Bayesian Inconsistency under Misspecification

Peter D. Grünwald
CWI, P.O. Box 94079, 1090 GB
Amsterdam, The Netherlands
www.grunwald.nl

Abstract

This is a synopsis of the work underlying the author's contributed plenary presentation at the *Valencia 8 meeting on Bayesian Statistics*, held in Benidorm, June 2006.

We show that Bayesian inference can be *inconsistent under misspecification*. Specifically, we exhibit a distribution P^*, a model $\mathcal{M}$ with $P^* \notin \mathcal{M}$, and a prior Π on $\mathcal{M}$ such that the prior puts significant mass on $\tilde{P}$, the best approximation to P^* within the set $\mathcal{M}$. Yet, if data are i.i.d. $\sim P^*$, then for all large samples, the Bayesian posterior puts its mass on a subset of $\mathcal{M}$ that only contains bad approximations to P^*. This result holds both if approximation quality is defined in terms of Kullback-Leibler divergence and if it is defined in terms of classification risk. We present several variations of this result, including one in which, with P^*-probability 1, for all large enough samples, predictions of the next outcome based on the Bayesian predictive distribution become worse than predictions based on purely random guessing.

1 Introduction

In this paper, we show that Bayesian inference can be inconsistent under misspecification in classification problems with random design. In Section 2 we introduce the setting in which our result holds. In Section 3 we give a specific model $\mathcal{M}$, a variety of priors Π and a variety of underlying distributions $P^* \notin \mathcal{M}$ for which Bayesian inference becomes inconsistent. The precise inconsistency results are discussed in Section 4. Section 5 discusses the theoretical and practical relevance of this result.

Whenever below a remark is accompanied by the $\circ$-sign, this indicates that extensive additional discussion and/or references can be found in [Grünwald and Langford 2004], a publication containing preliminary versions of some of the results presented here. The present work significantly extends these earlier results. In this synopsis we present our results in an informal fashion only. A formal treatment will be given in a longer journal version.

2 General Setting: Classification with Random Design

In this section, we describe the general setting of our misspecification result: we first describe the type of models $\mathcal{M}$, then the types of underlying distribution P^*, and finally, we show

the intimate connection between 0/1-classification loss and logarithmic score that holds in our setting.

Classification Models Let (X, Y) be a pair of random variables, where $Y \in \{0, 1\}$ is binary-valued and X is arbitrary. We consider a probabilistic classification problem where we are interested in learning the conditional distribution of Y given X. To this end we introduce a model $\mathcal{M}$ of conditional distributions identified with their mass functions $p_{c,\beta}(Y|X)$, parameterized by a pair (c, β) where c is a function from the domain of X to $\mathbb{R}$, taken from some set $\mathcal{C}$ (to be specified later), and β is a positive real number. Thus, $\mathcal{M} = \{p_{c,\beta} \mid c \in \mathcal{C}, \beta \in (0, \infty)\}$, where $p_{c,\beta}$ is defined as

$$P_{c,\beta}(Y = 1 \mid X = x) = p_{c,\beta}(1 \mid x) = \frac{e^{\beta c(x)}}{1 + e^{\beta c(x)}}, \tag{1}$$

and extended to m outcomes by independence. Models of this logistic form are often used for probabilistic classification. They are also frequently considered in the field of machine learning in order to equip nonprobabilistic classifiers $\mathcal{C}$ with an associated noise process and then apply Bayesian inference to them°. For example, $\mathcal{C}$ may be the set of support vector machines with respect to some kernel [Tipping 2001].

Underlying Distributions We observe a sample $S^{(m)} = ((X_1, Y_1), \ldots, (X_m, Y_m))$ drawn i.i.d. according to some P^*, where P^* is the distribution of the *joint* random vector (X, Y). This is a common setting in machine learning; in statistical terminology, we may call this setting 'classification with a random, i.i.d. design'.

We may express the discrepancy between any $p_{c,\beta} \in \mathcal{M}$ and the 'true' distribution P^* by the Kullback-Leibler (KL) divergence° $D(p^* \| p)$, extended to conditional distributions in the obvious way: $D(p^* \| p) := E_{P^*}[\log \frac{p^*(Y|X)}{p(Y|X)}]$, where $p^*(Y|X)$ is the conditional mass function induced by the (joint) P^*. Here and in the sequel, the expectation is over the joint $(X, Y) \sim P^*$. We will only consider P^* such that this conditional divergence is well-defined. Let $\tilde{p} \in \mathcal{M}$ be the distribution in $\mathcal{M}$ minimizing conditional KL-divergence to P^*, i.e. $\tilde{p} = \arg\min_{p \in \mathcal{M}} D(p^* \| p)$. We will only consider combinations of P^* and $\mathcal{M}$ for which $\tilde{p}$ exists and is unique.

