Bayesian Analysis (2017) **12**, Number 4, pp. 1069–1103

Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It

Peter Grünwald* and Thijs van Ommen†

Abstract. We empirically show that Bayesian inference can be inconsistent under misspecification in simple linear regression problems, both in a model averaging/selection and in a Bayesian ridge regression setting. We use the standard linear model, which assumes homoskedasticity, whereas the data are heteroskedastic (though, significantly, there are no outliers). As sample size increases, the posterior puts its mass on worse and worse models of ever higher dimension. This is caused by *hypercompression*, the phenomenon that the posterior puts its mass on distributions that have much larger KL divergence from the ground truth than their average, i.e. the Bayes predictive distribution. To remedy the problem, we equip the likelihood in Bayes' theorem with an exponent called the learning rate, and we propose the *SafeBayesian* method to learn the learning rate from the data. SafeBayes tends to select small learning rates, and regularizes more, as soon as hypercompression takes place. Its results on our data are quite encouraging.

1 Introduction

We empirically demonstrate a form of inconsistency of Bayes factor model selection, model averaging and Bayesian ridge regression under model misspecification on a simple linear regression problem with random design. We sample data $(X_1, Y_1), (X_2, Y_2), \ldots$ i.i.d. from a distribution P^*, where $X_i = (X_{i1}, \ldots, X_{ip_{\max}})$ are high-dimensional vectors, and we allow $p_{\max} = \infty$. We use nested models $\mathcal{M}_0, \mathcal{M}_1, \ldots$ where $\mathcal{M}_p$ is a standard linear model, consisting of conditional distributions $P(\cdot \mid \beta, \sigma^2)$ expressing that

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i \tag{1}$$

is a linear function of $p \leq p_{\max}$ covariates with additive independent Gaussian noise $\epsilon_i \sim N(0, \sigma^2)$. We equip each of these models with standard priors on coefficients and the variance, and also put a discrete prior on the models themselves. We specify a 'ground truth' P^* such that $\mathcal{M} := \bigcup_{p=0,\ldots,p_{\max}} \mathcal{M}_p$ does not contain the conditional ground truth $P^*(Y \mid X)$ (hence the model is 'misspecified'), but it does contain a $\tilde{P}$ that is 'best' in several respects: it is closest to P^* in KL (Kullback–Leibler) divergence, it represents the true regression function (leading to the best squared error loss predictions among all $P \in \mathcal{M}$) and it has the true marginal variance (explained in Section 2.3).

*CWI, Amsterdam and Leiden University, The Netherlands, pdg@cwi.nl
†University of Amsterdam, The Netherlands, thijsvanommen@gmail.com

 DOI: 10.1214/17-BA1085

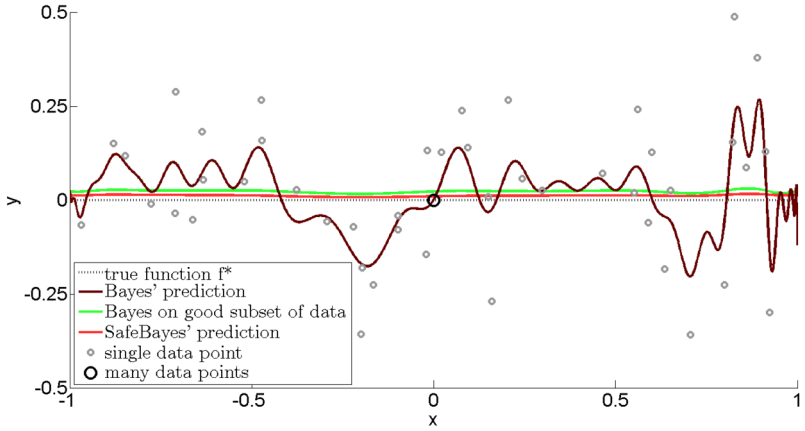

Figure 1: The conditional expectation $\mathbf{E}[Y \mid X]$ according to the full Bayesian posterior based on a prior on models $\mathcal{M}_0, \ldots, \mathcal{M}_{50}$ with polynomial basis functions, given 100 data points sampled i.i.d. $\sim P^*$ (about 50 of which are at $(0,0)$). Standard Bayes overfits, not as dramatically as maximum likelihood or unpenalized least squares, but still enough to show dismal predictive behaviour as in Figure 2. In contrast, SafeBayes (which chooses learning rate $\eta \approx 0.4$ here) and standard Bayes trained only at the points for which the model is correct (not $(0,0)$) both perform very well.

In fact we choose P^* such that $\tilde{P} \in \mathcal{M}_0$, and we choose our prior such that $\mathcal{M}_0$ receives substantial prior mass. Still, as n increases, the posterior puts most of its mass on complex $\mathcal{M}_p$'s with higher and higher p's, and, conditional on these $\mathcal{M}_p$'s, at distributions which are very far from P^* both in terms of KL divergence and in terms of L_2 risk, leading to bad predictive behaviour in terms of squared error. Figures 1 and 2 illustrate a particular instantiation of our results, obtained when X_{ij} are polynomial functions of S_i and $S_i \in [-1,1]$ uniformly i.i.d. We also show comparably bad predictive behaviour for various versions of Bayesian ridge regression, involving just a single, high-but-finite dimensional model. In that case Bayes eventually recovers and concentrates on $\tilde{P}$, but only at a sample size that is incomparably larger than what can be expected if the model is correct.

These findings contradict the folk wisdom that, if the model is incorrect, then "Bayes tends to concentrate on neighbourhoods of the distribution(s) $\tilde{P}$ in $\mathcal{M}$ that is/are closest to P^* in KL divergence." Indeed, the strongest actual theorems to this end that we know of, (Kleijn and Van der Vaart, 2006; De Blasi and Walker, 2013; Ramamoorthi et al., 2015), hold, as the authors emphasize, under regularity conditions that are substantially stronger than those needed for consistency when the model is correct (as by e.g. Ghosal et al. (2000) or Zhang (2006a)), and our example suggests that consistency may fail to hold even in relatively simple problems; to illustrate this further, in the supplementary material (Grünwald and Van Ommen, 2017), Section G.2, we show that the regularity conditions of De Blasi and Walker (2013) are violated in our setup.

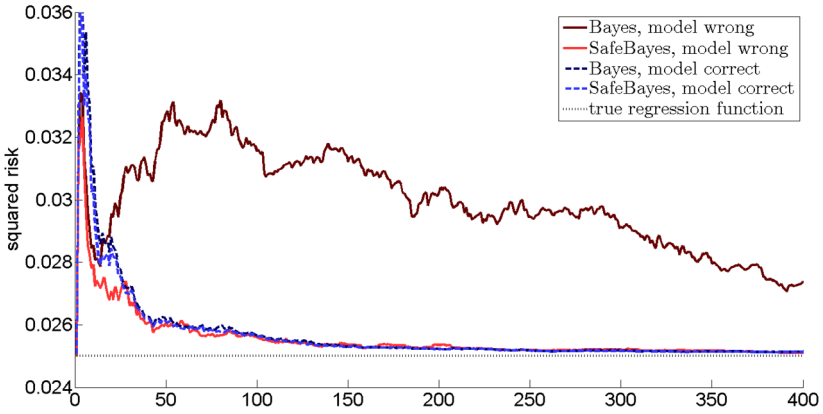

Figure 2: The expected squared error risk (defined in (4)), obtained when predicting by the full Bayesian posterior (brown curve), the SafeBayesian posterior (red curve) and the optimal predictions (black dotted curve), as a function of sample size for the setting of Figure 1. SafeBayes is the R-log-version of SafeBayes defined in Section 4.2. Precise definitions and further explanation in and above Section 5.1.

How inconsistency arises The explanation for Bayes' behaviour in our examples is illustrated in Figure 3, the essential picture to understand the phenomenon. As explained in the text (Section 3, with a more detailed analysis in the supplementary material), the figure indicates that there exists good or 'benign' and bad types of misspecification. Under bad misspecification, a phenomenon we call *hypercompression* can take place, and that explains why at the same time we can have a good log-score of the predictive distribution (as we must, by a result of Barron (1998)) yet a posterior that puts its mass on very bad distributions.

The solution: Generalized and SafeBayes Bayesian updating can be enhanced with a *learning rate* η, an idea put forward independently by several authors (Vovk, 1990; McAllester, 2003; Barron and Cover, 1991; Walker and Hjort, 2002; Zhang, 2006a) and suggested as a tool for dealing with misspecification by Grünwald (2011; 2012). η trades off the relative weight of the prior and the likelihood in determining the *η-generalized posterior*, where $\eta = 1$ corresponds to standard Bayes and $\eta = 0$ means that the posterior always remains equal to the prior. When choosing the 'right' η, which in our case is significantly smaller than 1 but of course not 0, η-generalized Bayes becomes competitive again. We give a novel interpretation of generalized Bayes in Section 4.1, showing that, for this 'right' η, it can be re-interpreted as standard Bayes with a different model, which now has 'good' rather than 'bad' misspecification. In general, this optimal η depends on the underlying ground truth P^*, and the remaining problem is how to determine the optimal η empirically, from the data.

Grünwald (2012) proposed the *SafeBayesian* algorithm for learning η. Even though lacking the explicit interpretation we give in Section 4.1, he mathematically showed that

it achieves good convergence rates in terms of KL divergence on a variety of problems.[1] Here we show empirically that SafeBayes performs excellently in our regression setting, being competitive with standard Bayes if the model is correct and very significantly outperforming standard Bayes if it is not. We do this by providing a wide range of experiments, varying parameters of the problem such as the priors and the true regression function and studying various performance indicators such as the squared error risk, the posterior on the variance etc.

A Bayesian's (and our) first instinct would be to learn η itself in a Bayesian manner. Yet this does not solve the problem, as we show in Section 5.4, where we consider a setting in which $1/\eta$ turns out to be exactly equivalent to the λ regularization parameter in the Bayesian Lasso and ridge regression approaches. We find that selecting η by (empirical) Bayes, as suggested by e.g. Park and Casella (2008), does not nearly regularize enough in our misspecification experiments. Instead, the SafeBayesian method learns η in a *prequential* fashion, finding the η which minimizes a sequential prediction error on the data. This would still be very similar to Bayesian learning of η if the error were measured in terms of the standard logarithmic score, but SafeBayes, which comes in two versions, uses a 'randomized' (R-log-SafeBayes) and an 'in-model' (I-log-SafeBayes) modification of log-score instead (Section 4.2). In the supplementary material we compare R- and I-log-SafeBayes to other existing methods for determining η: Section C.1 provides an illuminating comparison to leave-one-out cross-validation as used in the frequentist Lasso, and Section F briefly considers approaches from the recent Bayesian literature Bissiri et al. (2016); Holmes and Walker (2017); Miller and Dunson (2015); Syring and Martin (2017).

The type of misspecification The models are misspecified in that they make the standard assumption of homoskedasticity — σ^2 is independent of X — whereas in reality, under P^*, there is heteroskedasticity, there being a region of X with low and a region with (relatively) high variance. Specifically, in our simplest experiment the 'true' P^* is defined as follows: at each i, toss a fair coin. If the coin lands heads, then sample X_i from a uniform distribution on $[-1, 1]$, and set $Y_i = 0 + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_0^2)$. If the coin lands tails, then set $(X_i, Y_i) = (0, 0)$, so that there is no variance at all. The 'best' conditional density $\tilde{P}$, closest to $P^*(Y \mid X)$ in KL divergence, representing the true regression function $Y = 0$ and moreover 'reliable' in the sense of Section 2.3, is then given by (1) with all β's set to 0 and $\tilde{\sigma}^2 = \sigma_0^2/2$. In a typical sample of length n, we will thus have approximately $n/2$ points with X_i uniform and Y_i normal with mean 0, and approximately $n/2$ points with $(X_i, Y_i) = (0, 0)$. These points seem 'easy' since they lie exactly on the regression function one would hope to learn; but they really wreak severe havoc.

Heteroskedasticity, but no outliers While it is well-known that in the presence of outliers, Gaussian assumptions on the noise lead to problems, both for frequentist and Bayesian procedures, in the present problem we have 'in-liers' rather than outliers. Also, if we slightly modify the setup so that homoskedasticity holds, standard Bayes starts behaving excellently, as again depicted in Figures 1 and 2. Finally, while the figure shows

[1] An R package `SafeBayes` which implements the method for Bayesian ridge and Lasso Regression (De Heide, 2016a) is available at the Comprehensive R Archive Network (CRAN).

what happens for polynomials, we get essentially the same result with trigonometric basis functions; in the experiments reported in this paper, we used independent multivariate X's rather than nonlinear basis functions, again getting essentially the same results. In the technical report (Grünwald and Van Ommen, 2014) ([GvO] from now on) we additionally performed numerous variations on the experiments in this paper, varying priors and ground truths, and always getting qualitatively the same results. All this indicates that the inconsistency is really caused by misspecification, in particular the presence of in-liers, and not by anything else. We also note that our results are entirely different from the well-known Bayesian inconsistency results of Diaconis and Freedman (1986): whereas their results are based on a well-specified model having exponentially small prior mass in KL-neighbourhoods of the true P^*, our results hold for a misspecified model, but the 'pseudo-truth' $\tilde{P}$ can have a large prior mass (any point mass < 1 is sufficient to get our results); see also Section B in the supplement.

