Mach Learn (2007) 66:119–149
DOI 10.1007/s10994-007-0716-7

Suboptimal behavior of Bayes and MDL in classification under misspecification

Peter Grünwald · John Langford

Received: 29 March 2005 / Revised: 15 December 2006 / Accepted: 20 December 2006 / Published online: 30 January 2007
Springer Science + Business Media, LLC 2007

Abstract We show that forms of Bayesian and MDL inference that are often applied to classification problems can be *inconsistent*. This means that 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. From a Bayesian point of view, the result can be reinterpreted as saying that Bayesian inference can be inconsistent under misspecification, even for countably infinite models. We extensively discuss the result from both a Bayesian and an MDL perspective.

Keywords Bayesian statistics · Minimum description length · Classification · Consistency · Inconsistency · Misspecification

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 (Bernardo & Smith, 1994) and the Minimum Description Length (MDL) Principle (Rissanen, 1989; Barron, Rissanen, & Yu, 1998; Grünwald, 2005, 2007). 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

Editors: Olivier Bousquet and Andre Elisseeff

P. Grünwald (✉)
CWI, Amsterdam, the Netherlands
e-mail: pdg@cwi.nl

J. Langford
Yahoo Research, New York
e-mail: jl@hunch.net

or the full Bayesian posterior does not converge to the minimum achievable generalization error within the set of classifiers under consideration.

This result may be viewed as a challenge to Bayesian inference. Given a powerful piece of information (an objectively correct "prior" on a set of *classifiers*), transforming this information into a Bayesian prior on a set of *distributions* in a straightforward manner and applying Bayes rule yields significantly suboptimal behavior—while another simple approach yields optimal behavior. The key is the transformation from classifiers (functions mapping each input X to a discrete class label Y) to (conditional) distributions. Such a transformation is necessary because Bayes rule cannot be directly applied to classifiers. We do the conversion in a straightforward manner, crossing a prior on classifiers with a prior on error rates for these classifiers. This conversion method is a completely standard tool for Bayesians active in the field of machine learning—we tested this with some professing Bayesians, see Section 6.3.4—yet it inevitably leads to (sometimes subtly) 'misspecified' probability models not containing the 'true' distribution D. The result may therefore be re-interpreted as 'Bayesian inference can be inconsistent under misspecification for common classification probability models'. Since, in practice, Bayesian inference for classification tasks is frequently and inevitably based on misspecified probability models, the result remains relevant even if (as many Bayesians do, especially those not active in the field of machine learning) one insists that inference starts directly with a probability model, rather than a classification model that is then transformed into a probability model—see Section 6.

There are two possible resolutions to this challenge. Perhaps Bayesian inference is an incomplete characterization of learning: there exist pieces of information (e.g. prior information on deterministic classifiers rather than distributions) which can not be well integrated into a prior distribution and so learning algorithms other than Bayesian inference are sometimes necessary. Or, perhaps there is some less naive method allowing a prior to express the available information. We discuss this issue further in Section 6.

1.1 A preview

1.1.1 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.

1.1.2 The basic result

The 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), "Minimum Description Length classification" (MDL) and "Occam's Razor Bound classification" (ORB). 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 $\mathcal{C}$. The result applies even when the process creating classification problems draws the optimal classifier from $P(c)$. In Section 3 we state the 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 $\mathcal{X}$ and a prior $P(c)$ on a countable set of classifiers $\mathcal{C}$ such that:*

1. (inconsistency according to true distribution). *There exists a learning problem (distribution) D 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 suboptimal with respect to the ORB classifier $c_{\mathrm{ORB}(P,S)}$. That is, for $c^* \in \{c_{\mathrm{BAYES}(P,S)}, c_{\mathrm{MAP}(P,S)}, c_{\mathrm{MDL}(P,S)}\}$,*

$$\lim_{m\to\infty} \Pr_{S\sim D^m} (e_D(c_{\mathrm{ORB}(P,S)}) + 0.05 < e_D(c^*)) = 1. \qquad (1)$$

2. (inconsistency according to prior). *There exists a randomized algorithm selecting learning problems D in such a way that*

– (a) *the prior $P(c)$ is 'objectively correct' in the sense that, for all $c \in \mathcal{C}$, with probability $P(c)$, c is the optimal classifier, achieving* $\min_{c\in\mathcal{C}} e_D(c)$.

– (b) (1) *holds no matter what learning problem D/classifier c is selected. In particular,* (1) *holds with prior probability* 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 suboptimality become and how small can $e_D(c_{\mathrm{ORB}(P,S)})$ be?
3. Does this matter for logarithmic loss (which is what MDL approaches seek to minimize (Grünwald, 2007)) rather than 0-1 loss?
4. Is $c_{\mathrm{ORB}(P,S)}$ an algorithm which contains information specific to the distribution D?
5. Is this theorem relevant at small (and in particular noninfinite) sample sizes?

We will ask a number of more detailed questions from a Bayesian perspective in Section 6 and from an MDL perspective in Section 7. 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. The size of the suboptimality can be quite large, at least for the MAP and MDL classifiers (the number of 0.05 was just chosen for concreteness; other values are possible) and $e_D(c_{\mathrm{ORB}(P,S)})$ can be quite small—see Section 5.1 and Fig. 1. The short answer to (3) is "yes". A similar result holds for logarithmic loss, see Section 6.1.

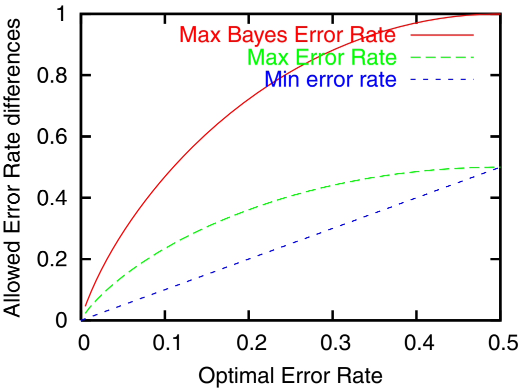

Fig. 1 A graph depicting the set of asymptotically allowed error rates for different classification algorithms. The x-axis depicts the optimal classifier's error rate μ (also shown as the straight line). The lower curve is just $0.5H(\mu)$ and the upper curve is $H(\mu)$. Theorem 2 says that any (μ, 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 answer to (4) is "no". The algorithm $c_{\mathrm{ORB}(P,S)}$, which minimizes the Occam's Razor bound (ORB) (see (Blumer et al., 1987) or Section 4.2), is asymptotically consistent for any D:

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 answer to (5) is that the result is very relevant for small sample sizes because the convergence to the probability 1 event occurs at a speed exponential in m. Although the critical example uses a countably infinite set of classifiers, on a finite set of n classifiers, the analysis implies that for $m < \log n$, Bayesian inference gives poorer performance than Occam's Razor Bound optimization.

Overview of the Paper. 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 the theorems. Technical discussion, including 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 7.

2 Some classification algorithms

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

2.1 The Bayesian classification algorithm

The Bayesian approach to inference starts with a prior probability distribution P over a set of distributions $\mathcal{P}$. P 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:

$$P(p\,|\,S) = \frac{p(S)P(p)}{P(S)}. \tag{2}$$

where $P(S) := E_{p\sim P}[p(S)] = \sum_{p\in\mathcal{P}} P(p)p(S)$, the sum being replaced by an integral when $\mathcal{P}$ is continuous and P admits a density. Note that in Bayesian statistics, $p(S)$ is usually denoted as $P(S\,|\,p)$.

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. In order to apply Bayesian inference, we somehow need to convert the set of classifiers into a corresponding set of distributions $\mathcal{P}$. With such a conversion, the prior $P(c)$ will induce a conventional Bayesian prior on $\mathcal{P}$ after all.

One common conversion (Jordan, 1995; Tipping, 2001; Grünwald, 1998) 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 the special case considered in this paper, $c(x) \in \{0, 1\}$ is binary valued. 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, we let $\mathcal{P}$ consist of the set of distributions $p_{c,\theta}$, where $c \in \mathcal{C}$ and $\theta \in [0, 1]$. These are defined as conditional probability distributions over the labels given the unlabeled data:

$$p_{c,\theta}(y^m\,|\,x^m) = \theta^{m\hat{e}_S(c)}(1-\theta)^{m-m\hat{e}_S(c)}. \tag{3}$$

This expresses that there exists some θ such that $\forall x\;\; P_{c,\theta}(c(X) \neq y\,|\,X = x) = \theta$. (homoskedasticity). Note that

$$p_{c,\theta}(y|x) = \begin{cases} \theta & \text{if } c(x) \neq y \\ 1-\theta & \text{if } c(x) = y. \end{cases}$$

For each fixed $\theta < 0.5$, the log likelihood $\log p_{c,\theta}(y^m\,|\,x^m)$ 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 (3) is maximized by setting $\theta := \hat{e}_S(c)$, giving

$$\log \frac{1}{p_{c,\hat{e}_S(c)}(y^m\,|\,x^m)} = mH(\hat{e}_S(c)), \tag{4}$$

where H is the binary entropy $H(\mu) = \mu \log \frac{1}{\mu} + (1-\mu) \log \frac{1}{1-\mu}$, which is strictly increasing for $\mu \in [0, 0.5)$. Here, as everywhere else in the paper, log stands for binary logarithm. Thus, the conversion is "reasonable" in the sense that, both with fixed $\theta < 0.5$ and with the likelihood-maximizing $\theta = \hat{e}_S(c)$ which varies with the data, the likelihood is a decreasing function in the empirical error rate of c, so that classifiers which achieve small error on the data correspond to a large likelihood of the data.

We further assume that *some* distribution p_x on $\mathcal{X}^m$ generates the x-values, and, in particular that this distribution is independent of c and θ. With this assumption, we can apply Bayes rule to get a posterior on $p_{c,\theta}$, denoted as $P(c, \theta\,|\,S)$, without knowing p_x, since the

$p_x(x^m)$-factors cancel:

$$P(c,\theta\,|\,S) = \frac{p_{c,\theta}(y^m|x^m)p_x(x^m)P(c,\theta)}{P(y^m\,|\,x^m)p_x(x^m)} = \frac{p_{c,\theta}(y^m|x^m)P(c,\theta)}{E_{c,\theta\sim P}[p_{c,\theta}(y^m\,|\,x^m)]}. \tag{5}$$

As is customary in Bayesian statistics, here as well as in the remainder of this paper we defined the prior P over (c,θ) rather than directly over $p_{c,\theta}$. The latter choice would be equivalent but notationally more cumbersome.

To make (5) applicable, we need to specify 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 the theorems hold.

Bayes rule (5) is formed into a classifier learning algorithm by choosing the most likely label given the input x and the posterior $P(\cdot|S)$:

$$c_{\text{BAYES}(P,S)}(x) := \begin{cases} 1 & \text{if } E_{c,\theta\sim P(\cdot|S)}[p_{c,\theta}(Y=1|X=x)] > \frac{1}{2}, \\ 0 & \text{if } E_{c,\theta\sim P(\cdot|S)}[p_{c,\theta}(Y=1|X=x)] < \frac{1}{2}. \end{cases} \tag{6}$$

If $E_{c,\theta\sim P(\cdot|S)}[p_{c,\theta}(Y=1|X=x)] = \frac{1}{2}$, then the value of $c_{\text{BAYES}(P,S)}$ is determined by an independent toss of a fair coin.