Scenarios In all our inconsistency scenarios, (a) $X \in [0, 1]$, i.e. X takes values in the unit interval; (b) the set $\mathcal{C} = \{c_0, c_1, \ldots\}$ is countable, (c) all $c_j \in \mathcal{C}$ are 'hard' classifiers such that for all x, $c_j(x) \in \{0, 1\}$, (d) P^* is designed so that c_0 is the classifier-component of the KL-optimal $\tilde{p}$. Hence $\min_{p \in \mathcal{M}} D(p^* \| p)$ is achieved for $\tilde{p} = p_{c_0,\beta_0}$ for some $\beta_0 \in \mathbb{R}$.

As a consequence of (c), the model of distributions (1) is particularly well-suited for dealing with misspecification, if the loss function of interest is the standard symmetric 0/1-loss, $\textsc{loss}(y, \delta) := |y - \delta|$, where $y, \delta \in \{0, 1\}$. Namely, if $c(x) \in \{0, 1\}$, we can reparameterize $p_{c,\beta}$ as $p_{c,\theta}$ with

$$\theta_\beta := \ln \beta - \ln(1 - \beta), \tag{2}$$

and then the (conditional) likelihood of $S^{(m)}$ according to $p_{c,\beta}$ becomes

$$p_{c,\theta}(Y^m \mid X^m) := p_{c,\theta}(Y_1, \ldots, Y_m \mid X_1, \ldots, X_m) = \theta^{m_1}(1 - \theta)^{m - m_1}, \tag{3}$$

where $\theta = \theta_\beta$ and m_1 is the number of outcomes on which c 'made a mistake', i.e. $m_1 = \sum_{i=1}^m |c(x_i) - y_i| = \sum_{i=1}^m \textsc{loss}(y_i, c(x_i))$. Now let $p = p(Y|X)$ represent a conditional distribution on Y and let $\delta_p(x) \in \{0, 1\}$ be the Bayes-optimal decision relative to p, predicting 1 if $p_{c,\beta}(Y = 1 \mid X = x) \geq 1/2$, and 0 otherwise. From (3) it follows$^\circ$ that for $p = p_{c,\beta}$, $\beta > 0$, we have $\delta_p(x) = c(x)$, i.e. the Bayes optimal act relative to p is just the classifier on which p is based. With each $p \in \mathcal{M}$, we associate its *classification risk*$^\circ$ $R(p)$ which is just the P^*-expected loss: $R(p) := E_{P^*}[\textsc{loss}(Y, \delta_p(X))]$. Now no matter what P^* is (in particular, we may have $P^* \notin \mathcal{M}$), we have that $\min_{p \in \mathcal{M}} R(p)$ is achieved$^\circ$ by $\tilde{p}$. To see the importance of this fact, note that, irrespective of the prior, if the posterior distribution converges at all with increasing m, then it will asymptotically put almost all of its mass on small KL-neighborhoods$^\circ$ around $\tilde{p} = \arg\min_{p \in \mathcal{M}} D(p^* \| p)$. But since $\tilde{p}$ is also $\arg\min_{p \in \mathcal{M}} R(p)$, it can be seen that the posterior predictive distribution, if it converges at all, also converges to the distribution with the smallest classification risk, which is what we are really interested in in 0/1-classification problems (see [Grünwald and Langford 2004], where we also explain that good classification performance and small KL divergence do not necessarily coincide under misspecification if the model $\mathcal{M}$ is not of the form (3), and discuss additional pleasant properties of models of the form (3)).

3 Specific Setting: 'hard' and 'easy' examples

We now describe a model $\mathcal{M}$ of the above form, together with a range of priors Π and a range of true distributions P^*, for which misspecification occurs.