Three remarks before we start We stress at the outset that, since this is experimental work and we are bound to experiment with finite sets of models ($p_{\max} < \infty$) and finite sample sizes n, we do not mathematically show formal inconsistency. Yet, as we explain in detail in Conclusion 2 in Section 5.2, our experiments with varying $p_{\max}$ and n strongly suggest that, if we could examine $p_{\max} = \infty$ and $n \to \infty$, then actual inconsistency will take place. Additional evidence (though of course, no proof) is provided by the fact that one of the weakest existing conditions that guarantee consistency under misspecification (De Blasi and Walker, 2013) does not hold for our model; see Section G.2 in the supplementary material. On the other hand, by checking existing consistency results for well-specified models one finds that, if one of the submodels $\mathcal{M}_p, p < p_{\max}$, is correct, then taking a prior over infinitely many models, $p_{\max} = \infty$, poses neither a problem for consistency nor for rates of convergence. Since, in addition, Grünwald and Langford (2004, 2007) did prove mathematically that consistency arises in a closely related (also featuring in-liers) but more artificial classification problem, we decided to call the phenomenon we report 'inconsistency' — but even if one thinks this term is not warranted, there remains a serious problem for finite sample sizes.

We also stress that, although both our experiments (as e.g. in Figure 2) and the implementation details of SafeBayes suggest a predictive–sequential setting, our results are just as relevant for the nonsequential setting of fixed-sample size linear regression with random design, which is a standard statistical problem. In such settings, one would like to have guarantees which, for the fixed, given sample size n, give some indication as to how 'close' our inferred distribution or parameter vector is from some 'true' or optimal vector. For example, the distance between the curve for 'Bayes, model wrong' and the curve for the true regression function at each fixed n on the x-axis in Figure 2 can be re-interpreted as the squared L_2-distance between the Bayes estimator of the regression function and the true regression function $\mathbf{0}$.

Finally, we stress that, if we modify the setup so that the 'easy' points $(0, 0)$ are at a different location, and have themselves a small variance, and the underlying regression function is not 0 everywhere but rather another function in the model, then all the phenomena we report here persist, albeit at a smaller scale (we performed additional experiments to this end in [GvO]; see also Section 6 and (Syring and Martin,

2017)). Also, recent work (De Heide, 2016b) reports on several real-world data sets for which SafeBayes substantially outperforms standard Bayes. This suggests that the phenomenon we uncovered is not merely a curiosity, and can really affect Bayesian inference in practice.

Contents and structure of this paper In Section 2, introduces our setting and the main concepts needed to understand our results, including the η-generalized posterior, and instantiates these to the linear model. In Section 3, we explain *how* inconsistency can arise under misspecification (essentially the only possible cause is 'bad misspecification' along with 'hypercompression'). Section 4 explains a potential solution, the generalized and SafeBayesian methods, and explains why they work. Section 5 discusses our experiments in detail. Section 6 provides an 'executive summary' of the experiments in this paper and the many additional experiments on which we report in the technical report [GvO]. In all experiments SafeBayesian methods behave much better in terms of squared error risk and reliability than standard Bayes if the model is incorrect, and hardly worse (sometimes still better) than standard Bayes if the model is correct.

Supplementary material Apart from inconsistency, there is one other issue with Bayes under misspecification: our inference task(s) of interest may not be associated with the KL-optimal $\tilde{P}$. We discuss this problem in the (main) Appendix B in the supplementary material, and show how adopting a Gibbs likelihood (to which we can then apply SafeBayes) sometimes, but not always solves the problem. On the other hand, we also discuss how SafeBayes can sometimes even help with well-specified models. We also discuss related work, pose several *Open Problems* and tentatively propose a generic theory of (pseudo-Bayesian) inference of misspecification, parts of which have already been developed in the companion papers Grünwald and Mehta (2016) and Grünwald (2017).

2 Preliminaries

We consider data $Z^n = Z_1, Z_2, \ldots, Z_n \sim$ i.i.d. P^*, where each $Z_i = (X_i, Y_i)$ is an independently sampled copy of $Z = (X, Y)$, X taking values in some set $\mathcal{X}$, Y taking values in $\mathcal{Y}$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We are given a *model* $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$ parameterized by (possibly infinite-dimensional) Θ, and consisting of conditional distributions $P_\theta(Y \mid X)$, extended to n outcomes by independence. For simplicity we assume that all P_θ have corresponding conditional densities f_θ, and similarly, the conditional distribution $P^*(Y \mid X)$ has a conditional f^*, all with respect to the same underlying measure. While we do not assume $P^*(Y \mid X)$ to be in (or even 'close' to) $\mathcal{M}$, we want to learn, from given data Z^n, a 'best' (in a sense to be defined below) element of $\mathcal{M}$, or at least, a distribution on elements of $\mathcal{M}$ that can be used to make adequate predictions about future data. While our experiments focus on linear regression, the discussion in this section holds for general conditional density models. The logarithmic score, henceforth abbreviated to *log-loss*, is defined in the standard manner: the loss incurred when predicting Y based on density $f(\cdot \mid x)$ and Y takes on value y, is given by $-\log f(y \mid x)$. A central quantity in our setup is then the *expected log-loss* or *log-risk*, defined as

$$\text{RISK}^{\log}(\theta) := \mathbf{E}_{(X,Y)\sim P^*}[-\log f_\theta(Y \mid X)]. \tag{2}$$

2.1 KL-optimal distribution

We let P^*_X be the marginal distribution of X under P^*. The *Kullback–Leibler (KL) divergence* $D(P^*\|P_\theta)$ between P^* and conditional distribution P_θ is defined as the expectation, under $X \sim P^*_X$, of the KL divergence between P_θ and the 'true' conditional $P^*(Y \mid X)$: $D(P^*\|P_\theta) = \mathbf{E}_{X\sim P^*_X}[D(P^*(\cdot \mid X) \,\|\, P_\theta(\cdot \mid X))]$. A simple calculation shows that for any θ, θ',

$$D(P^*\|P_\theta) - D(P^*\|P_{\theta'}) = \textsc{risk}^{\log}(\theta) - \textsc{risk}^{\log}(\theta'),$$

so that the closer P_θ is to P^* in terms of KL divergence, the smaller its log-risk, and the better it is, on average, when used for predicting under the log-loss.

Now suppose that $\mathcal{M}$ contains a unique distribution that is closest, among all $P \in \mathcal{M}$ to P^* in terms of KL divergence. We denote such a distribution, if it exists, by $\tilde{P}$. Then $\tilde{P} = P_\theta$ for at least one $\theta \in \Theta$; we pick any such θ and denote it by $\tilde{\theta}$, i.e. $\tilde{P} = P_{\tilde{\theta}}$, and note that it also minimizes the log-risk:

$$\textsc{risk}^{\log}(\tilde{\theta}) = \min_{\theta\in\Theta} \textsc{risk}^{\log}(\theta) = \min_{\theta\in\Theta} \mathbf{E}_{(X,Y)\sim P^*}[-\log f_\theta(Y \mid X)]. \tag{3}$$

We shall call such a $\tilde{\theta}$ *(KL-)optimal*.

Since, in regions of about equal prior density, the log Bayesian posterior density is proportional to the log likelihood ratio, we hope that, given enough data, with high P^*-probability, the posterior puts most mass on distributions that are close to $P_{\tilde{\theta}}$ in KL divergence, i.e. that have log-risk close to optimal. Indeed, all existing consistency theorems for Bayesian inference under misspecification express concentration of the posterior around $P_{\tilde{\theta}}$. While the minimum KL divergence point is not always of intrinsic interest, for some (not all) models, $\tilde{P}$ can be of interest for other reasons as well (Royall and Tsou, 2003): there may be *associated* inference tasks for which $\tilde{P}$ is also suitable. Examples of associated prediction tasks for the linear model are given in Section 2.3; we further consider non-associated tasks such as absolute loss in Appendix B.

2.2 A special case: The linear model

Fix some $p_{\max} \in \{0, 1, \dots\} \cup \{\infty\}$. We observe data $Z_1, \dots, Z_n$ where $Z_i = (X_i, Y_i)$, $Y_i \in \mathbf{R}$ and $X_i = (1, X_{i1}, \dots, X_{ip_{\max}}) \in \mathbf{R}^{p_{\max}+1}$. Note that this is as in (1) but from now on we adopt the standard convention to take $X_{i0} \equiv 1$ as a dummy random variable. We denote by $\mathcal{M}_p = \{P_{p,\beta,\sigma^2} \mid (p, \beta, \sigma^2) \in \Theta_p\}$ the standard linear model with parameter space $\Theta_p := \{(p, \beta, \sigma^2) \mid \beta = (\beta_0, \dots, \beta_p)^\top \in \mathbf{R}^{p+1}, \sigma^2 > 0\}$, where the entry p in (p, β, σ^2) is redundant but included for notational convenience. We let $\Theta = \bigcup_{p=0,\dots,p_{\max}} \Theta_p$. $\mathcal{M}_p$ states that for all i, (1) holds, where $\epsilon_1, \epsilon_2, \dots \sim$ i.i.d. $N(0, \sigma^2)$. When working with linear models $\mathcal{M}_p$, we are usually interested in finding parameters β that predict well in terms of the *squared error loss function* (henceforth abbreviated to *square-loss*): the square-loss on data (X_i, Y_i) is $(Y_i - \sum_{j=0}^p \beta_j X_{ij})^2 = (Y_i - X_i\beta)^2$. We thus want to find the distribution minimizing the expected square-loss, i.e. *squared error risk* (henceforth abbreviated to 'square-risk') relative to the underlying P^*:

$$\text{RISK}^{\text{sq}}(p,\beta) := \mathbf{E}_{(X,Y)\sim P^*}(Y - \mathbf{E}_{p,\beta,\sigma^2}[Y \mid X])^2 = \mathbf{E}_{(X,Y)\sim P^*}(Y - \sum_{j=0}^{p} \beta_j X_j)^2, \quad (4)$$

where $\mathbf{E}_{p,\beta,\sigma^2}[Y \mid X]$ abbreviates $\mathbf{E}_{Y\sim P_{p,\beta,\sigma^2}|X}[Y]$. Since this quantity is independent of the variance σ^2, σ^2 is not used as an argument of RISK^{sq}.

2.3 KL-associated prediction tasks for the linear model: Optimality; reliability

Suppose that an optimal $\tilde{P} \in \mathcal{M}$ exists in the regression model. We denote by $\tilde{p}$ the smallest p such that $\tilde{P} \in \mathcal{M}_p$, and define $\tilde{\sigma}^2, \tilde{\beta}$ such that $\tilde{P} = P_{\tilde{p},\tilde{\beta},\tilde{\sigma}^2}$. A straightforward computation shows that for all $(p,\beta,\sigma^2) \in \Theta$:

$$\text{RISK}^{\log}((p,\beta,\sigma^2)) = \frac{1}{2\sigma^2}\text{RISK}^{\text{sq}}((p,\beta)) + \frac{1}{2}\log(2\pi\sigma^2), \quad (5)$$

so that the (p,β) achieving minimum log-risk for each fixed σ^2 is equal to the (p,β) with the minimum square-risk. In particular, $(\tilde{p},\tilde{\beta},\tilde{\sigma}^2)$ must minimize not just log-risk, but also square-risk. Moreover, the conditional expectation $\mathbf{E}_{P^*}[Y \mid X]$ is known as the *true regression function*. It minimizes the square-risk among all conditional distributions for $Y \mid X$. Together with (5) this implies that, if there is some (p,β) such that $\mathbf{E}[Y \mid X] = \sum_{j=0}^{p} \beta_j X_j = X\beta$, i.e. (p,β) represents the true regression function, then $(\tilde{p},\tilde{\beta})$ also represents the true regression function. In all our examples, this will be the case: the model is misspecified only in that the true noise is heteroskedastic; but the model does invariably contain the true regression function.

Moreover, for each fixed (p,β), the σ^2 minimizing $\text{RISK}^{\log}$ is, as follows by differentiation, given by $\sigma^2 = \text{RISK}^{\text{sq}}(p,\beta)$. In particular, this implies that

$$\tilde{\sigma}^2 = \text{RISK}^{\text{sq}}(\tilde{p},\tilde{\beta}), \quad (6)$$

or in words: the KL-optimal model variance $\tilde{\sigma}^2$ is equal to the true expected (marginal, not conditioned on X) square-risk obtained if one predicts with the optimal $(\tilde{p},\tilde{\beta})$. This means that the optimal $(\tilde{p},\tilde{\beta},\tilde{\sigma}^2)$ is *reliable* in the sense of Grünwald (1998, 1999): its self-assessment about its square-loss performance is correct, independently of whether $\tilde{\beta}$ is equal to the true regression function or not. In other words, $(\tilde{p},\tilde{\beta},\tilde{\sigma}^2)$ *correctly predicts how well it predicts in the squared-error sense.*

Summarizing, for misspecified models, $(\tilde{p},\tilde{\beta},\tilde{\sigma}^2)$ is optimal not just in KL/log-risk sense, but also in terms of square-risk and in terms of reliability; in our examples, it also represents the true regression function. We say that, for linear models, square-risk optimality, square-risk reliability and regression-function consistency are *KL-associated prediction tasks*: if we can find the KL-optimal $\tilde{\theta}$, we automatically behave well in these associated tasks as well. Thus, whenever one is prepared to work with linear models and one is interested in squared error risk or reliability, then Bayesian inference would seem the way to go, even if one suspects misspecification... at least if there is consistency.