2.2 The MAP classification algorithm

The integrations of the full Bayesian classifier can be too computationally intensive, so in practice one often predicts using the Maximum A Posteriori (MAP) classifier. This classifier is given by:

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

with ties broken arbitrarily. Integration over $\theta \in [0,1]$ is easy compared to summation over $c \in \mathcal{C}$, so one sometimes uses a learning algorithm (SMP) which integrates over θ (like full Bayes) but maximizes over c (like MAP):

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

2.3 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 7, many implementations are essentially equivalent to the following algorithm (Quinlan & Rivest, 1989; Rissanen, 1989; Kearns et al., 1997; Grünwald, 1998):

$$c_{\text{MDL}(P,S)} = \arg\min_{c\in\mathcal{C}} \left\{ \log\frac{1}{P(c)} + \log\binom{m}{m\hat{e}_S(c)} \right\}. \tag{7}$$

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 Main theorems

We prove inconsistency for some countable set of classifiers $\mathcal{C} = \{c_0, c_1, \dots\}$ which we define later. The inconsistency is attained for priors with 'heavy tails'. Formally, for Theorem 2 (inconsistency of Bayes, MDL, MAP and SMP), we require $P(c_k)$ to be such that for all k,

$$\log \frac{1}{P(c_k)} \leq \log k + o(\log k). \tag{8}$$

This condition is satisfied, for example, by Rissanen's (1983) *universal prior for the integers*. Another simple prior satisfying (8) can be defined as follows: group the classifiers $c_1, c_2, \dots$ as $\mathcal{C}_0 := \{c_1\}, \mathcal{C}_1 := \{c_2, c_3\}, \mathcal{C}_2 := \{c_4, \dots, c_7\}$ and so on, so that $\mathcal{C}_a$ contains 2^a classifiers. Then the prior of any classifier in $\mathcal{C}_a$ is defined as

$$P(c) = \frac{1}{a(a+1)} 2^{-a}.$$

The sensitivity of our results to the choice of prior is analyzed further in Section 5.1. The prior on θ can be any distribution P on $[0, 1]$ with a density w that is continuously differentiable and bounded away from 0 on $[0, 0.5)$, i.e. for some $\gamma > 0$,

$$\text{for all } \theta \in [0, 0.5), w(\theta) > \gamma. \tag{9}$$

For example, we may take the uniform distribution on $[0, 1]$ with $w(\theta) \equiv 1$. We can also take the uniform distribution on $[0, 0.5)$, with $w(\theta) \equiv 2$.

For our result concerning the full Bayesian classifier, Theorem 2, Part (b), we need to make the further restriction[1]

$$P(\theta \geq 0.5) = 0. \tag{10}$$

For ease of comparison with other results (Section 6), we shall also allow discrete priors on θ that put all their mass on a countable set, $[0, 1] \cap \mathbb{Q}$. For such priors, we require that they satisfy, for all $a \in \mathbb{N}$, all $b \in \{0, 1, \dots, \lfloor a/2 \rfloor\}$:

$$P\left(\theta = \frac{b}{a}\right) \geq K_1 a^{-K_2} \tag{11}$$

for some fixed constants $K_1, K_2 > 0$. An example of a prior achieving (11) is $P(\theta = b/a) = 1/(a(a+1)\lfloor a/2 + 1 \rfloor)$.

[1] Without this restriction, we may put nonzero prior on distributions $p_{c,\theta}$ with $\theta > 1/2$. For such distributions, the log likelihood of the data increases rather than decreases as a linear function of the error that c makes on the data. This implies that with a uniform prior, under our definition of the Bayes MAP classifier, in some cases the MAP classifier may be the classifier with the largest, rather than the smallest empirical error. As pointed out by a referee, for this reason the term "Bayes MAP classifier" may be somewhat of a misnomer: it does not always coincide with the Bayes act corresponding to the MAP distribution. If one insists on defining the Bayes MAP classifier as the Bayes act corresponding to the MAP distribution, then one can achieve this simply by restricting oneself to priors satisfying (10), since all our results still hold under the restriction (10).

We assume that the priors $P(\theta)$ on $[0,1]$ and the prior $P(c)$ on $\mathcal{C}$ are fully dependent so that every classifier can have its own error rate. We require each classifier to have the same prior for θ. Thus, $P(c,\theta) = P(c)P(\theta|c)$ where for every c, $P(\theta|c)$ is set to $P(\theta)$. Note that given a sample S, the posterior $P(\theta\,|\,c,S)$ can depend on c. In the theorem, $H(\mu) = \mu \log \frac{1}{\mu} + (1-\mu)\log\frac{1}{1-\mu}$ stands for the binary entropy of a coin with bias μ. The function $g(\mu)$ appearing in Part (b) of the theorem is defined as

$$g(\mu) = \mu + \sup\{\nu\,|\,\nu \geq 0\ ;\ H(\nu) < H(\mu) - 2\mu\}. \tag{12}$$

This function will be analyzed later.

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* (8) *and* (9)*. Then, for all $\mu \in (0, 0.5)$,*

(a) *For all $\mu' \in [\mu, H(\mu)/2)$, there exists a D with* $\min_{c\in\mathcal{C}} e_D(c) = e_D(c_0) = \mu$ *such that, for all large m, with $a_m = 3\exp(-2\sqrt{m})$,*

$$\Pr_{S\sim D^m}(e_D(c_{\mathrm{MAP}(P,S)}) = \mu') \geq 1 - a_m$$

$$\Pr_{S\sim D^m}(e_D(c_{\mathrm{SMP}(P,S)}) = \mu') \geq 1 - a_m$$

$$\Pr_{S\sim D^m}(e_D(c_{\mathrm{MDL}(P,S)}) = \mu') \geq 1 - a_m. \tag{13}$$

(b) *If the prior P further satisfies* (10)*, then for all $\mu' \in [\mu, g(\mu))$, there exists a D with* $\min_{c\in\mathcal{C}} e_D(c) = e_D(c_0) = \mu$ *such that, for all large m, with $a_m = (6 + o(1))\exp(-(1 - 2\mu)\sqrt{m})$,*

$$\Pr_{S\sim D^m}(e_D(c_{\mathrm{BAYES}(P,S)}) \geq \mu') \geq 1 - a_m. \tag{14}$$

Since $H(\mu)/2 > \mu$ for all $\mu \in (0, 0.5)$ (Fig. 1), inconsistency of $c_{\mathrm{MAP}(P,S)}$, $c_{\mathrm{SMP}(P,S)}$ and $c_{\mathrm{MDL}(P,S)}$ can occur for all $\mu \in (0, 0.5)$. 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".[2] Differentiation of $0.5H(\mu) - \mu$ shows that the maximum discrepancy between $e_D(c_{\mathrm{MAP}(P,S)})$ 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 Fig. 1. Concerning $c_{\mathrm{BAYES}(P,S)}$, since $H(\mu) - 2\mu > 0$ for all $\mu \in (0, 0.5)$, $H(0) = 0$ and $H(\nu)$ is monotonically increasing between 0 and 0.5, we have $g(\mu) > \mu$ for all $\mu \in (0, 0.5)$. Hence, inconsistency can occur for all such μ. Since $H(\nu)$ is monotonically increasing in ν, the largest value of ν can be obtained for the μ for which $H(\mu) - 2\mu$ is largest. We already noted that this is maximized for $\mu = 0.2$. Then the largest ν such that $H(\nu) < H(\mu) - 2\mu = 2 \cdot 0.1609...$ can be numerically calculated as $\nu_{\max} = 0.0586....$ Thus, in that case we can get $e_D(c_{\mathrm{BAYES}(P,S)})$ arbitrarily close to $\mu + \nu_{\max} = 0.2586....$[3]

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

[3] While we have no formal proof, we strongly suspect that $g(\mu)$ can be replaced by $H(\mu)/2$ in Part 2 of the theorem, so that the suboptimality for $c_{\mathrm{BAYES}(P,S)}$ would be just as large as for the other three classifiers. For this reason we did not bother to draw the function $g(\mu)$ in Fig. 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? The next theorem, again summarized in Fig. 1, gives an upper bound which holds for all learning problems (distributions D), 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 $0 < \inf_{c\in\mathcal{C}} e_D(c) = \mu < 0.5$, for all $\delta > 0$, for all large m, with D^m-probability $\geq 1 - 2\exp(-2\sqrt{m})$,*

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

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 $e_D(c_{\text{BAYES}(P,S^{i-1})}(x_i)) \leq H(\mu)$ for *all* large i.

4 Proofs

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.

4.1.1 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 $e_D(c_j) = \mu' \geq \mu$. $x = x_0 x_1 \ldots \in \mathcal{X} = \{0, 1\}^\infty$ consists of one binary feature per classifier, and the classifiers simply output the value of their special feature. The underlying distribution D depends on two parameters $0 < p_{\text{h}} < 1$ and $\eta \in [0, 1/2)$. These are defined in terms of the μ and μ' mentioned in the theorem, in a way to be described later.

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_{h} which determines if this is a "hard" example or an "easy" example. Based upon these two coin flips, for $j \geq 1$, each x_j is independently generated according to the following 2 cases.

1. For a "hard" example, and for each classifier c_j with $j \geq 1$, set $x_j = |1 - y|$ with probability $1/2$ and $x_j = y$ otherwise. Thus, $x_1, x_2, \ldots$ becomes an infinite sequence of realizations of i.i.d. uniform Bernoulli random variables.
2. For an "easy" example, we flip a third coin Z with bias η. If $Z = 0$ ('example ok'), we set, for every $j \geq 1$, $x_j = y$. If $Z = 1$, we set, for all $j \geq 1$, $x_j = |1 - y|$ otherwise. Note that for an "easy" example, all bad classifiers make the same prediction.

The bad classifiers essentially predict Y by random guessing on hard examples. On easy examples, they all make the same prediction, which is correct with probability $1 - \eta > 0.5$.

It remains to define the input x_0 of the "good" classifier c_0 with true error rate μ. This is done simply by setting $x_0 = |1 - y|$ with probability μ and $x_0 = y$ otherwise.

The setting of p_{h} and η is different for, on the one hand, the (S)MAP and MDL inconsistency proofs, and on the other hand, the full Bayes inconsistency proof. To get a first, intuitive idea of the proof, it is best to ignore, for the time being, the parameter values for the full Bayes proof.

In the MAP and MDL inconsistency proof, we set $p_{\mathrm{h}} := 2\mu'$ and $\eta = 0$. In the Bayes proof, we first set $p_{\mathrm{h}} := 2\mu$. We then define $\nu := \mu' - \mu$ and we set $\eta := \nu/(1 - 2\mu)$. By the conditions of the theorem, we have $0 < H(\nu) < H(\mu) - 2\mu$. Note that for such ν, $H(\nu) < 1 - 2\mu$ and therefore $2\nu \leq 1 - 2\mu$ so that $\eta < 1/2$. As is easily checked in the (S)MAP and MDL case, and somewhat harder to check in the full Bayes case, the error rate of each 'bad' classifier is $e_D(c_j) = \mu'$ for all $j \geq 1$.