The Model $\mathcal{M}$ The model $\mathcal{M} = \{p_{c,\beta} \mid c \in \mathcal{C}, \beta \in (0, \infty)\}$ is a set of conditional distributions of form (1) corresponding to a set of classifiers $\mathcal{C} := \{c_0, c_1, \ldots\}$ that is defined as follows: First, c_0 and c_1 are defined, respectively, as

$$c_0(x) := \begin{cases} 1 & \text{if } x \in \left[0, \frac{1}{4}\right) \cup \left[\frac{3}{4}, 1\right], \\ 0 & \text{if } x \in \left[\frac{1}{4}, \frac{3}{4}\right). \end{cases} \quad ; \quad c_1(x) := \begin{cases} 1 & \text{if } x \in \left[0, \frac{1}{8}\right) \cup \left[\frac{3}{8}, \frac{5}{8}\right) \cup \left[\frac{7}{8}, 1\right] \\ 0 & \text{if } x \in \left[\frac{1}{8}, \frac{3}{8}\right) \cup \left[\frac{5}{8}, \frac{7}{8}\right). \end{cases}$$

Note that $c_0(x)$ can be depicted as a step function, switching value between 1 and 0 at $X = 1/4$ and $X = 3/4$. $c_1(x)$ is a step function that switches value four rather than two times, at $1/8, 3/8, 5/8$ and $7/8$. c_2 is defined as the corresponding step function that switches value eight times, c_3 switches value 16 times, and in general, c_k switches value 2^{k+1} times, with $c_k(x) = 1$ if $x \in [0, 2^{-k-2})$, $c_k(x) = 0$ if $x \in [2^{-k-2}, 2^{-k-2} + 2^{-k-1})$, and so on.

The Prior Π The prior on $\mathcal{M}$ can be any distribution such that (a) the prior on β and the prior on c_j are probabilistically independent; (b) the prior density $\pi(\beta)$ is continuous, has full support, and $\lim_{\beta \downarrow 0} \pi(\beta) > 0$; and (c) there exists an $\alpha > 0$ such that the prior mass function $\pi(c_k)$ satisfies, for all large k,

$$\pi(c_k) > \frac{1}{k^{1+\alpha}}, \tag{4}$$

so that the prior has 'polynomial tails'. The extent of misspecification we can achieve below depends on the parameter α. The prior Π is not allowed to depend on the sample size.

We note that requirements (a) and (b) are in fact much stronger than necessary, and are merely imposed to simplify the treatment. Indeed, the results still hold if the prior on β is restricted to rational-valued β, see below. Requirement (c) however, is essential.

The True Distribution P^* We will actually specify a whole range of distributions $P^*_{\gamma,\eta}$ depending on two parameters $\gamma \in [0,1]$ and $\eta \in (0,\infty)$. They are defined as follows: first, a biased coin $Z \in \{0,1\}$ is thrown, with bias $P(Z=1) = \gamma$. If $Z=1$, then we sample X from a uniform distribution on the unit interval, and, conditional on the sampled $x \in [0,1]$, we sample Y from the distribution $p_{c_0,\eta}(Y = \cdot \mid X = x)$, as given by (1). If $Z = 0$, then we set both $X = 1$ and $Y = 1$.

This means that, if we sample i.i.d. from $P^*_{\gamma,\eta}$, then with high probability about $m(1-\gamma)$ examples (corresponding to $Z=0$) will have $(X_i, Y_i) = (1,1)$. We may call those examples 'easy', since all classifiers predict them correctly: we have $c_j(1) = 1$ for all $c_j \in \mathcal{C}$. For the other, 'hard' examples (corresponding to $Z = 1$) , c_0 will make a correct prediction with probability θ_η as given by (2). Note that $\theta_\eta > 0.5$. It is easy to see that for such 'hard' exapmles, for all other classifiers c_j , $j > 1$, in such cases, the probability of predicting Y_i correctly given X_i is exactly $1/2$.

Note that if we set $\gamma = 1$, then the conditional distribution $P^*_{\gamma,\eta}(Y = \cdot \mid X = \cdot) \in \mathcal{M}$. Then the model is well-specified and indeed, Bayesian inference with our priors will be consistent. We only get trouble if we choose $\gamma < 1$.