2.4 The generalized posterior

General losses The original 'generalized' or 'Gibbs' posterior is a notion going back at least to Vovk (1990) and has been developed mainly within the so-called (frequentist) *PAC-Bayesian* framework (McAllester, 2003; Seeger, 2002; Catoni, 2007; Audibert, 2004; Zhang, 2006b; see also Jiang and Tanner (2008), Bissiri et al. (2016) and the extensive discussion in the supplementary material). It is defined relative to a prior on *predictors* rather than probability distributions. Depending on the decision problem at hand, predictors can be e.g. classifiers, regression functions or probability densities. Formally, we are given an abstract space of predictors represented by a set Θ, which obtains its meaning in terms of a loss function $\ell : \mathcal{Z} \times \Theta \to \mathbf{R}$, writing $\ell_\theta(z)$ as shorthand for $\ell(z, \theta)$. Following e.g. Zhang (2006b), for any prior Π on Θ with density π relative to some underlying measure ρ, we define the *generalized Bayesian posterior with learning rate η relative to loss function ℓ*, denoted as $\Pi \mid Z^n, \eta$, as the distribution on Θ with density

$$\pi(\theta \mid z^n, \eta) := \frac{e^{-\eta \sum_{i=1}^n \ell_\theta(z_i)} \pi(\theta)}{\int e^{-\eta \sum_{i=1}^n \ell_\theta(z_i)} \pi(\theta) \rho(d\theta)} = \frac{e^{-\eta \sum_{i=1}^n \ell_\theta(z_i)} \pi(\theta)}{\mathbf{E}_{\theta \sim \Pi}[e^{-\eta \sum_{i=1}^n \ell_\theta(z_i)}]}. \tag{7}$$

Thus, if θ_1 fits the data better than θ_2 by a difference of ϵ according to loss function ℓ, then their posterior ratio is larger than their prior ratio by an amount exponential in ϵ, where the larger η, the larger the influence of the data as compared to the prior.

Log-loss and likelihood Now consider the case that the set Θ represents a model of (conditional) distributions $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$. Then we may set $\ell_\theta(z_i) = -\log f_\theta(y_i \mid x_i)$ to be the log-loss as defined above. The definition of η-generalized posterior now specializes to the definition of 'generalized posterior' (in this context also called "fractional posterior") as known within the Bayesian literature (Walker and Hjort, 2002; Zhang, 2006a; Martin et al., 2017):

$$\pi(\theta \mid z^n, \eta) = \frac{(f(y^n \mid x^n, \theta))^\eta \pi(\theta)}{\int (f(y^n \mid x^n, \theta))^\eta \pi(\theta) \rho(d\theta)} = \frac{(f(y^n \mid x^n, \theta))^\eta \pi(\theta)}{\mathbf{E}_{\theta \sim \Pi}[(f(y^n \mid x^n, \theta))^\eta]}, \tag{8}$$

where here as in the remainder we use the notation $f(\cdot \mid \theta)$ and $f_\theta(\cdot)$ interchangeably. Obviously $\eta = 1$ corresponds to standard Bayesian inference, whereas if $\eta = 0$ the posterior is equal to the prior and nothing is ever learned. Our algorithm for learning η will usually end up with values in between. The rationale behind taking $\eta < 1$ even if the model is well-specified is discussed in Section F.2. A connection to misspecification was first made by Grünwald (2011) (see Section F) and Grünwald (2012). In the literature (7) is often called a 'Gibbs posterior'; whenever no confusion can arise, we will use the phrase 'generalized posterior' to refer to both (7) and (8).

Generalized predictive distribution We also define the predictive distribution based on the η-generalized posterior (8) as a generalization of the standard definition as follows: for $m \geq 1, m' \geq m$, we set

$$\bar{f}(y_{i+1}, \ldots, y_{i+m} \mid x_{i+1}, \ldots, x_{i+m'}, z^i, \eta)$$

$$\begin{aligned} &:= \mathbf{E}_{\theta\sim\Pi|z^i,\eta}[f(y_{i+1},\ldots,y_{i+m} \mid x_i,\ldots,x_{i+m'},\theta)] \\ &= \mathbf{E}_{\theta\sim\Pi|z^i,\eta}[f(y_{i+1},\ldots,y_{i+m} \mid x_i,\ldots,x_{i+m},\theta)], \end{aligned} \tag{9}$$

where the first equality is a definition and the second follows by our i.i.d. assumption. We always use the bar-notation $\bar{f}$ to indicate marginal and predictive distributions, i.e. distributions on data that are arrived at by integrating out parameters. If $\eta = 1$ then $\bar{f}$ and π become the standard Bayesian predictive density and posterior, and if it is clear from the context that we consider $\eta = 1$, we leave out the η in the notation.

2.5 Instantiating generalized Bayes to linear model selection and averaging

Now consider again a linear model $\mathcal{M}_p$ as defined in Section 2.3. We instantiate the generalized posterior and its marginals for this model. With prior $\pi(\beta, \sigma^2 \mid p)$ taken relative to Lebesgue measure, (8) specializes to:

$$\pi(\beta, \sigma \mid z^n, p, \eta) = \frac{(2\pi\sigma^2)^{-n\eta/2} e^{-\frac{\eta}{2\sigma^2}\sum_{i=1}^n (y_i - x_i\beta)^2} \pi(\beta, \sigma \mid p)}{\int (2\pi\sigma^2)^{-n\eta/2} e^{-\frac{\eta}{2\sigma^2}\sum_{i=1}^n (y_i - x_i\beta)^2} \pi(\beta, \sigma \mid p)\, d\beta\, d\sigma}.$$

In the numerator $1/\sigma^2$ and η are interchangeable in the exponent, but not in the factor in front: their role is subtly different. For Bayesian inference with a sequence of models $\mathcal{M} = \bigcup_{p=0,\ldots,p_{\max}} \mathcal{M}_p$, with $\pi(p)$ a probability mass function on $p \in \{0, \ldots, p_{\max}\}$, we get

$$\pi(\beta, \sigma, p \mid z^n, \eta) = \frac{(2\pi\sigma^2)^{-n\eta/2} e^{-\frac{\eta}{2\sigma^2}\sum_{i=1}^n (y_i - x_i\beta)^2} \pi(\beta, \sigma \mid p)\pi(p)}{\sum_{p=0}^{p_{\max}} \int (2\pi\sigma^2)^{-n\eta/2} e^{-\frac{\eta}{2\sigma^2}\sum_{i=1}^n (y_i - x_i\beta)^2} \pi(\beta, \sigma \mid p)\pi(p)\, d\beta\, d\sigma}. \tag{10}$$

The total generalized posterior probability of model $\mathcal{M}_p$ then becomes:

$$\pi(p \mid z^n, \eta) = \int \pi(\beta, \sigma, p \mid z^n, \eta)\, d\beta\, d\sigma. \tag{11}$$

The previous displays held for general priors. The experiments in this paper adopt widely used priors (see e.g. Raftery et al., 1997): normal priors on the β's and inverse gamma priors on the variance. These conjugate priors allow explicit analytical formulas for all relevant quantities for arbitrary η. We consider here the simple case of a fixed $\mathcal{M}_p$; the more complicated formulas with an additional prior on p are given in [GvO]. Let $\mathbf{X}_n = (x_1^\top, \ldots, x_n^\top)^\top$ be the design matrix, let the initial Gaussian prior on β conditional on σ^2 be given by $N(\bar{\beta}_0, \sigma^2\Sigma_0)$, and the prior on σ^2 by $\pi(\sigma^2) = \text{Inv-gamma}(\sigma^2 \mid a_0, b_0)$ for some a_0 and b_0. Here we use the following parameterization of the inverse gamma distribution:

$$\text{Inv-gamma}(\sigma^2 \mid a, b) = \sigma^{-2(a+1)} e^{-b/\sigma^2} b^a / \Gamma(a). \tag{12}$$

Then the generalized posterior on β is again Gaussian with mean

$$\bar{\beta}_{n,\eta} := \mathbf{E}_{\beta\sim\Pi|z^n,p,\eta}\,\beta = \Sigma_{n,\eta}(\Sigma_0^{-1}\bar{\beta}_0 + \eta\mathbf{X}_n^\top y^n) \tag{13}$$

and covariance matrix $\sigma^2\Sigma_{n,\eta}$, where $\Sigma_{n,\eta} = (\Sigma_0^{-1} + \eta\mathbf{X}_n^\top\mathbf{X}_n)^{-1}$ (note that the posterior mean of β given σ^2 does not depend on σ^2; also note that for $\eta = 1$, this is the standard posterior); the generalized posterior $\pi(\sigma^2 \mid z^n, p, \eta)$ is given by Inv-gamma$(\sigma^2 \mid a_{n,\eta}, b_{n,\eta})$ where $a_{n,\eta} = a_0 + \eta n/2$ and $b_{n,\eta} = b_0 + \frac{\eta}{2}\sum_{i=1}^n (y_i - x_i\bar{\beta}_{n,\eta})^2$. The posterior expectation of σ^2 can be calculated as

$$\bar{\sigma}^2_{n,\eta} := \frac{b_{n,\eta}}{a_{n,\eta} - 1}. \tag{14}$$

3 Bayesian inconsistency from bad misspecification

In this section and the next, we provide the necessary background on Bayesian inconsistency under misspecification. We first explain (Section 3.1) *when* it can arise and we then explain (Section 3.2) *why* it arises. Section 4.1 explains how a different learning rate can solve the problem, and Section 4.2 introduces SafeBayes and explains how it can find this learning rate. We focus on generalized Bayes with standard likelihoods (8), but stress that analogous problems arise with Gibbs posteriors (7), as explained in Appendix B.

3.1 Preparation: Benign vs. bad misspecification

The first thing to understand what goes on is to distinguish between two types of misspecification. The difference is depicted in cartoon fashion in Figure 3. In the figure, $\tilde{P} = \arg\min_{P\in\mathcal{M}} D(P^*\|P)$ is the distribution in model $\mathcal{M}$ that minimizes KL divergence to the 'true' P^* but, since the model is nonconvex, the distribution $\bar{\tilde{P}}$ that minimizes KL divergence to P^* within the convex hull of $\mathcal{M}$ may be very different from $\tilde{P}$. This means that also the Bayes predictive distribution $\bar{P}(Y_i \mid X_i, Z^{i-1})$ based on Z^{i-1}, with density as given by (9) with $\eta = 1$ and $m = 1$, may happen to be very different from any $P \in \mathcal{M}$, and in fact, closer to P^* than the KL-optimal $\tilde{P}$. If, as in the picture, P^* is such that $\inf_{P\in\mathcal{M}} D(P^*\|P)$ decreases if the infimum is taken over the convex hull of $\mathcal{M}$, then we speak of 'bad misspecification'; otherwise (e.g. if Q^* rather than P^* was the true distribution, so that $\tilde{Q}$ reached the minimum) the misspecification is 'benign'. We will see in the next subsection that inconsistency (posterior not concentrating near $\tilde{P}$) happens if and only if the Bayes predictive $\bar{P}(Y_i \mid X_i, Z^{i-1})$ is KL-closer to P^* than $\tilde{P}$ at many i, which in turn can happen only if we have bad misspecification. Figure 4 illustrates the strong potential for bad misspecification with our regression model. Two remarks are in order: (a) for convex probability models, one can only have benign misspecification. (b) Our regression model may seem convex since it is convex at the level of regression coefficients, but, as we illustrate further below, it is *not* convex at the level of conditional densities, which is what matters.

3.2 Hypercompression

A paradox? We now explain in more detail what can (and does, in our experiments) happen under bad misspecification. We first note that there does exist an almost

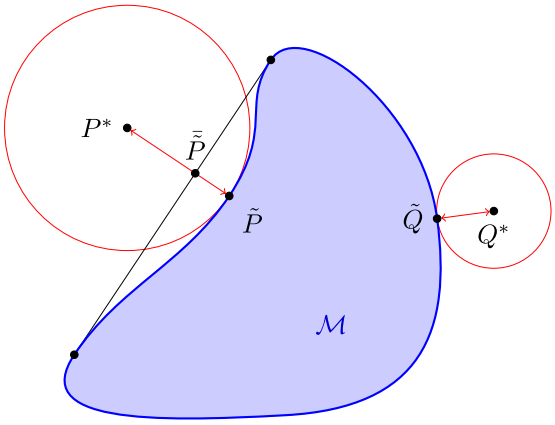

Figure 3: Benign vs. bad misspecification.

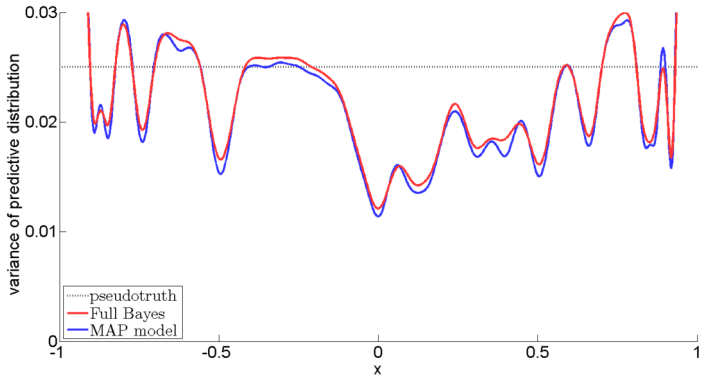

Figure 4: Variance of standard Bayes predictive distribution conditioned on a new input S as a function of S after 50 examples for the polynomial model-wrong experiment (Figure 1), shown both for the predictive distribution based on the full, model-averaging posterior and for the posterior conditioned on the MAP model $\mathcal{M}_{\breve{p}_{\text{map}}}$. For both posteriors, the posterior mean of Y is incorrect for $X \neq 0$, yet $\bar{f}(Y \mid Z^{50}, X)$ still achieves small risk because of its small variance at $X = 0$.