Discussion The inputs x are infinite-dimensional vectors. Nevertheless, the Bayesian posterior can be arbitrarily well approximated in finite time for any finite sample size 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 $m + O(\log m)$ bits. This follows because by (9) and (11), the prior $P(\theta)$ puts sufficient mass in a neighborhood of $\theta = 0.5$. For such θ, the distribution $p_{c,\theta}(y|x)$ becomes uniform, independently of c.

The (constructive) proof of Theorem 2 relies upon this problem, but it's worth mentioning that other hard learning problems exist and that this hard learning problem can be viewed in two other ways:

1. The input space has two bits (the hardness bit and the "correct value" bit) and the classifiers are stochastic. Stochastic classifiers might be reasonable (for example) if the task is inferring which of several stock "experts" are the best on average. The stock pickers occasionally make mistakes as modeled by the stochasticity.
2. The input space consists of one real valued feature. Each bit in the binary expansion of the real number is used by a different classifier as above.

4.1.2 Bayes and MDL are inconsistent

We now prove Theorem 2. Stage 1 and 2 do not depend on the specific choices for p_{h} and η, and are common to the proofs for MAP, SMP, MDL and full Bayes. In Stage 1 we show that for some function $k(m)$, for every value of m, with probability converging to 1, there exists some 'bad' classifier $c^* = c_j$ with $0 < j \leq k(m)$ that has 0 empirical error on hard examples, whereas the good classifier has empirical error close to its expected generalization error. Up to sublinear terms, we find that

$$\log k(m) = m p_{\mathrm{h}}. \tag{15}$$

The precise expression is given in (18). In Stage 2 we rewrite the log posterior odds ratio between the good classifier c_0 and c^*. Up to sublinear terms (see (26)), this ratio turns out to be equal to

$$m H(\hat{e}_S(c^*)) - m H(\hat{e}_S(c_0)) + m p_{\mathrm{h}}. \tag{16}$$

In Stage 3 we combine (15) and (16) to show that, with the choice $p_{\mathrm{h}} = 2\mu'$, $\eta = 0$, the posterior on c^* becomes exponentially larger than the posterior on c_0, from which inconsistency of MAP, SMP and MDL readily follows. In Stage 4, we show that with the choice $p_{\mathrm{h}} = 2\mu$,

$\eta = \nu/(1-2\mu)$, the posterior on c^* still becomes exponentially larger than the posterior on c_0, but now additionally, the classification performance of the Bayesian classifier (a mixture that puts nearly all its weight on bad classifiers), cannot exceed that of c^*.

Stage 1. Let m_{h} denote the number of hard examples generated within a sample S of size m. Let $\hat{e}_{S,\mathrm{h}}(c)$ be the number of mistakes that the classifier c makes on the subset S_{h} of S of hard examples, divided by $m_{\mathrm{h}} = |S_{\mathrm{h}}|$. Let k be a positive integer and $\mathcal{C}_k = \{c_j \in \mathcal{C} : 1 \le j \le k\}$. For all $\epsilon > 0$ and $m \ge 0$, we have:

$$\begin{aligned}
&\Pr_{S\sim D^m}(\forall c \in \mathcal{C}_k : \ \hat{e}_{S,\mathrm{h}}(c) > 0)\\
&\overset{(a)}{=} \Pr_{S\sim D^m}\left(\forall c \in \mathcal{C}_k : \ \hat{e}_{S,\mathrm{h}}(c) > 0 \,\middle|\, \frac{m_{\mathrm{h}}}{m} > p_{\mathrm{h}} + \epsilon\right) \Pr_{S\sim D^m}\left(\frac{m_{\mathrm{h}}}{m} > p_{\mathrm{h}} + \epsilon\right)\\
&\quad + \Pr_{S\sim D^m}\left(\forall c \in \mathcal{C}_k : \ \hat{e}_{S,\mathrm{h}}(c) > 0 \,\middle|\, \frac{m_{\mathrm{h}}}{m} \le p_{\mathrm{h}} + \epsilon\right) \Pr_{S\sim D^m}\left(\frac{m_{\mathrm{h}}}{m} \le p_{\mathrm{h}} + \epsilon\right)\\
&\overset{(b)}{\le} e^{-2m\epsilon^2} + \Pr_{S\sim D^m}\left(\forall c \in \mathcal{C}_k : \ \hat{e}_{S,\mathrm{h}}(c) > 0 \,\middle|\, \frac{m_{\mathrm{h}}}{m} \le p_{\mathrm{h}} + \epsilon\right)\\
&\overset{(c)}{\le} e^{-2m\epsilon^2} + (1 - 2^{-m(p_{\mathrm{h}}+\epsilon)})^k \overset{(d)}{\le} e^{-2m\epsilon^2} + e^{-k2^{-m(p_{\mathrm{h}}+\epsilon)}}.
\end{aligned} \tag{17}$$

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

$$k = k(m) = \frac{2m\epsilon_m^2}{2^{-m(p_{\mathrm{h}}+\epsilon_m)}}. \tag{18}$$

Note that, up to sublinear terms, this is equal to (15). With (18), (17) becomes

$$\Pr_{S\sim D^m}(\forall c \in \mathcal{C}_{k(m)} : \ \hat{e}_{S,\mathrm{h}}(c) > 0) \le 2e^{-2\sqrt{m}} \tag{19}$$

On the other hand, by the Chernoff bound we have $\Pr_{S\sim D^m}(\hat{e}_S(c_0) < e_D(c_0) - \epsilon_m) \le e^{-2\sqrt{m}}$ for the optimal classifier c_0. Combining this with (19) using the union bound, we get that, with D^m-probability larger than $1 - 3e^{-2\sqrt{m}}$, the following event holds:

$$\exists c \in \mathcal{C}_{k(m)} : \ \hat{e}_{S,\mathrm{h}}(c) = 0 \text{ and } \ \hat{e}_S(c_0) \ge e_D(c_0) - \epsilon_m. \tag{20}$$

Stage 2. In this stage, we calculate, for large m, the log ratio between the posterior on some $c^* \in \mathcal{C}_{k(m)}$ with $\hat{e}_{S,\mathrm{h}}(c^*) = 0$ and the posterior on c_0. We have:

$$\begin{aligned}
&\log \frac{\max_\theta P(c_0, \theta \,|\, x^m, y^m)}{\max_\theta P(c^*, \theta \,|\, x^m, y^m)} = \log \frac{\max_\theta P(c_0)P(\theta)P(y^m \,|\, x^m, c_0, \theta)}{\max_\theta P(c^*)P(\theta)P(y^m \,|\, x^m, c^*, \theta)}\\
&\quad = \log \max_\theta P(c_0)P(\theta)P(y^m \,|\, x^m, c_0, \theta) - \log \max_\theta P(c^*)P(\theta)P(y^m \,|\, x^m, c^*, \theta).
\end{aligned} \tag{21}$$

Using (4), (9) and (11), we see that, uniformly for all samples S with $\hat{e}_S(c_0) < 1/2$, the leftmost term is no larger than

$$\log\big(\max_\theta P(c_0)P(\theta)\big) \cdot \big(\max_{\theta'} P(y^m \,|\, x^m, c_0, \theta')\big) = -mH(\hat{e}_S(c_0)) + O(1). \tag{22}$$

Similarly, uniformly for all samples S with $\hat{e}_{S,\mathrm{h}}(c^*) = 0$, $\hat{e}_S(c^*) < 1/2$, the rightmost term in (21) satisfies

$$-\log \max_{\theta} P(c^*)P(\theta)P(y^m \,|\, x^m, c^*, \theta) \leq -\log P(c^*) + mH(\hat{e}_S(c^*)) + O(1), \tag{23}$$

where the constant in the O-notation does not depend on c^*. Using condition (8) on prior $P(c^*)$ and using $c^* \in \mathcal{C}_{k(m)}$, we find:

$$-\log P(c^*) = \log \frac{1}{P(c^*)} \leq \log k(m) + o(\log k(m)), \tag{24}$$

where $\log k(m) = \log 2\sqrt{m} + mp_{\mathrm{h}} + m^{0.75}$, so that

$$\log \frac{1}{P(c^*)} \leq mp_{\mathrm{h}} + o(m) \tag{25}$$

which implies that (23), is no larger than $mp_{\mathrm{h}} + mH(\hat{e}_S(c^*)) + o(m)$. Thus, for all large m, the difference between the leftmost term and the rightmost term in (21) satisfies

$$\log \frac{\max_\theta P(c_0, \theta \,|\, x^m, y^m)}{\max_\theta P(c^*, \theta \,|\, x^m, y^m)} \leq -mH(\hat{e}_S(c_0)) + mp_{\mathrm{h}} + mH(\hat{e}_S(c^*)) + o(m), \tag{26}$$

as long as $\hat{e}_S(c_0)$ and $\hat{e}_S(c^*)$ are both less than 0.5.

Stage 3(a). (**MAP**) Recall that, for the MAP result, we set $\eta := 0$ and $p_{\mathrm{h}} := 2\mu'$. Let us assume that the large probability event (20) holds. This will allow us to replace the two 'empirical entropies' in (26), which are random variables, by corresponding ordinary entropies, which are constants. By (20), $\hat{e}_{S,\mathrm{h}}(c^*) = 0$, so that we have (since $\eta = 0$) that $\hat{e}_S(c^*) = 0$ and then also $H(\hat{e}_S(c^*)) = 0$. Because $H(\mu)$ is continuously differentiable in a small enough neighborhood around μ, by (20) we also have, for some constant K,

$$H(\hat{e}_S(c_0)) \geq H(e_D(c_0)) - K\epsilon_m + O(1) = H(\mu) + O(m^{-1/2}).$$

Plugging these expressions for $H(\hat{e}_S(c^*))$ and $H(\hat{e}_S(c_0))$ into (26), and using the fact that we set $p_{\mathrm{h}} = 2\mu'$, we see that, as long as $\mu' < H(\mu)/2$, there exists a $c > 0$ such that for all large m, (26), and hence (21) is smaller than $-cm$. Thus, (21) is less than 0 for large m, implying that then $e_D(c_{\mathrm{MAP}(P,S)}) = \mu'$. We derived all this from (20) which holds with probability $\geq 1 - 3\exp(-2\sqrt{m})$. Thus, for all large m, $\Pr_{S \sim D^m}\left(e_D(c_{\mathrm{MAP}(P,S)}) = \mu'\right) \geq 1 - 3\exp(-2\sqrt{m})$, and the result follows.

Stage 3(b). (**SMP**) We are now interested in evaluating, instead of the posterior ratio (21), the posterior ratio with the error rate parameters integrated out:

$$\log \frac{P(c_0 \,|\, x^m, y^m)}{P(c^* \,|\, x^m, y^m)} = \log P(c_0)P(y^m \,|\, x^m, c_0) - \log P(c^*)P(y^m \,|\, x^m, c^*). \tag{27}$$

By Proposition 2 in the appendix, we see that, if (20) holds, then (21) is no larger than (27) plus an additional term of order $O(\log m)$. To see this, apply the first inequality of (54) to the term involving c_0, and the second inequality of (54) to the term involving c^*. The result now follows by exactly the same reasoning as in Stage 3(a).