4 Results

Classification Loss Let $p_{\mathrm{Bayes},S^{(m)}} := p_{\mathrm{Bayes}}(Y|X, S^{(m)})$ be the Bayesian predictive distribution defined with respect to a prior Π as given above. Let $H(\mu) = -\mu \log_2 \mu - (1-\mu)\log_2(1-\mu)$ be the binary entropy. Now, pick any μ_0, μ_1 with $0 < \mu_0 < 0.5$ and $\mu_0 < \mu_1 < H(\mu_0)/2$. We show that for all such (μ_0, μ_1), there exist choices of α, γ and η such that if $P^* = P^*_{\gamma,\eta}$, and Π is a prior satisfying (4), then the following holds: (a) $R(\tilde{p}) = R(p_{c_0,\beta_0}) = \mu_0$, (b) for all $j > 0$, $\inf_{\beta>0} R(p_{c_j,\beta}) = \mu_1 > \mu_0$. Yet, (c), (*Theorem 1(a)*) with P^*-probability 1, $R(p_{\mathrm{Bayes},S^{(m)}})$ converges to μ_1. Hence, when used for classification, the Bayesian posterior performs like the bad distributions $p_{c_j,\beta}$ for $j > 1$, and not like $\tilde{p} = p_{c_0,\beta_0}$, which is both optimal in terms of classification risk and in the sense of minimizing KL divergence to the true P^*. Since for $0 < \mu < 0.5$, $H(\mu)/2 > \mu$, Theorem 1(a) implies 'classification inconsistency' for all $0 < \mu_0 < 1/2$. Theorem 1(a) is a rephrasing of Theorem 2 of our preliminary work [Grünwald and Langford 2004]. Theorem 1(b) (as yet unpublished) strengthens the result by showing that, under a somewhat different definition of P^*, the same result still holds as long as $\mu_0 < \mu_1 < H(\mu_0)$. Since $H(1/2) = 1$, this implies that for μ_0 close to but smaller than $1/2$, $c_0 = \delta_{p_{c_0,\beta_0}}$ performs slightly better than randomly guessing Y, yet the full Bayesian classifier has classification error tending to 1, being much *worse* than random guessing! Unfortunately there is no space here to give the definition of this more malignant P^*.

Our result captures the worst possible behaviour of Bayesian classification: *Theorem 2* (published before as Theorem 3 in [Grünwald and Langford 2004]) essentially shows that for $Y \in \{0,1\}$, arbitrary X and arbitrary $\mathcal{M}$ and any prior with full support on $\mathcal{M}$, if

$\min_{p\in\mathcal{M}} R(p) = R(\tilde{p}) = \mu_0$, then asymptotically, with high P^*-probability the risk of the full Bayesian classifier cannot be larger than $H(\mu_0)$. The situation is somewhat less bad for the classifier δ_{map} based on the Bayesian maximum a posteriori (MAP) distribution and variations thereof: *Theorem 1(b)* (as yet unpublished, strengthening a similar result in [Grünwald and Langford 2004]) shows that we can still get classification inconsistency for δ_{map}, but, for large m the MAP classifier can never be worse than random guessing. We note that $\tilde{p} = p_{c_0,\beta_0}$ has a Bayes act c_0 that is equal to the 'true' Bayes optimal act δ_{p^*} – misspecification arises because $\mathcal{M}$ wrongly assumes homoskedastic noise.

KL-divergence We also extend [Grünwald and Langford 2004] by studying the Bayesian posterior behaviour in terms of KL divergence. *Theorem 1 part (c)* says that, in our scenario (for suitable choices of $\alpha > 0, 0 < \eta < \infty$ and $0 < \gamma < 1$), with P^*-probability 1, for all large m, the Bayesian posterior puts all its mass, except for an exponentially small part, on a set $\mathcal{P}_m$ consisting of distributions $p_{c_j,\beta}$ with large j and β. With increasing sample size m, the minimum j and the minimum β with $p_{j,\beta} \in \mathcal{P}_m$ both tend to infinity. Note that $\beta = \infty$ amounts to the assumption of no noise. As a consequence, if $R(\tilde{p}) = \mu_0 > 0$, then $\inf_{p\in\mathcal{P}_m} D(p^*\|p) \to \infty$ almost surely. This holds even for arbitrarily small $\mu_0 > 0$, for which $D(p^*\|\tilde{p})$ is also arbitrarily small. In stark contrast, and perhaps as a relief, the Bayesian predictive distribution $p_{\text{Bayes},S^{(m)}}$ is asymptotically close to P^* (here 'close' is in the sense of accumulated KL risk, see [Barron 1998]). In fact, for large m, with high P^*-probability, $D(p^*\|p_{\text{Bayes},S^{(m)}})$ will be *smaller* than $D(p^*\|\tilde{p}) = \inf_{p\in\mathcal{M}} D(p^*\|p)$. Thus, while (a) the posterior predictive distribution $p_{\text{Bayes},S^{(m)}}$, a *mixture* of distributions with very large KL divergence to P^*, is itself closer to P^* in KL-divergence than the $\tilde{p}$ that minimizes KL-divergence to P^* among all $p \in \mathcal{M}$, at the same time, (b) the Bayes act $\delta_{p_{\text{Bayes},S^{(m)}}}$ based on $p_{\text{Bayes},S^{(m)}}$ has worse classification risk than the classifier $c_0 = \delta_{\tilde{p}}$ based on $\tilde{p}$, which, among all $p \in \mathcal{M}$, minimizes classification risk relative to P^*. At first sight paradoxical, it turns out that this is really essential: In our final result *(Theorem 4, as yet unpublished)*, we show that, for arbitrary models of form (1) with countable $\mathcal{C}$ and arbitrary X, for each P^* *either* the posterior asymptotically concentrates on the optimal $\tilde{p}$, *or* for large m, $D(p^*\|p_{\text{Bayes},S^{(m)}})$ is smaller than $D(p^*\|\tilde{p})$, with high probability; exactly one of the two has to be the case. This implies that we can have inconsistency in our sense only if for large m, the Bayesian predictive distribution *"predicts too well"* in terms of log score.