condition-free 'consistency-like' result for Bayesian inference that even holds under misspecification, but it is different from standard consistency results in an essential way. This result, which in essence goes back to Barron (1998), says that, for any i.i.d. model $\mathcal{M}$, under no further conditions, for all n, the following holds:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(D(P^*\|\bar{P}(\cdot \mid Z^{i-1})) - D(P^*\|P_{\hat{\theta}})\right)\right] \tag{15}$$

$$= \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} \text{RISK}^{\log}(\bar{P}(\cdot \mid Z^{i-1}))\right] - \text{RISK}^{\log}(\tilde{\theta})$$

$$= \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} \big(-\log \bar{f}(Y_i \mid X_i, Z^{i-1})\big) - \big(-\log f_{\tilde{\theta}}(Y_i \mid X_i)\big)\right] \leq \text{SMALL}_n, \tag{16}$$

where the expectation is over $Z^n \sim P^*$. Here, extending (2), we used the notation $\text{RISK}^{\log}(\bar{P}(\cdot \mid Z^{i-1})) = \mathbf{E}_{(X_i,Y_i)\sim P^*}[-\log \bar{f}(Y_i \mid X_i, Z^{i-1})]$. The equalities are (essentially trivial, see (Grünwald, 2007)) rewritings; the real meat is in the final inequality. SMALL_n depends on the amount of prior mass in neighbourhoods of $\tilde{P}$, and in our case, is on the order of $(\log n)/n$ with a small constant in front (details in Section D.1 of the Supplementary Material). This implies that at most sample sizes i, $D(P^*\|\bar{P}(\cdot \mid Z^{i-1}))$ must be of order $1/i$, even if the model is misspecified; thus, at least in a time-averaged sense, the KL divergence between P^* and the Bayes predictive distribution converges at a fast rate of order $1/n$ to the smallest value attainable within model $\mathcal{M}$, or becomes even smaller. However, in the experiment of Figure 1, we see that Bayes is not putting significant posterior mass near the pseudo-true parameter $\tilde{\theta}$ at most n. Given (15)–(16), at first this may seem paradoxical or even impossible, but of course there is an explanation: because we have *bad misspecification* as in Figure 3, it can in fact happen that many terms $D(P^*\|\bar{P}(\cdot \mid Z^{i-1})) - D(P^*\|P_{\tilde{\theta}})$ in the sum in (15) are *negative*. Then Barron's bound (15)–(16) may be satisfied, not because the posterior concentrates on distributions close to $\tilde{P}$, but rather because — at many sample sizes i — it puts its mass on distributions which are very far from $\tilde{P}$, but mixed together are closer in KL-divergence to P^* than is $\tilde{P}$. As long as the prior puts sufficient mass near $\tilde{P}$ and the data and model are i.i.d., this is the *only* way in which inconsistency under misspecification can occur. By the law of large numbers, if (and only if) a substantial fraction of the terms in (15) are negative, we would also expect that the log-loss of predicting Y_i given X_i using the Bayes predictive is often lower than the log-loss of predicting Y_i given X_i using the KL-optimal $\tilde{P}$; in other words, we expect many of the terms on the left in (16) to be negative as well. As we show in Section 5.2, Figure 7, this indeed happens in our experiments — in such an extreme form that for $n < 100$ the entire sum in (16) is negative by a fair amount. If this sum is negative for empirical data, we say that *hypercompression* takes place. The name is borrowed from information theory — readers familiar with this field will recognize the sum over the $-\log f_{\tilde{\theta}}(Y_i \mid X_i)$ as the number of bits needed to code the n observed y-values given the x-values under the code which would be optimal to use in expectation if the data were sampled from $P_{\tilde{\theta}}$. The sum in (16) being negative implies that this code is outperformed by another code on the empirical data, something which is next to impossible if the model is correct — this is quantified by the *no-hypercompression inequality* (Grünwald, 2007) which we repeat in Section 5.3 to help interpret our results.

If we have hypercompression, then $\bar{P}(\cdot \mid Z^i)$ will be close in KL divergence to P^*. Small KL-divergence implies small log-risk, and in Section 2.3 was also related to small square-risk and good performance on other KL-associated prediction tasks. This at first sight may seem like another paradox: in Figure 2 we saw very large square-risk of the Bayes predictive. Again, this is no contradiction: the relation between log-risk and

square-risk (5) does not hold for arbitrary distributions, but only for members of the model. If, as in the figure, $\bar{P}(\cdot \mid Z^{i-1})$ is outside the model (in fact, it differs significantly from any element of the model), small KL-divergence is no longer an indication that $\bar{P}(\cdot \mid Z^{i-1})$ will also give good results for KL-associated prediction tasks.

To see where the hypercompression in our regression example comes from, note first that our model is not convex: the conditional densities indexed by θ are normals with mean $X\beta$ and fixed variance σ^2 for each given X; a mixture of two such conditional normals can be bimodal and hence is itself not a conditional normal, hence the model is not convex. In our setting the predictive is a mixture of infinitely many conditional normals. Its conditional density is a mixture of t-distributions, whose variance highly depends on X, thus making the highly heteroskedastic predictive very different from any of the — homoskedastic — distributions in the model. The striking difference is plotted in Figure 4. Hypercompression occurs because at $X = 0$, the variance of the predictive is smaller than $\tilde{\sigma}^2$, which substantially decreases the log-risk.

4 The Solution: How $\eta \ll 1$ can help, and how SafeBayes finds it

4.1 How η-generalized Bayes for $\eta \ll 1$ can avoid bad misspecification

We start with a re-interpretation of the η-generalized posterior: for small enough η, *it is formally equivalent to a standard posterior based on a modified joint probability model*[2]. Let $f^*(x, y)$ be the density of the true distribution P^* on (X, Y). Formally, we define the *η-reweighted distributions* $P^{(\eta)}$ as joint distributions on (X, Y) with densities $f^{(\eta)}$ given by

$$f^{(\eta)}(x, y \mid \theta) = f^*(x, y) \cdot \left(\frac{f(y \mid x, \theta)}{f(y \mid x, \tilde{\theta})} \right)^{\eta}, \tag{17}$$

extended to n outcomes by independence. Now, as follows from (Van Erven et al., 2015, Example 3.7), in our setting[3] there exists a critical value of $\bar{\eta}$ such that if we take any $0 < \eta \leq \bar{\eta}$, then for every $\theta \in \Theta$, $P^{(\eta)}(\cdot \mid \theta)$ is a (sub-) probability distribution, i.e. for all $\theta \in \Theta$,

$$\int \int f^{(\eta)}(x, y \mid \theta) dx dy \leq 1. \tag{18}$$

If for some $\theta \in \Theta$, (18) is strictly smaller than 1, say $1 - \epsilon$, then the corresponding $P^{(\eta)}(\cdot \mid \theta)$ can be thought of as a standard probability distribution by defining it on extended outcome space $(\mathcal{X} \times \mathcal{Y}) \cup \{\square\}$ and assuming that it puts mass ϵ on the special outcome $\square$ which in reality will never actually occur. We can thus think of $\mathcal{M}^{(\eta)} := \{P^{(\eta)}(\cdot \mid \theta) \mid \theta \in \Theta\}$ as a standard probability model. One immediately verifies that,

[2] In this explicit form, this insight is new and cannot be found in any of the earlier papers on generalized or safe Bayes, although the reweighted probabilities that we now define can be found in e.g. Van Erven et al. (2015).

[3] The story still goes through with some modifications if $\bar{\eta} = 0$ (Grünwald and Mehta, 2016).

for every $\eta > 0$, $D(P^* \| P^{(\eta)}(\cdot \mid \theta))$ is minimized for $\tilde{\theta}$, just as before — $P^{(\eta)}(\cdot \mid \tilde{\theta})$ now being equal to the 'true' P^* (!). By Bayes' theorem, with prior π on model $\mathcal{M}^{(\eta)}$, the posterior probability is given by:

$$\pi(\theta \mid z^n) = \frac{f^{(\eta)}(x^n, y^n \mid \theta) \cdot \pi(\theta)}{\int f^{(\eta)}(x^n, y^n \mid \theta)\pi(\theta)\rho(d\theta)} = \frac{\prod_{i=1}^n f^*(x_i, y_i) \cdot \left(\frac{f(y_i|x_i,\theta)}{f(y_i|x_i,\tilde{\theta})}\right)^\eta \cdot \pi(\theta)}{\int \prod_{i=1}^n f^*(x_i, y_i) \cdot \left(\frac{f(y_i|x_i,\theta)}{f(y_i|x_i,\tilde{\theta})}\right)^\eta \pi(\theta)\rho(d\theta)}$$
$$= \frac{\prod_{i=1}^n f(y_i \mid x_i, \theta)^\eta \cdot \pi(\theta)}{\int \prod_{i=1}^n f(y_i \mid x_i, \theta)^\eta \pi(\theta)\rho(d\theta)}, \tag{19}$$

which is seen to coincide with (8). Thus, as promised, for any $0 < \eta \leq \bar{\eta}$, we can equivalently think of the generalized posterior as a standard posterior on a different model. But for such a value of η, our use of generalized Bayesian updating is equivalent to using Bayes' theorem in the standard way with a correctly specified probability model (because $P(\cdot \mid \tilde{\theta}) = P^*$), and hence standard consistency and rate of convergence results such those by Ghosal et al. (2000) kick in, and convergence of the posterior must take place. We can also see this in terms of Barron's result, (15)–(16), which must also hold for the model $\mathcal{M}^{(\eta)}$, i.e. if we replace the standard predictive distribution $\bar{P}$ (and its density $\bar{f}$) for model $\mathcal{M}$ by the standard predictive for model $\mathcal{M}^{(\eta)}$. For this reweighted model, we have that

$$0 = D(P^* \| \tilde{P}^{(\eta)}) \leq D(P^* \| \bar{P}) \tag{20}$$

for any arbitrary mixture $\bar{P}$ of distributions in $\mathcal{M}^{(\eta)}$, and therefore also for every possible predictive distribution $\bar{P} := \bar{P}(\cdot \mid Z^i)$. This means that the terms in the sum in (15) are now all positive and (15)–(16) now *does* imply that, at most n, the Bayes predictive distribution is close to $\tilde{P}$ — so, generalized Bayes with $0 < \eta < \bar{\eta}$ should become competitive again. The 'best' value of η will typically be slightly smaller than, but not equal to $\bar{\eta}$: convergence of the posterior on reweighted probabilities $P^{(\eta)}$ of order $\text{SMALL}_n = (\log n)/n$ corresponds to a convergence of the original probabilities $P^{(1)}$ at order $(\log n)/(n\eta)$, so the price to pay for using a small η is that, although the posterior will now concentrate on the KL-optimal distribution in our model, it may take longer (by a constant factor) before this happens. The η at which the fastest convergence takes place will thus be close to $\bar{\eta}$, but in practice it may be slightly smaller, as we further explain in Appendix D.2. We proceed to address the one remaining question: how to determine $\bar{\eta}$ based on empirical data.

4.2 The SafeBayesian algorithm and How it finds the right η

We introduce SafeBayes via Dawid's prequential interpretation of Bayes factor model selection. As was first noticed by Dawid (1984) and Rissanen (1984), we can think of Bayes factor model selection as picking the model with index p that, when used for sequential prediction with a logarithmic scoring rule, minimizes the cumulative loss. To see this, note that for any distribution whatsoever, we have that, by definition of conditional probability,

$$-\log f(y^n) = -\log \prod_{i=1}^n f(y_i \mid y^{i-1}) = \sum_{i=1}^n -\log f(y_i \mid y^{i-1}).$$

In particular, for the standard Bayesian marginal distribution $\bar{f}(\cdot \mid p) = \bar{f}(\cdot \mid p, \eta = 1)$ as defined above, for each fixed p, we have

$$-\log \bar{f}(y^n \mid x^n, p) = \sum_{i=1}^{n} -\log \bar{f}(y_i \mid x^n, y^{i-1}, p) = \sum_{i=1}^{n} -\log \bar{f}(y_i \mid x_i, z^{i-1}, p), \quad (21)$$

where the second equality holds by (9). If we assume a uniform prior on model index p, then Bayes factor model selection picks the model maximizing $\pi(p \mid z^n)$, which by Bayes' theorem coincides with the model minimizing (21), i.e. minimizing cumulative log-loss. Similarly, in 'empirical Bayes' approaches, one picks the value of some parameter ρ that maximizes the marginal Bayesian probability $\bar{f}(y^n \mid x^n, \rho)$ of the data. By (21), which still holds with p replaced by ρ, this is again equivalent to the ρ minimizing the cumulative log-loss. This is the *prequential* interpretation of Bayes factor model selection and empirical Bayes approaches, showing that Bayesian inference can be interpreted as a sort of *forward* (rather than cross-) validation (Dawid, 1984; Rissanen, 1984; Hjorth, 1982).

We will now see whether we can use this approach with ρ in the role of the η for the η-generalized posterior that we want to learn from the data. We continue to rewrite (21) as follows (with ρ instead of p that can either stand for a continuous-valued parameter or for a model index but not yet for η), using the fact that the Bayes predictive distribution given ρ and z^{i-1} can be rewritten as a posterior-weighted average of f_θ:

$$\breve{\rho} := \arg\max_{\rho} \bar{f}(y^n \mid x^n, \rho) = \arg\min_{\rho} \sum_{i=1}^{n} \left(-\log \bar{f}(y_i \mid x_i, z^{i-1}, \rho)\right)$$
$$= \arg\min_{\rho} \sum_{i=1}^{n} \left(-\log \mathbf{E}_{\theta \sim \Pi \mid z^{i-1}, \rho}[f(y_i \mid x_i, \theta)]\right). \quad (22)$$

This choice for $\breve{\rho}$ being entirely consistent with the (empirical) Bayesian approach, our first idea is to choose $\hat{\eta}$ in the same way: we simply pick the η achieving (22), with ρ substituted by η. However, this will tend to pick η close to 1 and does not improve predictions under misspecification — this is illustrated experimentally in Section 5.4; see also (Grünwald and Van Ommen, 2014, Figure 13). But it turns out that a *slight* modification of (22) does the trick: we simply interchange the order of logarithm and expectation in (22) and pick the η minimizing

$$\sum_{i=1}^{n} \mathbf{E}_{\theta \sim \Pi \mid z^{i-1}, \eta} \left[-\log f(y_i \mid x_i, \theta)\right]. \quad (23)$$

In words, we pick the η minimizing the posterior-expected posterior-**R**andomized log-loss, i.e. the log-loss we expect to obtain, according to the η-generalized posterior, if we actually sample from this posterior. This modified loss function has also been called *Gibbs error* (Cuong et al., 2013); we simply call it the η-*R*-log-loss from now on.