Stage 3(c). (**MDL**) By part (1) of Proposition 2 in the appendix, the MDL procedure is equal to SMP with the uniform prior $w(\theta) \equiv 1$. Thus, the MDL case is a special case of the SMP case for which we already proved inconsistency above.

Stage 3(d). (**Bayes**) In order to prove the inconsistency for the full Bayesian classifier, we construct a setup where on on hard examples, all classifiers, even the 'good' classifier c_0, predict $c(X) = 1$ with probability 1/2, independently of the true value of Y. To this end, we refine our learning problem by setting $x_0 = |1 - y|$ with probability 1/2 for a "hard" example, and $x_0 = y$ with probability 1 for an "easy" example. By setting $p_{\mathrm{h}} := 2\mu$, we still get that $e_D(c_0) = \mu$. In order to make the error rate for the bad classifiers $c_1, c_2, \ldots$ still larger than for c_0, we now set η to a value larger strictly than 0.

We let $\hat{e}_{S,\mathrm{easy}}(c)$ denote the empirical error that classifier c achieves on the easy examples in S, i.e. the number of mistakes on the easy examples in S divided by $|S| - |S_{\mathrm{h}}|$. Now set ϵ_m as in Stage 1 and define the events (sets of samples S^m of length m) $\mathcal{A}$ and $\mathcal{B}$ as

$$\mathcal{A} = \{S^m \; : \; \exists j > 0 : |\hat{e}_{S,\mathrm{easy}}(c_j) - \eta| > \epsilon_m\}; \quad \mathcal{B} = \left\{ S^m \; : \; \frac{m_{\mathrm{h}}}{m} \geq p_{\mathrm{h}} + \epsilon_m \right\},$$

and let $\mathcal{A}^c$ and $\mathcal{B}^c$ be their respective complements. We have

$$\begin{aligned}\Pr_{S\sim D^m}(\mathcal{A}\cup\mathcal{B}) &= \Pr(\mathcal{A}\cup\mathcal{B}\,|\,\mathcal{B})\Pr(\mathcal{B}) + \Pr(\mathcal{A}\cup\mathcal{B}\,|\,\mathcal{B}^c)\Pr(\mathcal{B}^c) \leq \Pr(\mathcal{B}) + \Pr(\mathcal{A}\,|\,\mathcal{B}^c)\\ &\leq e^{-2\sqrt{m}} + 2e^{-2m(1-p_{\mathrm{h}}-\epsilon_m)\epsilon_m^2} = e^{-2\sqrt{m}} + 2e^{-2\sqrt{m}(1-p_{\mathrm{h}})+2m^{0.25}}\\ &\leq 3e^{-(1-p_{\mathrm{h}})\sqrt{m}}(1+o(1)), \quad (28)\end{aligned}$$

where the second inequality follows by applying the Chernoff bound to both terms (recall that on easy examples, all classifiers c_j with $j > 1$ output the same prediction for Y). Combining this with the result of Stage 1 using the union bound and using $p_{\mathrm{h}} = 2\mu$, we get that, with D^m-probability at least $1 - 6(1 + o(1))e^{-\sqrt{m}(1-2\mu)}$, (20) and $\mathcal{A}^c$ and $\mathcal{B}^c$ all hold at the same time. Let us assume for now that this large probability event holds. We must then have that some $c^* \in \mathcal{C}_{k(m)}$ achieves empirical error 0 on hard examples (which occur with probability 2μ) and at least $\eta + O(m^{-1/2})$ on easy examples (which occur with probability $1 - 2\mu$), so that

$$\hat{e}_S(c^*) = (1 - 2\mu)\eta + O(m^{-1/2}) = \nu + O(m^{-1/2}), \quad (29)$$

where $\nu = \mu' - \mu$. By continuity of H, we also have that $H(\hat{e}_S(c^*)) = H(\nu) + O(m^{-1/2})$.

Entirely analogously to the reasoning in Stage 3(a), we can now replace the empirical entropies in the expression (26) for the log-likelihood ratio between c^* and c_0 by the corresponding ordinary entropies. This gives

$$\log \frac{\max_\theta P(c_0, \theta\,|\,x^m, y^m)}{\max_\theta P(c^*, \theta\,|\,x^m, y^m)} \leq -mH(\mu) + m2\mu + mH(\nu) + o(m), \quad (30)$$

By definition of ν, $\nu = \mu' - \mu$ satisfies $-H(\mu) + H(\nu) + 2\mu < 0$, so that there exists a $c > 0$ such that for all large m, (30) is smaller than $-cm$. Reasoning entirely analogously to

Stage 3(b), we see that (30) still holds if we integrate out θ, rather than maximize over it: there exists a $c > 0$ such that for all large m,

$$\log \frac{P(c_0 \,|\, x^m, y^m)}{P(c^* \,|\, x^m, y^m)} \leq -mH(\mu) + m2\mu + mH(\nu) + o(m) \leq -cm. \tag{31}$$

Furthermore, by (29) and our condition on the prior, the posterior on θ given c^* must concentrate on ν (even though c^* varies with m): we must have that, for every open set A containing ν, the posterior distribution of θ given c^* and sample S satisfies

$$P(\theta \in A \,|\, c^*, S) \overset{m\to\infty}{\to} 1. \tag{32}$$

We now show that (31) and (32), both of which hold with high probability, imply that the full Bayesian classifier based on sample S errs with probability at least $\mu + \nu = \mu'$:

$$\begin{aligned}
&\Pr_{X,Y\sim D}(Y \neq c_{\text{BAYES}(P,S)}(X)) \\
&\quad = \Pr_{X,Y\sim D}(Y \neq c_{\text{BAYES}(P,S)}(X) \,|\, \text{Ex. hard}) p_{\text{h}} \\
&\qquad + \Pr_{X,Y\sim D}(Y \neq c_{\text{BAYES}(P,S)}(X) \,|\, \text{Ex. easy})(1 - p_{\text{h}}) \\
&\quad \geq \frac{1}{2} 2\mu + \Pr_{X,Y\sim D}(Y \neq c_{\text{BAYES}(P,S)}(X) \,|\, \text{Ex. easy})(1 - 2\mu) \\
&\quad \geq \mu + \Pr_{X,Y\sim D}(Y \neq c_{\text{BAYES}(P,S)}(X) \,|\, \text{Ex. easy, corrupted})(1 - 2\mu)\eta \\
&\quad = \mu + \Pr_{X,Y\sim D}(Y \neq c_{\text{BAYES}(P,S)}(X) \,|\, \text{Ex. easy, corrupted}, Y = 1)(1 - 2\mu)\eta.
\end{aligned} \tag{33}$$

Here the first inequality follows by symmetry: on hard examples, $Y = 1$ with probability $1/2$ and all classifiers independently output $Y = 1$ with probability $1/2$. The final equality follows again by symmetry between the case $Y = 1$ and $Y = 0$. Depending on the sample S, the probability in the final line of (33) is either equal to 1 or to 0. It is equal to 1 if $c_{\text{BAYES}(P,S)}(X) = 0$. By (6), a sufficient condition for this to happen is if S is such that

$$E_{c,\theta\sim P(\cdot|S)}[p_{c,\theta}(Y = 1|X = x)] < \frac{1}{2}. \tag{34}$$

This expectation can be rewritten as

$$\begin{aligned}
&\sum_{c\in\mathcal{C}} \int_{\theta\in[0,0.5)} P(\theta \,|\, c, S) P(c \,|\, S) p_{c,\theta}(Y = 1 \,|\, X = x) d\theta \\
&\quad = \sum_{c\in\mathcal{C}} P(c \,|\, S) u(c, S, x) \\
&\quad = P(c_0 \,|\, S) u(c_0, S, x) + P(c^* \,|\, S) u(c^*, S, x) + \sum_{c\notin\{c_0,c^*\}} P(c \,|\, S) u(c, S, x),
\end{aligned} \tag{35}$$

where $u(c, S, x) := \int_{\theta\in[0,0.5)} P(\theta \,|\, c, S) p_{c,\theta}(Y = 1|X = x) d\theta$. Note that we integrate here over [0, 0.5), reflecting the extra condition (10) that we required for the full Bayesian result. Since the example in the final line (33) is corrupted, for the x occurring there we have that $c_j(x) = 0$ for $j \geq 1$, so that $p_{c_j,\theta}(Y = 1 \,|\, X = x) < 1/2$ for all $\theta < 1/2$. It follows that for this x, (35) is no greater than

$$P(c_0 \mid S) + P(c^* \mid S) u(c^*, S, x) + \sum_{c \notin \{c_0, c^*\}} P(c \mid S) \frac{1}{2}.$$

By (31) and (32), for all $\delta > 0$, for all large m this is no greater than

$$a(e^{-cm} \cdot 1 + (1 - e^{-cm})(\nu + \delta)) + (1 - a)\frac{1}{2}. \tag{36}$$

for some a that may depend on m but that satisfies $0 < a < 1$ for all m. Therefore, and since $\nu < 1/2$, (36) is less than $1/2$ for all large m. But this implies (by the reasoning above (34)) that $c_{\mathrm{BAYES}(P,S)}(x) = 0$. It follows by (33) that, for large m,

$$\Pr_{X,Y \sim D}(Y \neq c_{\mathrm{BAYES}(P,S)}(X)) \geq \mu + (1 - 2\mu)\eta = \mu + \nu = \mu'.$$

All this is implied under an event that holds with probability at least $1 - 6(1 + o(1))e^{-\sqrt{m}(1-2\mu)}$ (see above), so that the result follows.

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). (Blumer et al., 1987) *For all priors P on a countable set of classifiers $\mathcal{C}$, for all distributions D, with probability* $1 - \delta$*:*

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

The algorithm stated here is in a suboptimal form, which is good enough for our purposes (see (McAllester, 1999) for more sophisticated versions):

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

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 $c \in \mathcal{C} = \{c_0, c_1, \dots\}$. 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 $1 - \delta_m = 1 - 1/m$,

$$e_D(\tilde{c}_m) \geq \hat{e}_S(\tilde{c}_m) - \sqrt{\frac{\ln(1/\delta_m)}{2m}} = \hat{e}_S(\tilde{c}_m) - \sqrt{\frac{\ln m}{2m}}, \tag{37}$$

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

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

Combining this with (37) 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 Corollary 1

The corollary relies on Theorem 1 and a slight generalization of the proof of Theorem 2. For Theorem 1 pick $K < 0.05$. In Theorem 2 choose $\mu = 1/5$, $\mu' = 1/5 + .15$. Now part 1 follows. For part 2, consider Theorem 2 with the same μ and μ'. From the theorem we see that for the learning problem for which (13) holds, c_0 is the optimal classifier. Denote this learning problem by D_0. We define D_j as the learning problem (distribution) in the proof (see Section 4.1.1), but with the role of x_0 and x_j interchanged. As a result, c_j will be the 'good' classifier with error rate μ and c_0 will be one of the bad classifiers with rate μ'. Then the good classifier and one of the bad classifiers will have a different prior probability, but otherwise nothing changes. Since the proof of Theorem 2 does not depend on the prior probability of the good classifier—it can be as large or small as we like as long as it is greater than 0 –, the proof goes through unchanged, and for all learning problems D_j, (13) will hold.