5 Relevance and Related Work

Theoretical Interest From a theoretical point of view, one might at first wonder why our result is important: it is well-known [Diaconis and Freedman 1986; Barron 1998] that Bayesian inference can be inconsistent even in the well-specified case, if $P^* \in \mathcal{M}$. So it does not seem surprising that we can get inconsistency under misspecification. However, the result becomes more surprising once we observe that our theorems still hold if we choose $\mathcal{M}$ to be *countable*: it is easy to check that if we only allow rational valued parameters β, i.e. by writing $\beta = p/q$ and putting any prior π on $(p, q) \in \mathbb{Z}^2$ polynomially decreasing in $|p|$ and $|q|$, then all our results still hold. With such a prior on $\mathcal{M}$, $\mathcal{M}$ becomes a countable set of distributions. Now standard Bayesian consistency theorems such as those

by Blackwell and Dubins [1962] or Barron [1985] imply that, if $P^* \in \mathcal{M}$ and $\mathcal{M}$ countable, we *must* have consistency. More precisely, in our setting, let $p^* = p^*(Y|X)$ be the conditional mass function corresponding to P^*. If $p^* \in \mathcal{M}$, then, by a theorem of Barron [1985], with P^*-probability 1, the posterior probability $\Pi(\{p^*\} \mid X^m, Y^m) \to 1$. Thus, our result is fundamentally different from earlier Bayesian inconsistency results of Barron [1998] and Diaconis and Freedman [1986], which are based on a situation with $P^* \in \mathcal{M}$ and therefore can only work for uncountable $\mathcal{M}$. Our result shows that if $P^* \notin \mathcal{M}$, one can have inconsistency even in the countable case. Thus, our result crucially depends on misspecification (although we can make the misspecification as small as we like). As indicated above (Theorem 4) it derives, in a sense, from the fact that under misspecification, the Bayesian posterior can predict *too well* (better than the best $\tilde{p} \in \mathcal{M}$) in the log score sense.

Practical Relevance In practice, Bayesian inference is employed under misspecification *all the time*, particularly so in machine learning applications°. While sometimes it works quite well under misspecification [Blei, Jordan, and Ng 2003; Kleijn and van der Vaart 2006], there are also cases where it does not [Clarke 2004; Fushiki 2005], so it seems important to determine precise conditions under which misspecification is harmful – even if such an analysis is based on frequentist assumptions.

References

Barron, A. (1985). *Logically Smooth Density Estimation*. Ph. D. thesis, Department of EE, Stanford University, Stanford, Ca.

Barron, A. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In A. D. J.M. Bernardo, J.O. Berger and A. Smith (Eds.), *Bayesian Statistics*, Volume 6, pp. 27–52.

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

Blei, D., M. Jordan, and A. Ng (2003). Hierarchical Bayesian models for applications in information retrieval. *Bayesian Statistics 7*, 25–43.

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.

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

Fushiki, T. (2005). Bootstrap prediction and Bayesian prediction under misspecified models. *Bernoulli 11*(4), 747–758.

Grünwald, P. D. and J. Langford (2004). Suboptimality of MDL and Bayes in classification under misspecification. In *Proceedings of the Seventeenth Conference on Learning Theory (COLT' 04)*. Springer-Verlag.

Kleijn, B. and A. van der Vaart (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics 34*(2).

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