We now give a heuristic explanation of why (22) (with $\rho = \eta$) fails while (23) works; a much more detailed explanation is in the supplementary material. The problem with (22) is that it tends to be small for η at which hypercompression takes place. By (15), at η at which bad misspecification and hence hypercompression takes place — some of the terms inside the expectation in (15) are negative — we also expect some of the

terms in (16) to be negative. In Figure 7 we will see that in our experiments, this indeed happens. (22) can thus become very small — smaller than the cumulative loss we would get if we would persistently predict with the optimal $\tilde{\theta}$ — for $\eta = 1$ or close to 1. These are the η for which we have already established that Bayes fails. In contrast, the Gibbs error (23) is small in expectation only if a substantial part of the posterior mass is assigned to θ close to $\tilde{\theta}$ in the sense that $D(P^*\|P_\theta) - D(P^*\|P_{\tilde{\theta}})$ is small. This stricter requirement clearly cannot favour η at which hypercompression takes place, and directly targets η^* at which the posterior mass concentrates around the optimal $\tilde{\theta}$ — so SafeBayes can be expected to perform well as long as such η^* exists. Barron's theorem in combination with (20) implies (ignoring some subtleties explained in the appendix) that for $\eta^* < \bar{\eta}$, for all large enough n (depending on η^*), the η^*-generalized posterior indeed concentrates around $\tilde{\theta}$, so that SafeBayes will tend to find such an η^*.

In practice, it is computationally infeasible to try all values of η and we simply have to try out a grid of values, where, as shown by Grünwald (2012), it is sufficient to let grid points decrease exponentially fast. For convenience we give a detailed description of the algorithm below, copied from Grünwald (2012). In the present paper, we will invariably apply the algorithm with $z_i = (x_i, y_i)$ as before, and $\ell_\theta(z_i)$ set to the (conditional) log-loss as defined before.

Algorithm 1: The (R-)SafeBayesian algorithm

Input: data $z_1, \ldots, z_n$, model $\mathcal{M} = \{f(\cdot \mid \theta) \mid \theta \in \Theta\}$, prior Π on Θ, step-size κ_{STEP}, max. exponent $\kappa_{\max}$, loss function $\ell_\theta(z)$

Output: Learning rate $\hat{\eta}$

$\mathcal{S}_n := \{ 1, 2^{-\kappa_{\text{STEP}}}, 2^{-2\kappa_{\text{STEP}}}, 2^{-3\kappa_{\text{STEP}}}, \ldots, 2^{-\kappa_{\max}} \}$;

for *all* $\eta \in \mathcal{S}_n$ **do**

  $s_\eta := 0$;

  for $i = 1 \ldots n$ **do**

    Determine generalized posterior $\Pi(\cdot \mid z^{i-1}, \eta)$ of Bayes with learning rate η.

    Calculate "posterior-expected posterior-randomized loss" of predicting actual next outcome:

$$r := \ell_{\Pi|z^{i-1},\eta}(z_i) = \mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}\left[\ell_\theta(z_i)\right] \tag{24}$$

    $s_\eta := s_\eta + r$;

  end

end

Choose $\hat{\eta} := \arg\min_{\eta\in\mathcal{S}_n} \{ s_\eta \}$ (if min achieved for several $\eta \in \mathcal{S}_n$, pick largest);

Variation We may also consider the η which, instead of the η-R-log-loss ((23) and (24)), minimizes the *η-in-model-log-loss* (or just *η-I-log-loss*), defined as

$$\sum_{i=1}^{n} \left[-\log f(y_i \mid x_i, \mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}[\theta]) \right]. \tag{25}$$

We call the version of SafeBayes which minimizes the alternative objective function (25) *in-model SafeBayes*, abbreviated to I-log-SafeBayes, and from now on use R-log-SafeBayes for the original version based on the R-log-loss. Like (23), (25) cannot exploit hypercompression: it measures prediction error using the log score of a distribution within $\mathcal{M}^{(\eta)}$, whereas hypercompression only occurs if a distribution outside $\mathcal{M}^{(\eta)}$ is used; otherwise though, it is quite different from (23), and further illuminating motivation is provided in Section C.1 in the supplementary material. The theoretical results of Grünwald (2012) do not give any clue as to whether to prefer the I- or the R-versions, and a secondary goal of the experiments in this paper is thus to see whether one of them is always preferable over the other (we find that the answer is no). Explicit formulas instantiating both versions of SafeBayes to the linear model are given in Section C.1 in the supplement; Section C.3 recalls some theoretical results on SafeBayes' convergence behaviour.

5 Main experiment

In this section we provide our main experimental results, based on linear models $\mathcal{M}_p$ as defined in Section 2.2. Figures 5 and 6 depict, and Section 5.2 discusses the results of model selection and averaging experiments, which choose or average between the models $0, \ldots, p_{\max}$, where we consider first an incorrectly and then a correctly specified model, both with $p_{\max} = 50$; Figures G.1 and G.2 in the supplement do the same for $p_{\max} = 100$. Section 5.4 contains and interprets additional experiments on Bayesian ridge regression, with a fixed p; a multitude of additional experiments checking whether our results hold under model, prior and ground truth variations is provided in [GvO]. The final Section 6 summarizes the relevant findings of both the experiments below and these additional experiments. But first we need to explain the priors π and the sampling ('true') distributions P^* with which we experiment: as to the priors, in our model selection/averaging experiments, we use a fat-tailed prior on the models given by

$$\pi(p) \propto \frac{1}{(p+2)(\log(p+2))^2}.$$

This prior was chosen because it remains well-defined for an infinite collection of models, even though we only use finitely many in our experiments. As a sanity check we repeated some of our experiments with a uniform prior on $0, \ldots, p_{\max}$ instead; the results were indistinguishable. Each model $\mathcal{M}_p$ has parameters β, σ^2, on which we put the standard conjugate priors as described in Section 2.5. We set the mean of the prior on β to $\bar{\beta}_0 = \mathbf{0}$, and its covariance matrix to $\sigma^2\Sigma_0$ setting Σ_0 to the identity matrix $\Sigma_0 = \mathbf{I}_{p+1}$; the hyperparameters on the variance are set to $a_0 = 1$ and $b_0 = 40$; in Appendix C.2 we explain the reasons for this choice and alternatives we experimented with as well.

Concerning ground truth P^*, our experiments fall into two categories: correct-model and wrong-model experiments. In the *correct-model experiments*, $X_1, X_2, \ldots$ are sampled i.i.d., with, for each individual $X_i = (X_{i1}, \ldots, X_{ip_{\max}})$, $X_{i1} \ldots, X_{ip_{\max}}$ i.i.d. $\sim N(0,1)$. Given each X_i, Y_i is generated as

$$Y_i = .1 \cdot (X_{i1} + \ldots + X_{i4}) + \epsilon_i, \tag{26}$$

where the ϵ_i are i.i.d. $\sim N(0, \sigma^{*2})$ with variance $\sigma^{*2} = 1/40$. In contrast, in the *wrong-model experiments*, at each time point i, a fair coin is tossed independently of everything else. If the coin lands heads, then the point is 'easy', and $(X_i, Y_i) := (\mathbf{0}, 0)$. If the coin lands tails, then X_i is generated as for the correct model, and Y_i is generated as (26), but now the noise random variables have variance $\sigma_0^2 = 2\sigma^{*2} = 1/20$. Thus, $Z_i = (X_i, Y_i)$ is generated as in the true model case but with a larger variance; this larger variance has been chosen so that the marginal variance of each Y_i is the same value σ^{*2} in both experiments.

From the results in Section 2.3 we immediately see that, for both experiments, the optimal model is $\mathcal{M}_{\tilde{p}}$ for $\tilde{p} = 4$, and the optimal distribution in $\mathcal{M}$ and $\mathcal{M}_{\tilde{p}}$ is parameterized by $\tilde{\theta} = (\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ with $\tilde{p} = 4$, $\tilde{\beta} = (\tilde{\beta}_0, \ldots \tilde{\beta}_4) = (0, .1, .1, .1, .1), \tilde{\sigma}^2 = 1/40$ (in the correct model experiment, $\tilde{\sigma}^2 = \sigma^{*2}$; in the wrong model experiment, since $\tilde{\sigma}^2$ must be reliable, it must be equal to the square-risk obtained with $(\tilde{p}, \tilde{\beta})$, which is $(1/2) \cdot (1/20) = 1/40$). $f(x) := x\tilde{\beta}$ is then equal to the *true* regression function $\mathbf{E}_{P^*}[Y \mid X]$.

Variations We have already seen a variation of these two experiments depicted in Figures 1 and 2. In the correct-model version of that experiment, P^* is defined as follows: set $X_j = P_j(S)$, where P_j is the Legendre polynomial of degree j and S is uniformly distributed on $[-1, 1]$, and set $Y = 0 + \epsilon$, where $\epsilon \sim N(0, \sigma^{*2})$, with $\sigma^{*2} = 1/40$; $(X_1, Y_1), \ldots$ are then sampled as i.i.d. copies of (X, Y). Note that the true regression function is 0 here. In [GvO] we briefly consider this and several other variations of these ground truths.

5.1 The statistics we report

Figure 5 reports the results of the wrong-model experiment, and Figure 6 of the correct-model experiment, both with $p_{\max} = 50$. For both experiments we measure three aspects of the performance of Bayes and SafeBayes, each summarized in a separate graph. First, we show the behaviour of several prediction methods based on SafeBayes relative to square-risk; second, we check a form of model identification consistency; third, we measure whether the methods provide a good assessment of their own predictive capabilities in terms of square-loss, i.e. whether they are reliable and not 'overconfident'. Below we explain these three performance measures in detail. We also provide a fourth graph in each case indicating what $\hat{\eta}$'s are typically selected by the two versions of SafeBayes.

Square-risk For a given distribution W on (p, β, σ^2), the *regression function based on* W, a function mapping covariate X to $\mathbf{R}$, abbreviated to $\mathbf{E}_W[Y \mid X]$, is defined as

$$\mathbf{E}_W[Y \mid X] := \mathbf{E}_{(p,\beta,\sigma)\sim W} \mathbf{E}_{Y \sim P_{p,\beta,\sigma}|X}[Y] = \mathbf{E}_{(p,\beta,\sigma)\sim W} \left[\sum_{j=0}^{p} \beta_j X_j \right]. \tag{27}$$

If we take W to be the η-generalized posterior, then (27) is also simply called the η-posterior regression function. The *square-risk* relative to P^* based on predicting by W is then defined as an extension of (4) as

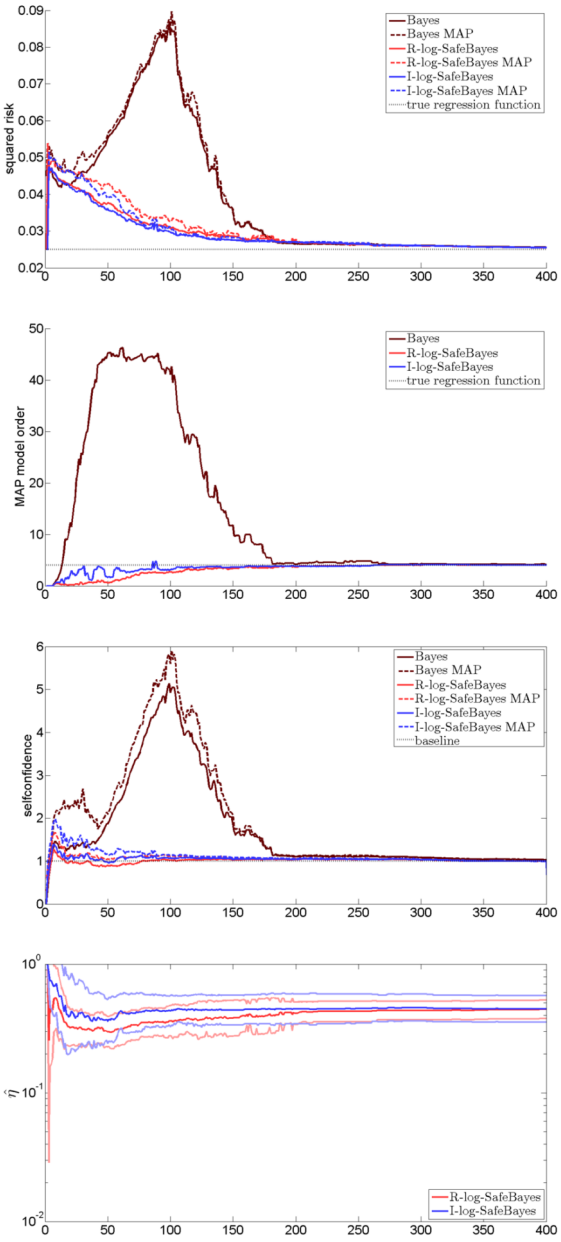

Figure 5: Four graphs showing respectively the square-risk, MAP model order, over-confidence (lack of reliability), and selected $\hat{\eta}$ at each sample size, each averaged over 30 runs, for the wrong-model experiment with $p_{\max} = 50$, for the methods indicated in Section 5.1. For the selected-$\hat{\eta}$ graph, the pale lines are one standard deviation apart from the average; all lines in this graph were computed over $\hat{\eta}$ indices (so that the lines depict the geometric mean over the values of $\hat{\eta}$ themselves).