We now generate a learning problem D_j by first sampling a classifier c_j according to $P(c)$, and then generating data according to D_j. Then, no matter what c_j we chose, it will be the optimal ('good') classifier, and, as we just showed, (13) will hold. Theorem 1 (with $K < 0.05$) can now be applied with $D = D_j$, and the result follows.

4.4 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 $P(y_i \mid x_i, S^{i-1}) = E_{c,\theta \sim P(\cdot|S^{i-1})} p_{c,\theta}(y_i|x_i)$. 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, if for some $\alpha > 0$, $\hat{e}_S(c_0) < 0.5 - \alpha$, then

$$\sum_{i=1}^{m} -\log P(y_i \mid x_i, S^{i-1}) \geq \sum_{i=1}^{m} |y_i - c_{\mathrm{BAYES}(P,S^{i-1})}(x_i)|. \tag{38}$$

On the other hand we have

$$
\begin{aligned}
\sum_{i=1}^{m} - \log P(y_i \,|\, x_i, S^{i-1}) &= - \log \prod_{i=1}^{m} P(y_i \,|\, x_i, x^{i-1}, y^{i-1}) \\
&= - \log \prod_{i=1}^{m} P(y_i \,|\, x^m, y^{i-1}) \\
&= - \log \prod_{i=1}^{m} \frac{P(y^i | x^m)}{P(y^{i-1} | x^m)} \\
&= - \log P(y^m \,|\, x^m) \\
&= - \log \sum_{j=0,1,2\ldots} P(y^m \,|\, x^m, c_j) P(c_j) \\
&\overset{(a)}{\leq} - \log P(y^m \,|\, x^m, c_0) - \log P(c_0) \overset{(b)}{\leq} m H(\hat{e}_S(c_0)) + O(\log m),
\end{aligned}
\tag{39}
$$

where the constant in the $O(\log m)$ term may depend on α. Here inequality (a) follows because a sum is larger than each of its terms, and (b) follows by Proposition 2 in the appendix. 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 $\epsilon_m = m^{-0.25}$. Using the fact that $H(\mu)$ in (39) is continuously differentiable in a neighborhood of μ and $\mu < 1/2$, it follows that with probability larger than $1 - 2\exp(-2\sqrt{m})$, for all large m,

$$
\sum_{i=1}^{m} - \log P(y_i \,|\, x_i, S^{i-1}) \leq m H(e_D(c_0)) + K m^{0.75} + O(\log m), \tag{40}
$$

where K is a constant not depending on m. Combining (40) with (38) we find that with probability $\geq 1 - 2\exp(-2\sqrt{m})$, $\sum_{i=1}^{m} |y_i - c_{\mathrm{BAYES}(P,S^{i-1})}(x_i)| \leq m H(e_D(c_0)) + o(m)$, which is what we had to prove.

5 Technical discussion

5.1 Variations of Theorem 2 and dependency on the prior

Prior on classifiers. The requirement (8) that $-\log P(c_k) \geq \log k + o(\log k)$ is needed to obtain (25), 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

$$
- \log P(c_k) = d \log k + o(\log k), \tag{41}
$$

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 (41). 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 this argument.

Prior on error rates. Condition (9) 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 7). 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 (Cover & Thomas, 1991) between $p_{c,\theta}$ and the 'true' conditional distribution p_D:

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

and note that 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

$$\min_{\theta} \Delta(p_{c,\theta}) = \Delta(p_{c,e_D(c)}) = H(e_D(c)) - K_D, \tag{42}$$

where $K_D = E[-\log p_D(y|x)]$ does not depend on c or θ, and $H(\mu)$ is the binary entropy.

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

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

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

Proof: Property 1 follows from (42) and the fact that $H(\mu)$ is monotonically increasing for $\mu < 1/2$. Property 2 follows directly from the information inequality (Cover & Thomas, 1991), using the fact that we assume $p_D(y|x)$ to be well-defined for all x, which implies that X has a density $p_D(x)$ with $p_D(x) > 0$ for all x. □

Proposition 1 implies that the 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 $\alpha \in A$, $-m^{-1} \log p_\alpha(y^m \mid x^m)P(\alpha) - K_D \rightarrow \mathrm{KL}(p_D \| p_\alpha)$. 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 this 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 (Grünwald, 1998), 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 (Grünwald, 1998). We can only get trouble as in Theorem 2 if we allow $\mathcal{C}$ to be of infinite VC-dimension.

Analogy to Regression. In ordinary (real-valued) regression, $\mathcal{Y} = \mathbb{R}$, and one tries to learn a function $f \in \mathcal{F}$ from the data. Here $\mathcal{F}$ is a set of candidate functions $\mathcal{X} \rightarrow \mathcal{Y}$. In order to apply Bayesian inference to this problem, one assumes a probability model $\mathcal{P}$ expressing $Y = f(X) + Z$, where Z is independent noise with mean 0 and variance σ^2. $\mathcal{P}$ then consists of conditional density functions p_{f,σ^2}, one for each $f \in \mathcal{F}$ and $\sigma^2 > 0$. It is well known that if one assumes Z to be normally distributed independently of X, then the p_{f,σ^2} become Gaussian densities and the log likelihood becomes *a linear function of the mean squared error* (Rissanen, 1989):

$$- \ln p_{f,\sigma^2}(y^n \mid x^n) = \beta_\sigma \sum_{i=1}^{n} (y_i - f(x_i))^2 + n \ln Z(\beta_\sigma). \tag{43}$$

where we wrote $\beta_\sigma = 1/2\sigma^2$ and $Z(\beta) = \int_{y \in \mathcal{Y}} \exp(-\beta y^2) dy$. Because least squares is an intuitive, mathematically well-behaved and easy to perform procedure, it is often assumed in Bayesian regression that the noise is normally distributed—even in cases where in reality, it is not (Grünwald, 1998; Kleijn & van der Vaart, 2004).

Completely analogously to the Gaussian case, the transformation used in this paper maps classifiers c and noise rates θ to distributions $p_{c,\theta}$ so that the log likelihood becomes a *linear function of the* $0/1$*-error*, since it can be written as:

$$- \ln p_{c,\theta}(y^n \mid x^n) = \beta_\theta \sum_{i=1}^{n} |y_i - c(x_i)| + n \ln Z(\beta_\theta). \tag{44}$$

where we wrote $\beta_\theta = \ln(1 - \theta) - \ln \theta$ and $Z(\beta) = \sum_{y \in \mathcal{Y}} \exp(-\beta y)$ (Grünwald, 1998; Meir & Merhav, 1995). Indeed, the models $\{p_{c,\theta}\}$ are a special case of *logistic regression models*, which we now define:

Logistic regression interpretation. let $\mathcal{C}$ be a set of functions $\mathcal{X} \rightarrow \mathcal{Y}$, 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

$$p_{c,\beta}(y \neq x \mid x) := \frac{e^{-\beta |y - c(x)|}}{1 + e^{-\beta}} \tag{45}$$

This is the standard construction used to convert classifiers with real-valued output such as support vector machines and neural networks into conditional distributions (Jordan, 1995; Tipping, 2001), 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 (44), we see that the construction is a special case of the logistic regression transformation (45). It may seem that (45) does not treat $y = 1$ and $y = 0$ on equal footing, but this is not so: we can alternatively define a symmetric version of (45) by defining, for each $c \in \mathcal{C}$, a corresponding $c' : \mathcal{X} \rightarrow \{-1, 1\}$, $c'(x) := 2c(x) - 1$. Then we can set

$$p_{c,\beta}(1 \mid x) := \frac{e^{\beta c(x)}}{e^{\beta c(x)} + e^{-\beta c(x)}}; \quad p_{c,\beta}(-1 \mid x) := \frac{e^{-\beta c(x)}}{e^{\beta c(x)} + e^{-\beta c(x)}}. \tag{46}$$

By setting $\beta = 2\beta'$ we see that $p_{c,\beta}$ as in (45) is identical to $p_{c,\beta'}$ as in (46), so that the two models really coincide.

6 Interpretation from a Bayesian perspective

We already addressed several questions concerning the relevance of our result directly below Corollary 1. Here we provide a more in-depth analysis from a Bayesian point of view.

6.1 Bayesian consistency

It is well-known that Bayesian inference is strongly consistent under very broad conditions (Doob, 1949; Blackwell & Dubins, 1962); see also (Barron, 1985). Such Bayesian consistency results take on a particularly strong form if the set of distributions under consideration is *countable*. In our setting we can achieve this by adopting a discrete prior satisfying (11). In that case, the celebrated (Doob, 1949) consistency theorem[4] says the following for our setting. Let $\mathcal{C}$ be countable and suppose D is such that, for some $c^* \in \mathcal{C}$ and $\theta^* \in [0, 1] \cap \mathbb{Q}$, 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, Doob's theorem 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 premise $\Delta(p_{c_0,e_D(c_0)}) = 0$ 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* $\theta \in [0, 1] \cap \mathbb{Q}$, $p_{c,\theta}$ is equal to $p_D(\cdot|\cdot)$—in statistical terms, the model $\mathcal{P} = \{p_{c,\theta} : c \in \mathcal{C}, \theta \in [0, 1] \cap \mathbb{Q}\}$ is *misspecified*. Thus, the result can be interpreted in two ways:

1. *'ordinary' Bayesian inference can be inconsistent under misspecification*: We exhibit a simple logistic regression model $\mathcal{P}$ and a true distribution D such that, with probability 1, the Bayesian posterior does not converge to the distribution $p_{c_0,e_D(c_0)} \in \mathcal{P}$ that minimizes, among all $p \in \mathcal{P}$, the KL-divergence to D (equivalently, $p_{c_0,e_D(c_0)}$ minimizes the D-expected log loss among all distributions in $\mathcal{P}$). Thus, the posterior does not converge to the optimal $p_{c_0,e_D(c_0)}$ even though $p_{c_0,e_D(c_0)}$ has substantial prior mass and is *partially* correct in the sense that c_0, the Bayes optimal classifier relative to $p_{c_0,e_D(c_0)}$, has true error rate $e_D(c_0)$, which is the *same* true error rate that it would have if $p_{c_0,e_D(c_0)}$ were 'true'.
2. *'pragmatic' Bayesian inference for classification can be suboptimal*: a standard way to turn classifiers into distributions so as to make application of Bayesian inference possible may give rise to suboptimal performance.

6.2 Two types of misspecification

$p_{c_0,e_D(c_0)}$ can be misspecified in *two different ways*. To see this, note that $p_{c_0,e_D(c_0)}$ expresses that

$$y = c_0(x) \texttt{ xor } z, \tag{47}$$

where z is a noise bit generated independently of x. This statement may be wrong under distribution D either because (a) c_0 is not the Bayes optimal classifier according to D; or (b) c_0 is Bayes optimal, but z is dependent on x under D. Let us consider both of them in more detail.

(a) *no Bayes optimal classifier in $\mathcal{C}$.* The way we defined the learning problem D used in the proof of Theorem 2 (Section 4.1) is an example of this case.

