyner for stimulating conversations during the workshop.

References

1. Andrew R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics*, volume 6, pages 27–52. Oxford University Press, 1998.
2. Andrew R. Barron, Jorma Rissanen, and Bin Yu. The MDL Principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6):2743–2760, 1998.
3. A.R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics*, pages 561–576. Kluwer Academic Publishers, 1990.
4. A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, 1991.
5. J.M. Bernardo and A.F.M Smith. *Bayesian theory.* John Wiley, 1994.
6. D. Blackwell and L. Dubins. Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, 33:882–886, 1962.
7. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, April 1987.
8. O. Bunke and X. Milhaud. Asymptotic behaviour of Bayes estimates under possibly incorrect models. *The Annals of Statistics*, 26:617–644, 1998.
9. T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley, 1991.
10. P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statitics*, 14(1):1–26, 1986.
11. P. D. Grünwald. MDL tutorial. In P. D. Grünwald, I. J. Myung, and M. A. Pitt, editors, *Minimum Description Length: recent developments in theory and practice*, chapter 1. MIT Press, 2004. to appear.
12. P.D. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty.* PhD thesis, University of Amsterdam, The Netherlands, 1998.
13. D. Heckerman, D.M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2000.
14. M.I. Jordan. Why the logistic funtion? a tutorial discussion on probabilities and neural networks. Computational Cognitive Science Tech. Rep. 9503, MIT, 1995.
15. M. Kearns, Y. Mansour, A.Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.
16. Bas Kleijn and Aad van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. submitted, 2004.
17. J.K. Li. *Estimation of Mixture Models.* PhD thesis, Yale University, Department of Statistics, 1997.
18. D. McAllester. PAC-Bayesian model averaging. In *Proceedings COLT '99*, 1999.
19. R. Meir and N. Merhav. On the stochastic complexity of learning realizable and unrealizable rules. *Machine Learning*, 19:241–261, 1995.
20. J. Quinlan and R. Rivest. Inferring decision trees using the minimum description length principle. *Information and Computation*, 80:227–248, 1989.
21. J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, 1989.
22. M.E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.
23. M. Viswanathan., C.S. Wallace, D.L. Dowe, and K.B. Korb. Finding cutpoints in noisy binary sequences - a revised empirical evaluation. In *Proc. 12th Australian Joint Conf. on Artif. Intelligence*, volume 1747 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 405–416, Sidney, Australia, 1999.
24. K. Yamanishi. A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Trans. Inform. Theory*, 44(4):1424–1439, 1998.