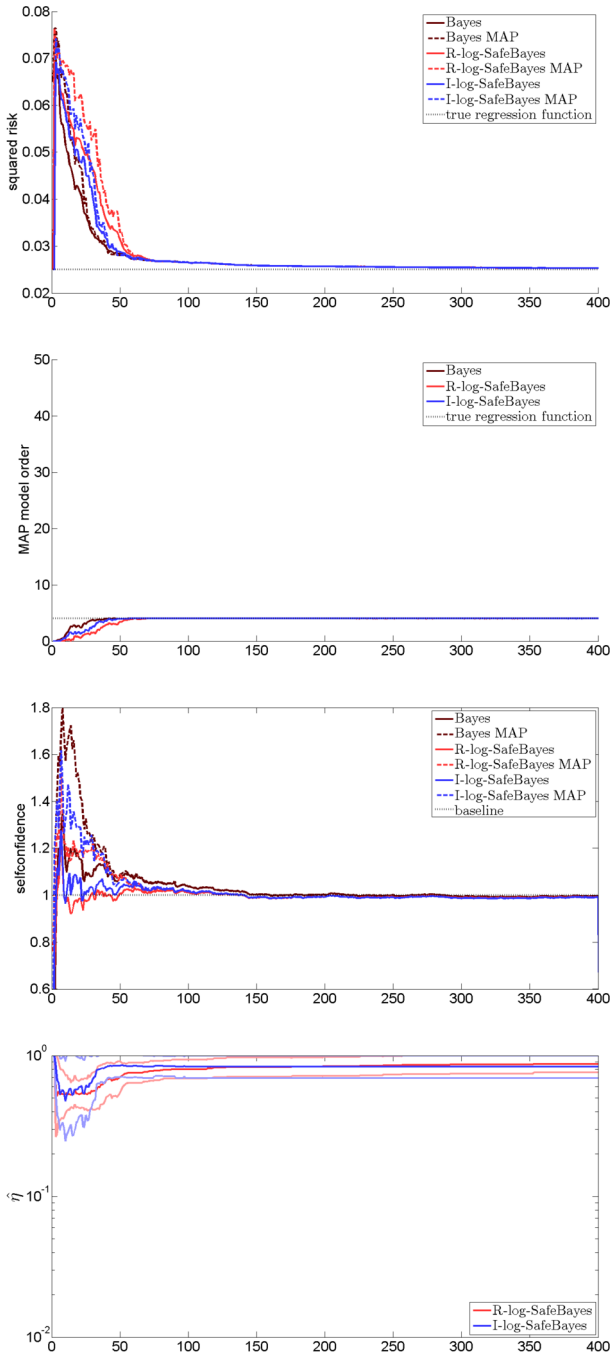

Figure 6: Same graphs as in Figure 5 for the correct-model experiment with $p_{\max} = 50$.

$$\text{RISK}^{\text{sq}}(W) := \mathbf{E}_{(X,Y)\sim P^*}(Y - \mathbf{E}_W[Y \mid X])^2. \tag{28}$$

In the experiments below we measure the square-risk relative to P^* at sample size $i-1$ achieved by, respectively, (1), the η-generalized posterior, and (2), the η-generalized posterior conditioned on the MAP (maximum a posteriori) model, i.e.

$$\mathbf{E}_{Z^{i-1}\sim P^*}[\text{RISK}^{\text{sq}}(W)], \text{ with } W = \Pi \mid Z^{i-1}, \eta \; ; \; W = \Pi \mid Z^{i-1}, \eta, \breve{p}_{\text{map}(Z^{i-1},\eta)} \tag{29}$$

respectively, where the MAP model $\breve{p}_{\text{map}(Z^{i-1},\eta)}$ is defined as the p achieving $\max_{p\in 0,\ldots,p_{\max}} \pi(p \mid Z^{i-1}, \eta)$, with $\pi(p \mid Z^{i-1}, \eta)$ defined as in (11). We do this for three values of η: (a) $\eta = 1$, corresponding to the standard Bayesian posterior, (b) $\eta := \hat{\eta}(Z^{i-1})$ set by the R-log SafeBayesian algorithm run on the past data Z^{i-1}, and (c) η set by the I-log SafeBayesian algorithm. In the figures of Section 5.2, 1(a) is abbreviated to *Bayes*, 1(b) is *R-log-SafeBayes*, 1(c) is *I-log-SafeBayes*, 2(a) is *Bayes MAP*, 2(b) is *R-log-SafeBayes MAP*, and 2(c) is *I-log-SafeBayes MAP*.

Concerning the two square-risks that we record: The first choice is the most natural, corresponding to the prediction (regression function) according to the 'standard' η-generalized posterior. The second corresponds to the situation where one first selects a single submodel $\breve{p}_{\text{map}}$ and then bases all predictions on that model; it has been included because such methods are often adopted in practice.

In Figure 5 and subsequent figures below, we depict these quantities by sequentially sampling data $Z_1, Z_2, \ldots, Z_{\max}$ i.i.d. from a P^* as defined above, where max is some large number. At each i, after the first $i-1$ points Z^{i-1} have been sampled, we compute the two square-risks given above. We repeat the whole procedure a number of times (called 'runs'); the graphs show the average risks over these runs.

MAP-model identification / Occam's razor When the goal of inference is model identification, 'consistency' of a method is often defined as its ability to identify the smallest model $\mathcal{M}_{\tilde{p}}$ containing the 'pseudo-truth' $(\tilde{\beta}, \tilde{\sigma}^2)$. To see whether standard Bayes and/or SafeBayes are consistent in this sense, we check whether the MAP model $\breve{p}_{\text{map}(Z^{i-1},\eta)}$ is equal to $\tilde{p}$.

Reliability vs. overconfidence Does Bayes learn how good it is in terms of squared error? To answer this question, we define, for a predictive distribution W as in (28) above, $U_i^{[W]}$ (a function of X_i, Y_i and (through W) of Z^{i-1}), as

$$U_i^{[W]} = (Y_i - \mathbf{E}_W[Y_i \mid X_i])^2.$$

This is the error we make if we predict Y_i using the regression function based on prediction method W. In the graphs in the next sections we plot the *self-confidence ratio* $\mathbf{E}_{X_i,Y_i\sim P^*}[U_i^{[W]}] / \mathbf{E}_{X_i\sim P^*}\mathbf{E}_{Y_i\sim W|X_i}[U_i^{[W]}]$ as a function of i for the three prediction methods / choices of W defined above. We may think of this as the ratio between the actual expected prediction error (measured in square-loss) one gets by using a predictor who based predictions on W and the marginal (averaged over X) subjectively expected prediction error by this predictor. We previously, in Section 2.3, showed

that the KL-optimal $(\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ is *reliable*: this means that, if we would take W the point mass on $(\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ and thus, irrespective of past data Z^{i-1}, would predict by $\mathbf{E}_{(\tilde{p},\tilde{\beta},\tilde{\sigma}^2)}[Y_i \mid X_i] = \sum_{j=0}^{\tilde{p}} \tilde{\beta}_j X_{ij}$, then the ratio would be 1. For the W learned from data considered above, a value larger than 1 indicates that W does not implement a 'reliable' method in the sense of Section 2.3, but, rather, is overconfident: it predicts its predictions to be better than they actually are, in terms of square-risk.

5.2 Main model selection/averaging experiment

We run the SafeBayesian algorithm of Section 4.2 with $z_i = (x_i, y_i)$ and $\ell_\theta(z_i) = -\log f_\theta(y_i \mid x_i)$ is the (conditional) log-loss as described in that section. As to the parameters of the algorithm, in all experiments we set the step-size $\kappa_{\text{STEP}} = 1/3$ and $\kappa_{\max} := 8$, i.e. we tried the following values of η: $1, 2^{-1/3}, 2^{-2/3}, \ldots, 2^{-8}$. The results of, respectively, the wrong-model and correct-model experiment as described above, with $p_{\max} = 50$, are given in Figures 5 and 6; results for $p_{\max} = 100$ can be found in Figures G.1 and G.2 in the supplement.

Conclusion 1: Bayes performs well in model-correct, and dismally in model-incorrect experiment The four figures show that standard Bayes behaves excellently in terms of all quality measures (square-risk, MAP model identification and reliability) when the model is correct, and dismally if the model is incorrect.

Conclusion 2: If (and only if) model incorrect, then the higher $p_{\max}$, the worse Bayes gets We see from Figures 6 and G.2 that standard Bayes behaves excellently in terms of all quality measures (square-risk, MAP model identification and reliability) when the model is correct, both if $p_{\max} = 50$ and if $p_{\max} = 100$, the behaviour at $p_{\max} = 100$ being essentially indistinguishable from the case with $p_{\max} = 50$. These and other (unreported) experiments strongly suggests that, when the data are sampled from a low-dimensional model, then, when the model is correct, standard Bayes is unaffected (does not get confused) by adding additional high-dimensional models to the model space. Indeed, the same is suggested by various existing Bayesian consistency theorems, such as those by Doob (1949); Ghosal et al. (2000); Zhang (2006a). At the same time, from Figures 5 and G.1 we infer that standard Bayes behaves very badly in all three quality measures in our (admittedly very 'evilly chosen') model-wrong experiment. Eventually, at very large sample sizes, Bayes recovers, but the main point here to notice is that the n at which a given level of recovery (measured in, say, square-loss) takes place is much higher for the case $p_{\max} = 100$ (Figure G.1) than for the case $p_{\max} = 50$ (Figure 5). This strongly suggests that, when the model is incorrect but the best approximation lies in a low-dimensional submodel, then standard Bayes gets confused by adding additional high-dimensional models to the model space — recovery takes place at a sample size that increases with $p_{\max}$. Indeed, the graphs suggest that in the case that $p_{\max} = \infty$ (with which we cannot experiment), Bayes will be inconsistent in the sense that the risk of the posterior predictive will never ever reach the risk attainable with the best submodel. Grünwald and Langford (2007) showed that this can indeed happen with a

simple, but much more unnatural classification model; the present result indicates (but does not mathematically prove) that it can happen with our standard model as well; see also the discussion in Section G.2.

Conclusion 3: R-log-SafeBayes and I-log-SafeBayes generally perform well Comparing the four graphs for R-log-SafeBayes and I-log-SafeBayes, we see that they behave quite well for *both* the model-correct and the model-wrong experiments, being slightly worse than, though still competitive to, standard Bayes when the model is correct and incomparably better when the model is wrong. Indeed, in the wrong-model experiments, about half of the data points are identical and therefore do not provide very much information, so one would expect that if a 'good' method achieves a given level of square-risk at sample size n in the correct-model experiment, it achieves the same level at about $2n$ in the incorrect-model experiment, and this is indeed what happens. Also, we see from comparing Figures G.1 and G.2 on the one hand to Figures 5 and 6 on the other that adding additional high-dimensional models to the model space hardly affects the results — like standard Bayes when the model is correct, SafeBayes does not get confused by living in a larger model space.

Secondary conclusions We see that both types of SafeBayes converge quickly to the right (pseudo-true) model order, which is pleasing since they were not specifically designed to achieve this. Whether this is an artefact of our setting or holds more generally would, of course, require further experimentation. We note that at small sample sizes, when both types of SafeBayes still tend to select an overly simple model, I-log-SafeBayes has significantly more variability in the model chosen-on-average; it is not clear though whether this is 'good' or 'bad'. We also note that the η's chosen by both versions are very similar for all but the smallest sample sizes, and are consistently smaller than 1. When instead of the full η-generalized posteriors, the η-generalized posterior conditioned on the MAP $\breve{p}_{\mathrm{map}}$ is used, the behaviour of all method consistently deteriorates a little, but never by much.

5.3 Experimental demonstration of hypercompression for standard Bayes

Figure 7 and Figure 8 show the predictive capabilities of standard Bayes in the wrong model example in terms of *cumulative* and *instantaneous log-loss* on a simulated sample. The graphs clearly demonstrate hypercompression: the Bayes predictive cumulatively performs *better* than the best single model / the best distribution in the model space, until at about $n \approx 100$ there is a phase transition. At individual points, we see that it sometimes performs a little worse, and sometimes (namely at the 'easy' $(0, 0)$ points) much better than the best distribution. We also see that, as anticipated above, for $\eta = 1$ randomized and in-model Bayesian prediction (used respectively by R- and I-log-SafeBayes to choose $\hat{\eta}$) do *not* show hypercompression and in fact perform terribly on the log-loss until the phase transition at $n = 100$, when they become almost as good as prediction with the Bayesian mixture.

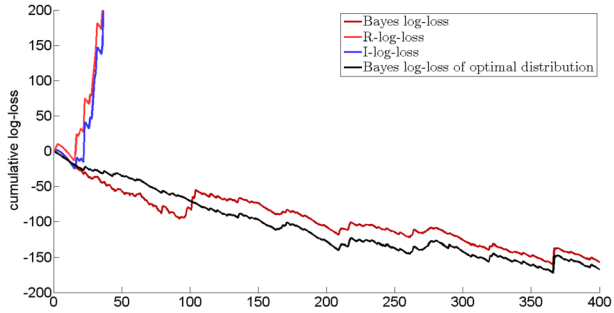

Figure 7: Cumulative standard, R-, and I-log-loss as defined in (23) and (25) respectively of standard Bayesian prediction ($\eta = 1$) on a single run for the model-averaging experiment of Figure 5. We clearly see that standard Bayes achieves *hypercompression*, being better than the best single model. And, as predicted by theory, randomized Bayes is never better than standard Bayes, whose curve has negative slope because the densities of outcomes are > 1 on average.

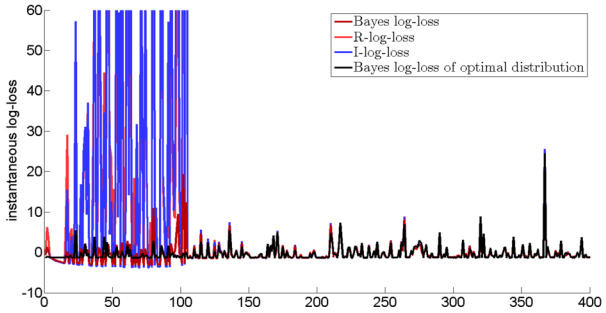

Figure 8: Instantaneous standard, R- and I-log-loss of standard Bayesian prediction for the run depicted in Figure 7.