This type of misspecification is subtle, because if we consider the optimal c_0 *in isolation*, ignoring the features X_i which do not influence the prediction made by c_0, then the conditional distribution $P(Y = 1 \mid c_0(X), \theta^*)$ becomes correct after all, in the sense that

[4] In particular, see Eq. (3.6) in Doob (1949) combined with the remark at the end of Section 3 of Doob's paper.

it is identical to the true conditional probability. That is: for all $x_0 \in \{0, 1\}$, we have

$$p_{c_0, e_D(c_0)}(Y = 1 \mid X_0 = x_0) = \Pr_{X,Y \sim D}(Y = 1 \mid c_0(X) = x_0, X_0 = x_0),$$

so $p_{c_0, e_D(c_0)}(\cdot \mid X_0 = x_0)$ is 'true'. This may imply that the set of distributions corresponding to $\mathcal{C}$ is well-specified, since c_0 only 'listens' to feature X_0. Yet still, misspecification occurs because for some $x \in \{0, 1\}^{\infty}$,

$$p_{c_0, e_D(c_0)}(Y = 1 \mid X = x) \neq \Pr_{X,Y \sim D}(Y = 1 \mid c_0(X) = x_0, X = x).$$

(b) *$\mathcal{C}$ contains Bayes act, but D is heteroskedastic.* It may seem that our theorem is only applicable to misspecification of type (a). But it is easy to see that it is just as applicable to the - arguably less serious—misspecification of type (b). Namely, in the proof of Theorem 2 (Section 4.1), we could have equally used the following slightly modified learning problem : step 1 and step 2 remain identical, so $c_1, c_2, \ldots$ are defined as before. The optimal c_0 is now defined by modifying step 3 as follows: for an easy example, we set $x_0 = y$. For a hard example, we set $x_0 = |1 - y|$ with probability $\mu/2\mu'$. Then the proof of Theorem 2 holds unchanged. But now c_0 *is* the Bayes optimal classifier relative to D, as is easy to see.

6.3 Why is the result interesting for a Bayesian?

Here we answer several objections that a cautious Bayesian might have to this work.

6.3.1 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 (Grünwald, 1998). 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.

6.3.2 It is already well-known that Bayesian inference can be inconsistent even if $\mathcal{P}$ is well-specified, i.e. if it contains D (Diaconis & Freedman, 1986; Barron, 1998). So why is this result interesting?

The (in)famous inconsistency results by Diaconis and Freedman (1986) are based on nonparametric inference with uncountable sets $\mathcal{P}$. It follows from Barron (1998) that their theorems require that the true distribution D has 'extremely small' prior density in the following sense: the prior mass of ϵ-Kullback-Leibler balls around D is exponentially small in $1/\epsilon$. Since such priors do not allow one to compress the data, from an MDL perspective it is not at all surprising that they lead to inconsistent inference (Barron, 1998). In contrast, in our result, rather than small prior densities we require misspecification. Since Diaconis and Freedman do not require misspecification, in a sense, our result is weaker. On the other hand, in our setting, the prior on the $p \in \mathcal{P}$ closest in KL divergence to the true conditional distribution p_D can be arbitrarily close to 1, whereas Diaconis and Freedman require the prior of the 'true' p_D

to be exponentially small in the sense explained above. In this sense, our result is stronger than theirs.

Barron (1998) exhibits an example of Bayesian inconsistency that is closer in spirit to ours. In his example, the prior *density* of KL-neighborhoods of the true D can be substantial. Nevertheless, his example requires that $\mathcal{P}$ contains uncountably many distributions. It is not possible to extend Barron's example to a case with only countably many distributions, since in that case, the posterior *must* concentrate[5] on the true D by Doob's result. Our result shows that even in the countable case, as soon as one allows for slight misspecification, the posterior may not converge to the best distribution in $\mathcal{P}$. Indeed, by an appropriate setting of the parameters μ and μ' it is seen from Theorem 2 that for every $\epsilon > 0$, no matter how small, we can exhibit a D with

$$\min_{c,\theta} \mathrm{KL}(p_D \| p_{c,\theta}) = \epsilon$$

for which Bayes is inconsistent with D-probability 1. This is interesting because even under misspecification, Bayes is consistent under fairly broad conditions (Bunke & Milhaud, 1998; Kleijn & van der Vaart, 2004), in the sense that the posterior concentrates on a neighborhood of the distribution that minimizes KL-divergence to the true D. We showed that if such conditions are violated, then consistency may fail dramatically. Thus, we feel our result is relevant at least from the *inconsistency under misspecification* interpretation.

6.3.3 So how can the 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 occur if $\mathcal{C}$ were finite or had finite VC-dimension for example.
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. Theorem 5.5 of Li (1997) shows that for general convex i.i.d. families (not just Gaussian mixtures), under conditions on the priors, two-part MDL (essentially the version of MDL that we consider here) is consistent in the sense of expected Kullback-Leibler risk. Although we have no formal proof, Li's result strongly suggests that with such priors, the Bayesian MAP and full Bayesian approach will also be consistent.

Our setup violates both conditions: $\mathcal{C}$ has infinite VC-dimension, and the corresponding $\mathcal{P}$ is not closed under taking mixtures. The latter issue is discussed further in Example 1.

6.3.4 How 'standard' is the conversion from classifiers to probability distributions on which the results are based?

One may argue that the 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, the transformation coincides with the logistic

[5] More precisely, the posterior mass on the set of all distributions in $\mathcal{P}$ that are mutually singular with D must go to 0 with D-probability 1.

regression transformation, which is a standard Bayesian tool; see for example (Jordan, 1995; Platt, 1999; Tipping, 2001). But the theorems are based on classifiers with 0/1-output. With the exception of decision trees, such classifiers have not been addressed frequently in the Bayesian literature. Decision trees have usually been converted to conditional distributions somewhat differently: one uses the same logistic transformation as we do, but one assumes a different noise rate *in each leaf* of the decision tree (Heckerman et al., 2000); thus, the transformation is done locally for each leaf rather than globally for the whole hypothesis. Since the noise rate can depend on the leaf, the set of all decision trees of arbitrarily length on a given input space $\mathcal{X}$ coincides with the set of all conditional distributions on $\mathcal{X}$. Thus it avoids the misspecification, and therefore the inconsistency problem, but at the cost of using a much larger model space.

Thus, here is a potentially weak point in the analysis: we use a transformation that has mostly been applied to real-valued classifiers, whereas here the classifiers are 0/1-valued. Nevertheless, to get an idea of how reasonable our transformation is, we simply *tested* it with three professing Bayesians. We did this in the following way: we first described the set of classifiers $\mathcal{C}$ used in the learning problem, and we said that we would like to perform Bayesian inference based on some prior over $\mathcal{C}$. We then asked the Bayesian how (s)he would handle this problem. All three Bayesians said that they would construct conditional distributions according to the logistic transformation, just as we did. We take this as evidence that the logistic transformation is reasonable, even for classifiers with binary outputs.

Whether the inconsistency results can be extended in a natural way to classifiers with real-valued output such as support vector machines remains to be seen. The fact that the Bayesian model corresponding to such neural networks will still typically be misspecified strongly suggests (but does not prove) that similar scenarios may be constructed.

6.3.5 Is there an alternative, more sophisticated transformation that avoids inconsistencies?

Even though the transformation we perform is standard, there may exist some other method of transforming a set of classifiers+prior into a set of distributions+prior that avoids the problems. There are only two obvious options which suggest themselves:

1. *Avoiding misspecification.* First, we can try to avoid misspecification; then by the strong Bayesian consistency theorems referred to in Question 6.3.2, we should be guaranteed to converge to the optimal classifier. However, as we explain below, this is often not practical.
2. *Ensuring $\mathcal{P}$ is convex.* Second, rather than using the set of transformed classifiers $\mathcal{P}$, we could put a prior on its convex closure $\overline{\mathcal{P}}$ (this is the set of all finite and infinite mixture distributions that can be formed from elements of $\mathcal{P}$. Note in particular that $\mathcal{P}$ and $\overline{\mathcal{P}}$ are sets of distributions defined on one outcome, not on a sample of $m > 1$ outcomes). Then, we can once again apply the consistency theorem for convex $\mathcal{P}$ referred to in Question 6.3.3, and we should be guaranteed to converge to the optimal distribution. Computational difficulties aside, this approach will not work, because now the distribution we converge to may not be the distribution we are interested in, as we describe further below.

Thus, the only two straightforward solutions to the transformation problem are either impractical or do not work. We discuss both of these in detail below. There may of course exist some clever alternative method that avoids all problems, but we have no idea how it would look like.

1. Can we ensure consistency by avoiding misspecification? From a subjective Bayesian perspective, one might require the learning agent to think hard enough about his or her prior probabilities so that the set of conditional distribution $\mathcal{P}$ does contain D, the true state of nature. In practice this means that one should ensure that $\mathcal{C}$ contains the Bayes optimal classifier with respect to D, and that $\mathcal{P}$ should contain distributions in which the noise z (Eq. (47)) can depend on the feature value x. In practical machine learning applications one will often have no idea how the Bayes optimal classifier behaves or how the noise depends on x. Thus, the only way to proceed seems to design a prior on *all* possible classifiers and *all* possible noise rate functions. Now the inconsistency problem is solved, because the ('nonparametric') model thus constructed is guaranteed to contain the true (conditionalized) distribution D, so common Bayesian consistency theorems (see above) apply. However, the cost may be enormous: the model space is now *much* larger and it seems that a lot more data may be needed before a reasonable approximation of D is learned—although interestingly, recent work by Hutter (2005) suggests that under suitable priors, reasonable approximations may be learned quite fast. It is not clear whether or not something like this can be done in our context.

2. Can we ensure consistency by using convex models? Suppose we first use the logistic transformation to transform the classifiers $\mathcal{C}$ into a set of conditional distributions $\mathcal{P}$, and we then put a prior on its convex closure $\overline{\mathcal{P}}$ and use Bayesian inference based on $\overline{\mathcal{P}}$. Now , Li's result (Section 6.3.3) suggests that the Bayesian posterior predictive distribution is (under weak conditions on the prior) guaranteed to converge to the closest distribution p^* to D within $\overline{\mathcal{P}}$, as measured in KL-divergence. However, as the following example shows, p^* 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. Thus, existing theorems suggest that with a prior on $\overline{\mathcal{P}}$, the Bayesian posterior will *converge*, but below we show that if it does converge, then it will sometimes converge to a distribution that is suboptimal in the performance measure we are interested in.

Example 1 (Classification error and taking mixtures). We consider the following learning problem. There are three classifiers $\mathcal{C} = \{c_1, c_2, c_3\}$ and three features X_1, X_2, X_3 taking values in $\{0, 1\}$. Each classifier simply outputs the value of the corresponding feature. The underlying distribution D is constructed as follows. We distinguish between three 'situations' s_1, s_2, s_3 (these are the values of some random variable $S' \sim D$ that is not observed). 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 fair three-sided coin to determine what situation we are in, so $S' = s_j$ with probability $1/3$, for $j \in \{1, 2, 3\}$. Now if we happen to be in situation s_j, we

1. Set $x_j = y$ (so c_j will predict Y correctly).
2. Flip a fair coin, determine the outcome $z \in \{0, 1\}$, and set $x_{j'} = z$ for the two values of $j' \in \{1, 2, 3\}$ that are not equal to j.