The no-hypercompression inequality In fact, Figure 7 shows a phenomenon that is virtually impossible if the Bayesian's model and prior are 'correct' in the sense that data Z^n would behave like a typical sample from them: it easily follows from Markov's inequality (for details see Grünwald, 2007, Chapter 3) that, letting Π denote the Bayesian's joint distribution on $\Theta \times \mathcal{Z}^n$, for each $K \geq 0$,

$$\Pi\left\{(\theta, Z^n) \ : \ \sum_{i=1}^{n} \left(-\log \bar{f}(Y_i \mid X_i, Z^{i-1})\right) \leq \sum_{i=1}^{n} \left(-\log f_\theta(Y_i \mid X_i, Z^{i-1})\right) - K\right\} \leq e^{-K},$$

i.e. the probability that the Bayes predictive $\bar{f}$ cumulatively outperforms f_θ, with θ drawn from the prior, by K log-loss units is exponentially small in K. Figure 7 thus shows that at sample size $n \approx 90$, an a priori formulated event has happened of probability less than e^{-30}, clearly indicating that something about our model or prior is quite wrong.

Since the difference in cumulative log-loss between $\bar{f}$ and f_θ can be interpreted as the amount of bits saved when coding data from f_θ with a lossless code that would be optimal in expectation under $\bar{f}$ rather than f_θ, this result has been called the *no-hypercompression inequality* by Grünwald (2007). The figure shows that for our data, we have substantial hypercompression.

5.4 Second experiment: Bayesian ridge regression

We repeat the model-wrong and model-correct experiments of Figures 5 and 6, with just one difference: all posteriors are conditioned on $p := p_{\max} = 50$. Thus, we effectively consider just a fixed, high-dimensional model, whereas the best approximation $\tilde{\theta} = (50, \tilde{\beta}, \tilde{\sigma}^2)$ viewed as an element of $\mathcal{M}_p$ is 'sparse' in that it has only $\beta_1, \ldots, \beta_4$ not equal to 0. We note that the MAP model index graphs of Figures 5 and 6 are meaningless in this context (they would be equal to the constant 50) so they are left out of the new Figures 9 and 10. Results for *Cesàro-averaged* posteriors are shown instead; we refer to Section C.3 in the supplementary material for their definition and relevance.

Connection to Bayesian (b)ridge regression From (13) we see that the posterior mean parameter $\bar{\beta}_{i,\eta}$ is equal to the posterior MAP parameter and depends on η but not on σ^2, since σ^2 enters the prior in the same way as the likelihood. Therefore, the square-loss obtained when using the generalized posterior for prediction is always given by $(y_i - x_i \bar{\beta}_{i-1,\eta})^2$ irrespective of whether we use the posterior mean, or MAP, or the value of σ^2. Interestingly, if we fix some λ and perform standard (nongeneralized) Bayes with a modified prior, proportional to the original prior raised to the power $\lambda := \eta^{-1}$, then the prior, conditioned on σ^2, becomes normal $N(\bar{\beta}_0, \sigma^2 \Sigma_0')$ where $\Sigma_0' = \eta \Sigma_0$ and the standard posterior given z^i is then (by (13)) Gaussian with mean

$$\left((\Sigma_0')^{-1} + \mathbf{X}_n^\top \mathbf{X} \right)^{-1} \cdot \left((\Sigma_0')^{-1} \bar{\beta}_0 + \mathbf{X}_n^\top y^n \right) = \bar{\beta}_{i,\eta}. \tag{30}$$

Thus we see that in this special case, the (square-risk of the) η-generalized Bayes posterior mean coincides with the (square-risk of the) standard Bayes posterior mean with prior $N(\bar{\beta}_0, \sigma^2 \eta \Sigma_0)$ given σ^2, for arbitrary prior on σ^2. We first note that if Σ_0 is the identity matrix, then this implies that, for fixed η, the η-generalized Bayes posterior mean coincides with the ridge regression estimate with penalty parameter $\lambda := \eta^{-1}$. For general Σ_0, it implies that the posterior on β coincides with the posterior one gets with *Bayesian ridge regression*, as defined, by[4], e.g., Park and Casella (2008), conditioned on setting the λ-parameter in Bayesian ridge regression equal to η^{-1}. Now, the general Bayesian ridge regression method has λ as a free parameter, determined either by empirical Bayes or equipped with a prior. It is thus of interest to see what happens if, rather than using SafeBayes, we determine η (equivalently, λ) in this way. In addition to the graphs discussed earlier in Section 5.1, we thus also show the results for η set by empirical Bayes (in separate experiments not shown here we confirmed that putting a

[4]To be precise, they call this method Bayesian 'bridge' regression with $q = 2$; the choice of $q = 1$ in their formula gives their celebrated 'Bayesian Lasso'.

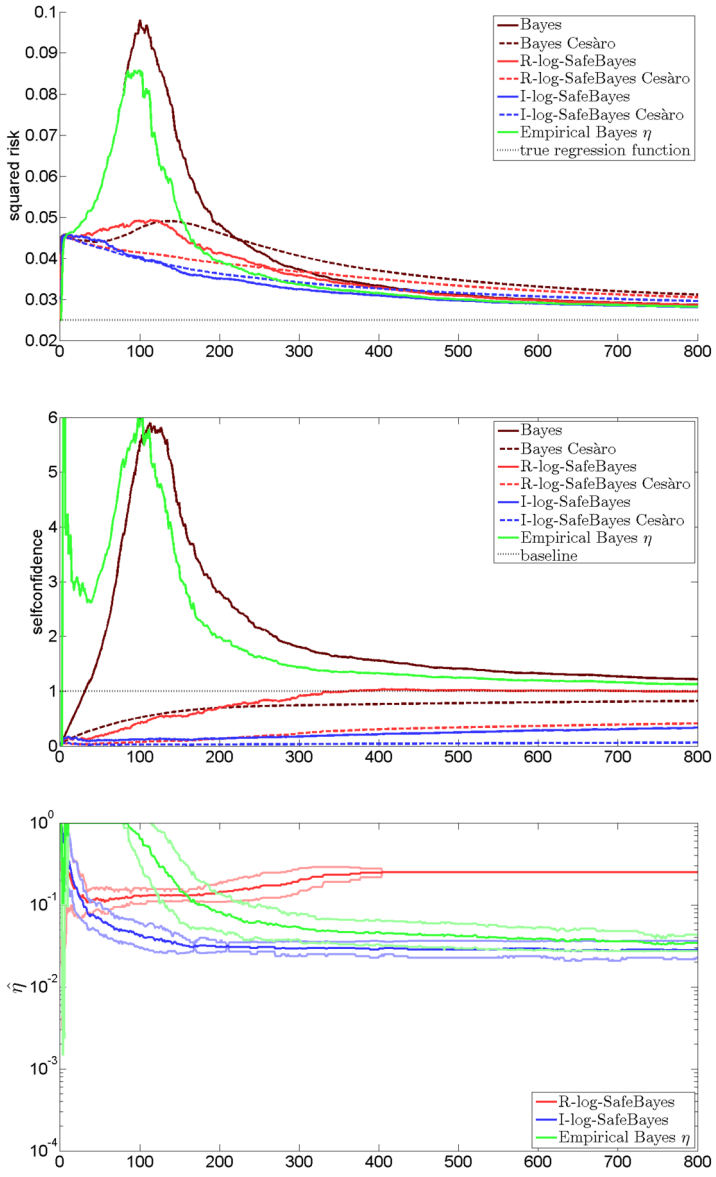

Figure 9: Bayesian ridge regression: Model-wrong experiment conditioned on $p := p_{\max} = 50$. The graphs (square-risk, self-confidence ratio and chosen η as function of sample size) are as in Figures 5 and 6, except for the third graph there (MAP model order), which has no meaning here. The meaning of the curves is given in Section 5.1 except for *empirical Bayes*, explained in Section 5.4.

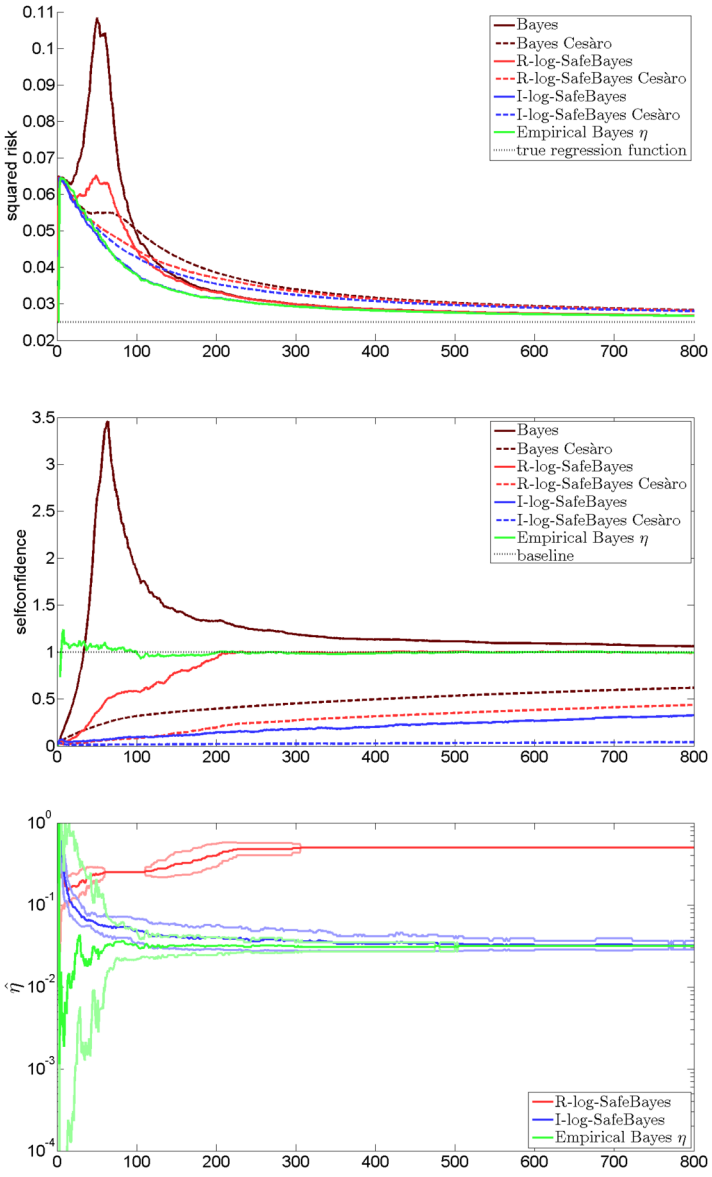

Figure 10: Bayesian ridge regression: Same graphs as in Figure 9, but for the model-correct experiment conditioned on $p := p_{\max} = 50$.

prior on η and using the full posterior on η gives essentially the same results). Whereas the empirical-Bayesian ridge regression is known to be a very competitive method in many practical settings (indeed in our model-correct experiment, Figure 10, it performs best in all respects), we see in Figure 9 (the green line) that, just like other versions of

Bayes, it breaks down under our type of misspecification. We hasten to add that the correspondence between the η-generalized posterior means and the standard posterior means with prior raised to power $1/\eta$ only holds for the $\bar{\beta}_{i,\eta}$ parameters. It does not hold for the $\bar{\sigma}^2_{i,\eta}$ parameters, and thus, for fixed η, the self-confidence ratio of both methods may be quite different.

Conclusions for model-wrong experiment For most curves, the overall picture of Figure 9 is comparable to the corresponding model averaging experiment, Figure 5: when the model is wrong, standard Bayes shows dismal performance in terms of risk and reliability up to a certain sample size and then very slowly recovers, whereas both versions of SafeBayes perform quite well even for small sample sizes. We do not show variations of the graph for $p = p_{\max} = 100$ (i.e. the analogue of Figure G.1), since it relates to Figure 9 in exactly the same way as Figure G.1 relates to Figure 5: with $p = 100$, bad square-risk and reliability behaviour of Bayes goes on for much longer (recovery takes place at much larger sample size) and remains equally good as for $p = 50$ with the two versions of SafeBayes.

We also see that, as we already indicated in the introduction, choosing the learning rate by empirical Bayes (thus implementing one version of Bayesian bridge regression) behaves terribly. This complies with our general theme that, to 'save Bayes' in general misspecification problems, the parameter η cannot be chosen in a standard Bayesian manner.

Conclusions for model-correct experiment The model-correct experiment for ridge regression (Figure 10) offers a surprise: we had expected Bayes to perform best, and were surprised to find that the SafeBayeses obtained smaller risk. Some followup experiments (not shown here), with different true regression functions and different priors, shed more light on the situation. Consider the setting in which the coefficients of the true function are drawn randomly according to the prior. In this setting standard Bayes performs at least as good in expectation as any other method including SafeBayes (the Bayesian posterior now represents exactly what an experimenter might ideally know). SafeBayes (still in this setting) usually chooses $\eta = 1/2$ or $1/4$, and the difference in risks compared to Bayes is small. On the other hand, if the true coefficients are drawn from a distribution with substantially smaller variance than a priori expected by the prior (a factor 1000 in the 'correct'-model experiment of Figure 10), then SafeBayes performs much better than Bayes. Here Bayes can no longer necessarily be expected to have the best performance (the model is correct, but the prior is "wrong"), and it is possible that a slightly reduced learning rate gives (significantly) better results. It seems that this situation, where the variance of the true function is much smaller than its prior expectation, is not exceptional: for example, Raftery et al. (1997) suggest choosing the variance of the prior in such a way that a large region of parameter values receives substantial prior mass. Following that suggestion in our experiments already gives a variance that is large enough compared to the true coefficients that SafeBayes performs better than Bayes even if the model is correct.

A joint observation for the model-wrong and model-correct experiments Finally we observe an interesting difference between the two SafeBayes versions here: I-log-SafeBayes seems better for risk, giving a smooth decreasing curve in both experiments. R-log-SafeBayes inherits a trace of standard Bayes' bad behaviour in both experiments, with a nonmonotonicity in the learning curve. On the other hand, in terms of reliability, R-log-SafeBayes is consistently better than I-log-SafeBayes (but note that the latter is underconfident, which is arguably preferable over being overconfident, as Bayes is). All in all, there is no clear winner between the two methods.