Thus, the value of $x_{j'}$ is determined completely at random, but must be the same for both features not equal to j. We thus have for $j = 1, 2, 3$:

$$e_D(c_j) = \frac{1}{3} \cdot 0 + \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3}, \tag{48}$$

$$\begin{aligned} \mathrm{KL}(p_D \| p_{c_j, e_D(c)}) &= E_{(x,y)\sim D}[-\log p_{c_j, e_D(c_j)}(y|x) + \log p_D(y|x)] \\ &= H(e_D(c_j)) - K_D = H\left(\frac{1}{3}\right) - K_D > .9 - K_D. \end{aligned} \tag{49}$$

Equation (49) follows by (42), and as in that equation, H is the binary entropy as defined above Theorem 2, and K_D is the conditional entropy of y given x according to D, which does not depend on j.

Thus, the distribution(s) in $\mathcal{P} := \{p_{c_j,\theta} \mid j \in \{1,2,3\}, \theta \in [0,1]\}$ closest to the underlying D in KL-divergence have KL divergence $H(\frac{1}{3}) - K_D$ to D.

Now consider the set of conditional distributions $\overline{\mathcal{P}}$ defined as the convex closure of $\mathcal{P}$. It is easy to see that each element of $\overline{\mathcal{P}}$ can be written as a three-component mixture

$$\overline{p}_{\vec{\alpha},\vec{\theta}} := \alpha_1 p_{c_1,\theta_1} + \alpha_2 p_{c_2,\theta_2} + \alpha_3 p(c_3, \theta_3),$$

for some vector $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ with nonnegative entries summing to 1, and $(\theta_1, \theta_2, \theta_3)$ with nonnegative entries ≤ 1. Thus, the distribution in $\overline{P}$ that is closest to D in KL divergence is the distribution that achieves the minimum over $\vec{\alpha}$ and $\vec{\theta}$ of the expression

$$\mathrm{KL}(p_D \| p_{\vec{\alpha},\vec{\theta}}) = E_{(x,y)\sim D}\left[\log \frac{1}{p_{\vec{\alpha},\vec{\theta}}(y|x)}\right] - K_D. \tag{50}$$

This expression is uniquely minimized for some $\overline{p}^*$ with parameters

$$\alpha_1^* = \alpha_2^* = \alpha_3^* = \frac{1}{3} \text{ and some } \theta^* \in [0,1] \text{ with } \theta^* = \theta_1 = \theta_2 = \theta_3. \tag{51}$$

To see this, note that by symmetry of the problem, $\mathrm{KL}(p_D \| p_{\vec{\alpha},\vec{\theta}}) = \mathrm{KL}(p_D \| p_{\vec{\alpha}',\vec{\theta}'})$ where $\vec{\alpha}' := (\alpha_2, \alpha_1, \alpha_3)$ and $\vec{\theta}' := (\theta_2, \theta_1, \theta_3)$. Since $\overline{\mathcal{P}}$ is closed under mixing, for any $\gamma \in [0,1]$, $p_\gamma := \gamma p_{\vec{\alpha},\vec{\theta}} + (1-\gamma) p_{\vec{\alpha}',\vec{\theta}'}$ must be in $\overline{\mathcal{P}}$. By strict convexity of KL divergence (Cover & Thomas, 1991) and symmetry of the problem, $\mathrm{KL}(p_D \| p_\gamma)$ is uniquely minimized for $\gamma = 1/2$, and then p_γ satisfies $\alpha_1 = \alpha_2$ and $\theta_1 = \theta_2$. In the same way one shows that the minimizing $\vec{\alpha}$ and $\vec{\theta}$ have to satisfy $\alpha_2 = \alpha_3, \theta_2 = \theta_3$ and $\alpha_1 = \alpha_3, \theta_1 = \theta_3$, and (51) follows. Now plugging the minimizing parameters (51) into (50) gives

$$\begin{aligned}\mathrm{KL}(p_D \| \overline{p}^*) &= \min_{\theta \in [0,1]} \mathrm{KL}(p_D \| p_{\vec{\alpha}^*,\vec{\theta}^*}) \\ &= \min_{\theta} -\frac{1}{2}[\log(1-\theta) + \log(1+\theta) - \log 3] - K_D \\ &= \frac{1}{2}\log 3 - K_D < .8 - K_D,\end{aligned} \tag{52}$$

which is strictly smaller than (49). Therefore, while (a) Li's consistency result (Section 6.3.3) for convex $\overline{P}$ suggests that both the Bayesian posterior and Bayesian MAP conditional distribution will converge (in expected KL-divergence), to $\overline{p}^*$, it turns out that, (b) the classification error rate of the Bayes classifier c_{p^*} corresponding to the resulting conditional distribution $\overline{p}^*$ is equal to

$$E_{x,y\sim D}[|y - c_{p^*}(x)|] = \left[\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 0\right] = \frac{1}{2},$$

which is *worse* than the optimal classification error rate that can be obtained within $\overline{\mathcal{P}}$: since $\mathcal{P} \subset \overline{\mathcal{P}}$, by (48) this error rate must be $\leq 1/3$.

Concluding, with D-probability 1, for large m, the error rate of the Bayes classifier based on the Bayesian posterior relative to $\overline{\mathcal{P}}$ will have classification error that is *larger* than that of the Bayesian posterior relative to $\mathcal{P}$: it is clear that by enlarging the model $\mathcal{P}$ to its convex closure, rather than sometimes not converging at all, we may now converge to a suboptimal distribution: instead of solving the problem, we merely replaced it by another one.

6.3.6 Isn't the example just "unnatural"?

Upon hearing our results, several people objected that our learning problem is "unnatural". We agree that it is unlikely that one will ever deal with such a scenario in practice. However, this does not rule out the possibility that related phenomena do occur in more practical settings; see (Clarke, 2004) for an example in a regression context. Part of the problem here is of course that it is not really clear what "unnatural" means. Indeed, it is certainly not our aim to show that "Bayesian inference is bad". Instead, one of our main messages is that more research is needed to determine under what types of misspecification Bayes performs well, and under what types it does not.

7 Interpretation from an MDL perspective

We now discuss the interpretation of our result from an MDL Perspective. Similar to the Bayesian analysis, we do this by answering objections that a cautious description length minimizer might have to this work.

7.1.1 Why is the two-part code (7) the appropriate formula to work with? Shouldn't we use more advanced versions of MDL based on one-part codes?

Equation (7) has been used for classification by various authors; see, e.g., (Rissanen, 1989; Quinlan & Rivest, 1989; Kearns et al., 1997). Grünwald (1998, Chapter 5) first noted that in this form, by using Stirling's approximation, (7) is essentially equivalent to MAP classification based on the models $p_{c,\theta}$ as defined in Section 2. Of course, there exist more refined versions of MDL based on one-part rather than two-part codes (Barron, Rissanen, & Yu, 1998). To apply these to classification, one somehow has to map classifiers to probability distributions explicitly. This was already anticipated by Meir and Merhav (1995) who used the transformation described in this paper to define one-part MDL codes. The resulting approach is closely related to the Bayesian posterior approach $c_{\text{BAYES}(P,S)}$, suggesting that a version of the inconsistency Theorem 2 still applies. Rissanen (1989) considered mapping classifiers $\mathcal{C}$ to distributions $\{p_{c,\theta^*}\}$ to a *single* value of θ^*, e.g., $\theta^* = 1/3$. As discussed in Section 5.1, a version of Theorem 2 still applies to the resulting distributions.

We should note that both Wallace and Patrick (1993) and Quinlan and Rivest(1989) really use an extension of the coding scheme expressed by (7), rather than the exact formula (7) itself: both publications deal with decision trees, and apply (7) on the level of the leaf nodes of the decision trees. The actual codelength for the data given a decision tree becomes a sum of expressions of the form (7), one for each leaf. This means that they are effectively estimating error rates separately for each leaf. Since their model consists of the set of all decision trees of arbitrary depth, they can thus essentially model almost any conditional distribution of Y given X. This makes their approach nonparametric, and therefore, broadly speaking, immune to misspecification as long as data are i.i.d., and therefore immune to our results: inconsistency can only arise if the coding scheme (7) is applied to a model that can only present homoskedasticity, whereas the data generating distribution is heteroskedastic. It is not clear though whether the use of nonparametric models such as decision trees always solves the problem in *practice*, as we already discussed in Section 6.3.5., Question 1. As a further (but inessential) difference, Quinlan and Rivest (1989) use one extra bit on top of (7) for each leaf node of the decision tree. Wallace and Patrick (1993) point out that this is unnecessary, and use more general codes based on Beta-priors, of which our code (7) is a special case, obtained with the uniform prior (see Proposition 2 in the Appendix). As can be seen from the proof of Theorem 2, the use of general Beta-priors in the definition of MDL will not affect the inconsistency result.

7.1.2 Does the coding scheme for hypotheses make sense from an MDL perspective?

MDL theory prescribes the design of codes for hypothesis spaces (roughly corresponding to priors) that minimize worst-case regret or redundancy (Barron, Rissanen, & Yu, 1998; Grünwald, 2007) of the resulting codelength of hypothesis + data. It may seem that our coding scheme for hypotheses does not satisfy this prescription. But in fact, it does: if no natural grouping of the hypotheses in subclasses exists (such as with Markov chains, the class of k-th order Markov chains being a natural subclass of the class of $k + 1$-st order chains), then the 'best', from an MDL perspective, code one can assign is a code such that the code length of c_i goes to infinity as slowly as possible with increasing index i (Grünwald, 2007), such as Rissanen's universal code for the integers (Eq. (8)). But this is exactly the type of codes to which our Theorem 2 applies!

Lest the reader disagree with this: according to 'standard' MDL theory, if $\mathcal{P}$ is well-specified and countable then the coding scheme should even be asymptotically irrelevant: *any* coding scheme for the hypothesis where the codelength of any $P \in \mathcal{P}$ does not depend on n, will lead to asymptotically consistent MDL inference under very weak conditions (Barron & Cover, 1991); see also Chapter 5, Theorem 5.1 of Grünwald (2007). Special types of codes minimizing worst-case regret are only needed to *speed up* up learning with small samples; for large samples, any code will do. Thus, our result shows that if a set of classifiers $\mathcal{C}$ is used (corresponding to a misspecified probability model $\mathcal{P}$), then the choice of prior becomes of crucial importance, even with an infinite amount of data.

7.1.3 It seems that MDL can already be inconsistent even if $\mathcal{P}$ is well-specified. So why is the result interesting?

This question mirrors Question 6.3.2. In Section 1.2 of Wallace and Dowe (1999b), a very simple problem is discussed for which a straightforward implementation of a two-part code estimator behaves quite badly, even though the true distribution is contained in the model $\mathcal{P}$, and $\mathcal{P}$ only contains 1 continuous-valued parameter. This suggests that MDL may be inconsistent in a setting that is much simpler than the one we discuss here. But this is not quite the case: if the true distribution is contained in $\mathcal{P}$, then *any* two-part code will be asymptotically consistent, as long as the code is 'universal'; see Theorem 15.3 in Chapter 15 of Grünwald (2007). Under the definition of MDL that has generally been adopted since Barron, Rissanen, and Yu (1998), an estimator based on a two-part code can only be called 'MDL estimator' if the code is universal. Thus, it may either be the case that the two-part code defined by Wallace and Dowe (1999b) is not universal, and hence not an MDL code, or their two-part code must be asymptotically consistent after all. We suspect that the latter is the case. From an MDL perspective, the interest in our example is that, under misspecification, we can get inconsistency, even though we do use a universal two-part code.