6 Executive summary: Main conclusions from experiments

Here we summarize the most important conclusions from our main experiments described above, as well as from the many variations of these main experiments that we performed to check the robustness of our conclusions in the technical report [GvO].

Standard Bayes In almost all our experiments, standard Bayesian inference fails in its KL-associated prediction tasks (squared error risk, reliability) when the model is wrong. Adopting a different prior (such as the g-prior) does not help, with two exceptions in model averaging: (a) if using a prior that follows the suggestions of Raftery et al. (1997), then Bayes works quite well with 50% 'easy' points, but starts to fail dramatically again (in contrast to SafeBayes) once the percentage of easy points is increased a bit further; (b) when it is run with a fixed variance that is significantly larger than the 'best' (pseudo-true) variance $\tilde{\sigma}^2$. Moreover, in the ridge regression experiments, as reported above, we find that standard Bayes can even perform much worse than SafeBayes when the model is correct — so all in all we tentatively conclude that SafeBayes is safer to use for linear regression.

SafeBayes The two SafeBayes methods discussed here behave reasonably well in all our experiments, and there is no clear winner among them. I-log-SafeBayes usually behaves excellently in terms of square-risk but is underconfident about its own performance (which is perhaps acceptable, overconfidence being a lot more dangerous). R-log-SafeBayes is usually good in terms of square-risk though not as good as I-log-SafeBayes, yet it is highly reliable. In [GvO] we additionally experiment with versions of SafeBayes for fixed-variance models, or, equivalently, for generalized posteriors of the generic form (7) with a squared loss function (the I-version (C.1) is discussed in the supplementary material, Section C.1). We find that (and explain why) for such cases, the R-method is not competitive with the I-method; the I-method works well for squared error prediction but, since it cannot learn σ^2 from the data, cannot be used to assess reliability (further explained in Section D.2).

Learning η in Bayes- or likelihood way/Bayesian ridge regression fails Despite its intuitive appeal, fitting η to the data by e.g. empirical Bayes fails in the model-wrong

ridge experiment with a prior on σ^2, where, as explained above, it amounts to Bayesian ridge regression (Figure 9). In [GvO] we also consider Bayesian ridge regression with a fixed variance, i.e. a degenerate prior that concentrates on a single σ^2. As we explain there, in this setting empirical Bayes learning of η reduces to empirical Bayes learning of the variance (since η^{-1} now plays just the same role as $\sigma^2/2$), and again it leads to very bad results in our model-wrong experiment. This latter empirical Bayes method for learning η is also related to one of the methods advocated by Bissiri et al. (2016), to which we return in Section F.

Robustness of experiments It does not matter whether the $X_{i1}, X_{i2}, \ldots$ are independent Gaussian, uniform or represent polynomial basis functions: all phenomena reported here persist for all choices. If the 'easy' points are not precisely $(0, 0)$, but have themselves a small variance in both dimensions, then all phenomena reported here persist, but on a smaller scale.

Centring We repeated several of our experiments with centred data, i.e. preprocessed data so that the empirical average of the Y_i is exactly 0 (Raftery et al., 1997; Hastie et al., 2001). In none of our experiments did this affect any results. We also looked at the case where the true regression function has an intercept far from 0, and data are *not* centred. This hardly affected the SafeBayes methods.

Other methods We also repeated the wrong-model experiment for several other methods of model selection: AIC (Akaike's Information Criterion), BIC (Bayesian Information Criterion), and various forms of cross-validation. Briefly, we found that all these have severe problems with our data as well. In these experiments, the mentioned methods were used to identify a model index p and η played no role. We also did an experiment where we used leave-one-out cross-validation to learn η itself. When we tried it with log-loss (as a likelihoodist or information-theorist might be tempted to do), it behaved terribly. However, with the squared error loss it worked fine in the sense that it achieved small squared error risk, which is not too surprising given its close similarity to the I-version of SafeBayes with a fixed variance (see Section C.1 in the Supplementary Material). Thus, we tentatively conclude (not surprisingly perhaps) that cross-validation based on the squared error loss works well if one is interested in the squared error risk; but it cannot be used for the other KL-associated prediction task, i.e. assessing reliability (Section D.2).

Real-world data Finally, we note that (De Heide, 2016b) is a preliminary report on some regression experiments with trigonometric basis functions, performed using the R-package `SafeBayes` (De Heide, 2016a). De Heide found that the phenomenon described in this paper also takes place for some real world data sets, namely daily maximum temperatures at the Seattle airport and air pollution data from the Openair Project at King's College, London (Carslaw and Ropkins, 2012). These data sets do appear heteroskedastic, and she finds that standard Bayes overfits and is substantially outperformed by SafeBayes, just as in Figure 2. Similarly, Quadrianto and Ghahramani

(2015) and Devaine et al. (2013) report good performance of SafeBayes-like methods in practice.

Supplementary Material

Supplementary material of "Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It" (DOI: 10.1214/17-BA1085SUPP; .pdf). In this paper, we described a problem for Bayesian inference under misspecification, and proposed the SafeBayes algorithm for solving it. The main appendix, Appendix B, places SafeBayes in proper context by giving a six point overview of what can go wrong in Bayesian inference from a frequentist point of view, and what can be done about it, both in the well- and in the misspecified case. Specifically we clarify the one other problem with Bayes under misspecification — interest in non-KL-associated tasks — and its relation to Gibbs posteriors. The remainder of the supplement is devoted to discussing these six points in great detail, explicitly stating several *Open Problems*, related work, and ideas for a general *Bayesian misspecification theory* as we go along. We also provide further details on SafeBayes (Appendix C), additional experiments (Appendix G) and refine and explain in more detail the notion of bad misspecification and hypercompression (Appendix D).

References

Audibert, J. Y. (2004). "PAC-Bayesian statistical learning theory." Ph.D. thesis, Université Paris VI. 1077

Barron, A. R. (1998). "Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics*, volume 6, 27–52. Oxford: Oxford University Press. 1071, 1080

Barron, A. R. and Cover, T. M. (1991). "Minimum Complexity Density Estimation." *IEEE Transactions on Information Theory*, 37(4): 1034–1054. MR1111806. 1071

Bissiri, P. G., Holmes, C., and Walker, S. G. (2016). "A General Framework for Updating Belief Distributions." *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 78(5): 1103–1130*. 1072, 1077, 1099

Carslaw, D. C. and Ropkins, K. (2012). "Openair – an R package for air quality data analysis." *Environmental Modelling and Software*, 27(18): 52–61. 1099

Catoni, O. (2007). *PAC-Bayesian Supervised Classification*. Lecture Notes-Monograph Series. IMS. MR2483528. 1077

Cuong, N. V., Lee, W. S., Ye, N., Chai, K. M. A., and Chieu, H. L. (2013). "Active Learning for Probabilistic Hypotheses Using the Maximum Gibbs Error Criterion." In *Advances in Neural Information Processing Systems 26*. 1084

Dawid, A. P. (1984). "Present Position and Potential Developments: Some Personal

Views, Statistical Theory, The Prequential Approach." *Journal of the Royal Statistical Society, Series A*, 147(2): 278–292. MR0763811. 1083, 1084

De Blasi, P. and Walker, S. G. (2013). "Bayesian asymptotics with misspecified models." *Statistica Sinica*, 23: 169–187. 1070, 1073

Devaine, M., Gaillard, P., Goude, Y., and Stoltz, G. (2013). "Forecasting electricity consumption by aggregating specialized experts; a review of the sequential aggregation of specialized experts, with an application to Slovakian and French country-wide one-day-ahead (half-)hourly predictions." *Machine Learning*, 90(2): 231–260. MR3015743. 1100

Diaconis, P. and Freedman, D. (1986). "On the Consistency of Bayes Estimates." *The Annals of Statistics*, 14(1): 1–26. MR0829555. 1073

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*, 23–27. Paris. 1091

Van Erven, T., Grünwald, P., Mehta, N., Reid, M., and Williamson, R. (2015). "Fast Rates in Statistical and Online Learning." *Journal of Machine Learning Research*. Special issue in Memory of Alexey Chervonenkis. MR3417799. 1082

Ghosal, S., Ghosh, J., and Van der Vaart, A. (2000). "Convergence rates of posterior distributions." *Annals of Statistics*, 28(2): 500–531. MR1790007. 1070, 1083, 1091

Grünwald, P. D. (1998). "The Minimum Description Length Principle and Reasoning under Uncertainty." Ph.D. thesis, University of Amsterdam, The Netherlands. Available as ILLC Dissertation Series 1998-03; see www.grunwald.nl. 1076

Grünwald, P. D. (1999). "Viewing all Models as "Probabilistic"." In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, 171–182. 1076

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. Cambridge, MA: MIT Press. 1081, 1093, 1094

Grünwald, P. D. (2011). "Safe Learning: Bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity." In *Proceedings of the Twenty-Fourth Conference on Learning Theory (COLT' 11)*. 1071, 1077

Grünwald, P. D. (2012). "The Safe Bayesian: Learning the learning rate via the mixability gap." In *Proceedings 23rd International Conference on Algorithmic Learning Theory (ALT '12)*. Springer. 1071, 1077, 1085, 1086

Grünwald, P. D. (2017). "Safe Probability." *Journal of Statistical Planning and Inference*. doi: http://dx.doi.org/10.1016/j.jspi.2017.09.014. To appear. 1074

Grünwald, P. D. and Langford, J. (2004). "Suboptimality of MDL and Bayes in classification under misspecification." In *Proceedings of the Seventeenth Conference on Learning Theory (COLT' 04)*. New York: Springer-Verlag. 1073

Grünwald, P. D. and Langford, J. (2007). "Suboptimal behavior of Bayes and

MDL in classification under misspecification." *Machine Learning*, 66(2–3): 119–149. doi: http://dx.doi.org/10.1007/s10994-007-0716-7. 1073, 1091

Grünwald, P. D. and Mehta, N. A. (2016). "Fast Rates with Unbounded Losses." *arXiv preprint arXiv:1605.00252*. 1074, 1082

Grünwald, P. D. and Van Ommen, T. (2014). "Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It." *arXiv preprint arXiv:1412.3730*. 1073, 1084

Grünwald, P. D. and Van Ommen, T. (2017). "Supplementary material of "Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It"." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/17-BA1085SUPP. 1070

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag. 1099

De Heide, R. (2016a). "*R*-Package `SafeBayes`." Freely available at CRAN Repository. 1072, 1099

De Heide, R. (2016b). "The Safe–Bayesian Lasso." Master's thesis, Leiden University. 1074, 1099

Hjorth, U. (1982). "Model Selection and Forward Validation." *Scandinavian Journal of Statistics*, 9: 95–105. 1084

Holmes, C. and Walker, S. (2017). "Assigning a value to a power likelihood in a general Bayesian model." *Biometrika*, 104(2): 497–503. MR3698270. 1072

Jiang, W. and Tanner, M. (2008). "Gibbs posterior for variable selection in high-dimensional classification and data mining." *Annals of Statistics*, 36(5): 2207–2231. 1077

Kleijn, B. and Van der Vaart, A. (2006). "Misspecification in infinite-dimensional Bayesian statistics." *Annals of Statistics*, 34(2). MR2283395. 1070

Martin, R., Mess, R., and Walker, S. G. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. MR3624879. 1077

McAllester, D. (2003). "PAC-Bayesian Stochastic Model Selection." *Machine Learning*, 51(1): 5–21. 1071, 1077

Miller, J. and Dunson, D. (2015). "Robust Bayesian Inference via Coarsening." Technical report, arXiv. Available at *arXiv:1506.06101*. 1072

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 1072, 1094

Quadrianto, N. and Ghahramani, Z. (2015). "A very simple safe-Bayesian random forest." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(6). 1099

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian model averaging for

linear regression models." *Journal of the American Statistical Association*, 92(437): 179–191. 1078, 1097, 1098, 1099

Ramamoorthi, R. V., Sriram, K., and Martin, R. (2015). "On posterior concentration in misspecified models." *Bayesian Analysis*, 10(4): 759–789. 1070

Rissanen, J. (1984). "Universal coding, information, prediction and estimation." *IEEE Transactions on Information Theory*, 30: 629–636. MR0755791. 1083, 1084

Royall, R. and Tsou, T.-S. (2003). "Interpreting statistical evidence by using imperfect models: Robust adjusted likelihood functions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 391–404. 1075

Seeger, M. (2002). "PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification." *Journal of Machine Learning Research*, 3: 233–269. MR1971338. 1077

Syring, N. and Martin, R. (2017). "Calibrating general posterior credible regions." *arXiv preprint arXiv:1509.00922*. 1072, 1073

Vovk, V. G. (1990). "Aggregating strategies." In *Proc. COLT' 90*, 371–383. 1071, 1077

Walker, S. and Hjort, N. L. (2002). "On Bayesian consistency." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4): 811–821. 1071, 1077

Zhang, T. (2006a). "From ϵ-entropy to KL entropy: Analysis of minimum information complexity density estimation." *Annals of Statistics*, 34(5): 2180–2210. 1070, 1071, 1077, 1091

Zhang, T. (2006b). "Information Theoretical Upper and Lower Bounds for Statistical Estimation." *IEEE Transactions on Information Theory*, 52(4): 1307–1321. 1077

Acknowledgments

A large part of this work was done while the authors were visiting UC San Diego. We would like to thank the UCSD CS department for hosting us. Wouter Koolen, Tim van Erven and Steven de Rooij played a crucial role in the development of the mixability gap which underlies the SafeBayesian algorithm. Many thanks also to Rianne de Heide for doing an additional experiment and to Larry Wasserman for encouragement and to the referees and associate editor for highly useful feedback. This research was supported by NWO VICI Project 639.073.04.