7.1.4 Haven't Kearns et al. (1997) already shown that MDL is no good for classification?

It may seem that the results are in line with the investigation of Kearns et al. (1997). This, however, is not clear—Kearns et al. consider a scenario in which two-part code MDL for classification shows quite bad experimental performance for large (but not infinite!) sample sizes. However, according to Viswanathan et al. (1999), this is caused by the coding method used to encode hypotheses. This method does not take into account the precision of parameters involved (whereas taking the precision into account is a crucial aspect of MDL!). In the paper (Viswanathan et al., 1999), a different coding scheme is proposed. With this coding scheme, MML (an inference method that is related to MDL, see below) apparently behaves quite well on the classification problem studied by Kearns et al. In contrast to Kearns' example, in our case (a) there is no straightforward way to improve the coding scheme; (b) MDL fails even on an infinite sample.

7.1.5 What about MML?

The *Minimum Message Length (MML) Principle* (Wallace & Boulton, 1968; Comley & Dowe, 2005; Wallace, 2005) is a method for inductive inference that is both Bayesian and compression-based. The similarities and differences with MDL are subtle; see, for example, Section 10.2 of Wallace (2005) or Section 17.4 of Grünwald (2007), or (Wallace & Dowe, 1999a,b). An anonymous referee raised the possibility that MML may be consistent for the combination of the learning problem and the misspecified probability model discussed in this paper. We suspect that this is not the case, but we are not sure of this, and for the time being, the question of whether or not MML can be inconsistent under misspecification in classification contexts remains open. For the well-specified case, it is conjectured on page 282 of Wallace and Dowe (1999a) that only MML or closely related techniques can infer fully-specified models with both statistical consistency and invariance under one-to-one parameterization.

Related Work. Yamanishi (1998) and Barron (1990) proposed modifications of the two-part MDL coding scheme so that it would be applicable for inference with respect to general classes of predictors and loss functions, including classification with 0/1-loss as a special case. Both Yamanishi and Barron prove the consistency (and give rates of convergence) for their procedures. Similarly, McAllester's (1999) PAC-Bayesian method can be viewed as a modification of Bayesian inference that is provably consistent for classification, based on sophisticated extensions of the Occam's Razor bound, Theorem 4. These modifications anticipate our result, since it must have been clear to the authors that without the modification, MDL (and discrete Bayesian MAP) are not consistent for classification. Nevertheless, we seem to be the first to have explicitly formalized and proved this.

8 Conclusion and future work

We showed that some standard versions of MDL and Bayesian inference can be inconsistent for a simple classification problem, and we extensively discussed the interpretation of this result. As possible future work, it would be interesting to investigate

1. Whether there is a more natural learning problem, especially a more natural feature space, with respect to which an analogue to our result still holds.
2. Whether a similar result holds for regression rather than classification problems. We conjecture that the answer is *yes, but the suboptimality will be less dramatic*.

Appendix: Proposition 2 and its Proof

Proposition 2. *Consider any given sample S of arbitrary size m.*

1. *Let $c \in \mathcal{C}$ be an arbitrary classifier and let $P(\theta|c)$ be given by the uniform prior with $P(\theta\,|\,c) \equiv 1$. Then*

$$\begin{aligned} -\log P(y^m \mid x^m, c) &= -\log \int_0^1 P(y^m \mid x^m, c, \theta) d\theta \\ &= \log(m+1) + \log \binom{m}{m\hat{e}_S(m)}. \qquad (53) \end{aligned}$$

   so that, if the uniform prior is used, then $c_{\mathrm{MDL}(P,S)} = c_{\mathrm{SMP}}$.

2. *Suppose that $P(\theta\,|\,c)$ satisfies* (9) *or* (11)*, and that for some $\alpha > 0$, $\hat{e}_S(c) < 0.5 - \alpha$. Then*

$$\begin{aligned} mH(\hat{e}_S(c)) = \log \frac{1}{P(y^m\,|\,x^m, c, \hat{e}_S(c))} &\le \log \frac{1}{P(y^m\,|\,x^m, c)} \\ &\le \log \frac{1}{P(y^m\,|\,x^m, c, \hat{e}_S(c))} + f_\alpha(m) = mH(\hat{e}_S(c)) + f_\alpha(m), \qquad (54) \end{aligned}$$

   where $f_\alpha(m) = O(\log m)$, and the constant in the O-term may depend on α.

Proof: We recognize the integral in (53) as being a beta-integral. Straightforward evaluation of the integral (e.g. by partial integration) gives the result of part (1). For part (2), the leftmost and rightmost equalities follow by straightforward rewriting. The first inequality follows because

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

since the likelihood $P(y^m \,|\, x^m, c, \theta)$ is maximized at $\theta = \hat{e}_S(c)$. For the second inequality, we first consider the case that $P(\theta|c)$ satisfies (9). Then using (53),

$$\begin{aligned}\log \frac{1}{P(y^m \,|\, x^m, c)} &\leq -\log \int_0^{0.5} P(y^m \,|\, x^m, c, \theta) d\theta - \log \gamma \\ &\leq -\log \int_0^1 P(y^m \,|\, x^m, c, \theta) d\theta + \log \frac{\int_0^1 P(y^m \,|\, x^m, c, \theta) d\theta}{\int_0^{0.5} P(y^m \,|\, x^m, c, \theta) d\theta} - \log \gamma \\ &= \log(m+1) + \log \binom{m}{m\hat{e}_S(m)} - \log \gamma + o(1), \end{aligned} \tag{55}$$

where the constant in the $o(1)$ depends on α. The result for $P(\theta)$ satisfying (9) now follows upon noting that for all $s \in \{0, 1, \ldots, m\}$, $mH(s/m) \geq \log \binom{m}{s}$. This is the case because $mH(s/m)$ is the number of bits needed to encode m outcomes with s ones, using a Bernoulli distribution with parameter s/m; whereas $\log \binom{m}{s}$ is the number of bits needed to encode m outcomes with s ones, using a Bernoulli distribution with parameter s/m, conditioned on the relative frequency of 1s being s/m—thus, the same sequence is encoded using the same code, but conditioned on extra information, so that equally many or less bits are needed.

Now consider the case that $P(\theta|c)$ satisfies (11). Then

$$\begin{aligned} P(y^m \,|\, x^m, c) = \sum_{\theta \in [0,1] \cap \mathbb{Q}} P(y^m \,|\, x^m, c, \theta) P(\theta|c) &\geq P(y^m \,|\, x^m, c, \hat{e}_S(c)) P(\hat{e}_S(c) = \theta | c) \\ &\geq P(y^m \,|\, x^m, c, \hat{e}_S(c)) K_1 m^{-K_2}, \end{aligned} \tag{56}$$

for some constants K_1 and K_2. The result now follows by taking negative logarithms. □

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. We would especially like to thank Marcus Hutter who pointed out an error in a previous version of Theorem 2.

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

Barron, A. R. (1985). *Logically smooth density estimation*. PhD thesis, Department of EE, Stanford University, Stanford, Ca.

Barron, A. R. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian statistics*, vol. 6 (pp. 27–52). Oxford University Press.

Barron, A. R., Rissanen, J., & Yu, B. (1998). The MDL principle in coding and modeling. *IEEE Trans. Inform. Theory*, *44*(6), 2743–2760.

Barron, A. R. (1990). Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics* (pp. 561–576). Kluwer Academic Publishers.

Barron, A. R., & Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, *37*(4), 1034–1054.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. John Wiley.

Blackwell, D., & Dubins, L. (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, *33*, 882–886.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1987). Occam's razor. *Information Processing Letters*, *24*, 377–380.

Bunke, O., & Milhaud, X. (1998). Asymptotic behaviour of Bayes estimates under possibly incorrect models. *The Annals of Statistics*, *26*, 617–644.

Clarke, B. (2004). Comparing Bayes and non-Bayes model averaging when model approximation error cannot be ignored. *Journal of Machine Learning Research*, *4*(4), 683–712.

Comley, J. W., & Dowe, D. L. Minimum message length and generalised bayesian nets with asymmetric languages. In P. D. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: theory and applications*. MIT Press, 2005.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.

Diaconis, P., & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, *14*(1), 1–26.

Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul de Probabilités et ses Applications. Colloques Internationaux du Centre National de la Recherche Scientifique* (pp. 23–27), Paris.

Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

Grünwald, P. D. (1998). *The minimum description length principle and reasoning under uncertainty*. PhD thesis, University of Amsterdam, The Netherlands.

Grünwald, P. D. (2005). MDL tutorial. In P. D. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: theory and applications*. MIT Press

Heckerman, D., Chickering, D.M., Meek, C., Rounthwaite, R., & Kadie, C. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, *1*, 49–75.

Hutter, M. (2005). Fast non-parametric Bayesian inference on infinite trees. In *Proceedings of the 15th international workshop on artificial intelligence and statistics (AISTATS '05)*.

Jordan, M. I. (1995). Why the logistic function? a tutorial discussion on probabilities and neural networks. Computational Cognitive Science Tech. Rep. 9503, MIT.

Kearns, M., Mansour, Y., Ng, A.Y., & Ron, D. (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning*, *27*, 7–50.

Kleijn, B., & van der Vaart, A. (2004). Misspecification in infinite-dimensional Bayesian statistics. submitted.

Li, J. K. (1997). *Estimation of mixture models*. PhD thesis, Yale University, Department of Statistics.

McAllester, D. (1999). PAC-Bayesian model averaging. In *Proceedings COLT '99*.

Meir, R., & Merhav, N. (1995). On the stochastic complexity of learning realizable and unrealizable rules. *Machine Learning*, *19*, 241–261.

Platt, J. C. (1999). Probabilities for SV machines. In A. Smola, P. Bartlett, B. Schöelkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61–74). MIT Press.

Quinlan, J., & Rivest, R. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation*, *80*, 227–248.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, *11*, 416–431.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. World Scientific.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, *1*, 211–244.

Viswanathan, M., Wallace, C. S., Dowe, D. L., & Korb, K. B. (1999). Finding cutpoints in noisy binary sequences - a revised empirical evaluation. In *Proc. 12th Australian joint conf. on artif. intelligence*, vol. 1747 of *Lecture notes in artificial intelligence (LNAI)* (pp. 405–416), Sidney, Australia.

Wallace, C. S. (2005). *Statistical and Inductive Inference by Minimum Message Length*. New York: Springer.

Wallace, C. S., & Boulton, D. M. (1968). An information measure for classification. *Computing Journal*, *11*, 185–195.

Wallace, C. S., & Dowe, D. L. (1999a). Minimum message length and Kolmogorov complexity. *Computer Journal*, *42*(4), 270–283. Special issue on Kolmogorov complexity.

Wallace, C. S., & Dowe, D. L. (1999b). Refinements of MDL and MML coding. *Computer Journal*, *42*(4), 330–337. Special issue on Kolmogorov complexity.

Wallace, C. S., & Patrick, J. D. (1993). Coding decision trees. *Machine Learning*, *11*, 7–22.

Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Trans. Inform. Theory*, *44*(4), 1424–1439.