Safe-Bayesian Generalized Linear Regression

Rianne de Heide
Leiden University & CWI

Alisa Kirichenko
University of Oxford

Nishant A. Mehta
University of Victoria

Peter D. Grünwald
CWI & Leiden University

Abstract

We study generalized Bayesian inference under misspecification, i.e. when the model is 'wrong but useful'. Generalized Bayes equips the likelihood with a learning rate η. We show that for generalized linear models (GLMs), η-generalized Bayes concentrates around the best approximation of the truth within the model for specific $\eta \neq 1$, even under severely misspecified noise, as long as the tails of the true distribution are exponential. We derive MCMC samplers for generalized Bayesian lasso and logistic regression and give examples of both simulated and real-world data in which generalized Bayes substantially outperforms standard Bayes.

1 INTRODUCTION

Over the last ten years it has become abundantly clear that Bayesian inference can behave quite badly under misspecification, i.e., if the model $\mathcal{F}$ under consideration is 'wrong but useful' (Grünwald and Langford, 2007; Erven et al., 2007; Müller, 2013; Syring and Martin, 2017; Yao et al., 2018; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017). For example, Grünwald and Langford (2007) exhibit a simple nonparametric classification setting in which, even though the prior puts positive mass on the unique distribution in $\mathcal{F}$ that is closest in KL divergence to the data generating distribution P, the posterior never concentrates around this distribution. Grünwald and Van Ommen (2017) give a simple misspecified setting in which standard Bayesian ridge regression, model selection and model averaging severely overfit small-sample data.

Grünwald and Van Ommen (2017) also propose a remedy for this problem: equip the likelihood with an exponent or *learning rate* η (see (1) below). Such a *generalized Bayesian* (also known as *fractional* or *tempered* Bayesian) approach was considered earlier by e.g. Barron and Cover (1991); Walker and Hjort (2002); Zhang (2006b). In practice, η will usually (but not always — see Section 5.1 below) be chosen smaller than one, making the prior have a stronger regularizing influence. Grünwald and Van Ommen (2017) show that for Bayesian ridge regression and model selection/averaging, this results in excellent performance, being competitive with standard Bayes if the model is correct and very significantly outperforming standard Bayes if it is not. Extending Zhang's (2006a; 2006b) earlier work, Grünwald and Mehta (2019) (GM from now on) show that, under what was earlier called the $\bar{\eta}$*-central condition* (Definition 1 below), generalized Bayes with a specific finite learning rate $\bar{\eta}$ (usually $\neq 1$) will indeed concentrate in the neighborhood of the 'best' $f \in \mathcal{F}$ with high probability. Here, the 'best' f means the one closest in KL divergence to P.

Yet, three important parts of the story are missing in this existing work: (1) Can Grünwald-Van Ommen-type examples, showing failure of standard Bayes ($\eta = 1$) and empirical success of generalized Bayes with the right η, be given more generally, for different priors π (say of lasso-type ($\pi(f) \propto \exp(-\lambda \|f\|_1)$) rather than ridge-type $\pi(f) \propto \exp(-\lambda \|f\|_2^2)$), and for different models, say for *generalized* linear models (GLMs)? (2) Can we find examples of generalized Bayes outperforming standard Bayes with real-world data rather than with toy problems such as those considered by Grünwald and Van Ommen? (3) Does the central condition — which allows for good theoretical behavior of generalized Bayes — hold for GLMs, under reasonable further conditions?

We answer all three questions in the affirmative: in Section 2.1 below, we give (a) a toy example on which the Bayesian lasso and the Horseshoe estimator fail; later in the paper, in Section 5 we also (b) give a toy example on which standard Bayes logistic regression fails, and (c) two real-world data sets on which Bayesian lasso and Horseshoe regression fail; in all cases, (d) generalized Bayes with the right η shows

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

much better performance. In Section 3, we show (e) that for GLMs, even if the noise is severely misspecified, as long as the distribution of the predictor variable Y has exponentially small tails (which is automatically the case in classification, where the domain of Y is finite), the central condition holds for some $\eta > 0$. In combination with (e), GM's existing theoretical results suggest that generalized Bayes with this η should lead to good results — this is corroborated by our experimental results in Section 5. These findings are not obvious: one might for example think that the sparsity-inducing prior used by Bayesian lasso regression circumvents the need for the additional regularization induced by taking an $\eta < 1$, especially since in the original setting of Grünwald and Van Ommen, the standard Bayesian lasso ($\eta = 1$) succeeds. Yet, Example 1 below shows that under a modification of their example, it fails after all. In order to demonstrate the failure of standard Bayes and the success of generalized Bayes, we devise (in Section 4) MCMC algorithms (f) for generalized Bayes posterior sampling for Bayesian lasso and logistic regression. (a)-(f) are all novel contributions.

In Section 2 we first define our setting more precisely. Section 2.1) gives a first example of bad standard-Bayesian behavior and Section 2.2) recalls a theorem from GM indicating that under the $\bar{\eta}$-central condition, generalized Bayes for $\eta < \bar{\eta}$ should perform well. We present our new theoretical results in Section 3. We next (Section 4), present our algorithms for generalized Bayesian posterior sampling, and we continue (Section 5) to empirically demonstrate how generalized Bayes outperforms standard Bayes under misspecification. All proofs are in Appendix B.

2 THE SETTING

A *learning problem* can be characterized by a tuple $(P, \ell, \mathcal{F})$, where $\mathcal{F}$ is a set of predictors, also referred to as a *model*, P is a distribution on sample space $\mathcal{Z}$, and $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$ is a loss function. We denote by $\ell_f(z) \coloneqq \ell(f, z)$ the loss of predictor $f \in \mathcal{F}$ under outcome $z \in \mathcal{Z}$. If $Z \sim P$, we abbreviate $\ell_f(Z)$ to ℓ_f. In all our examples, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We obtain e.g. standard (random-design) regression with squared loss by taking $\mathcal{Y} = \mathbb{R}$ and $\mathcal{F}$ to be some subset of the class of all functions $f : \mathcal{X} \to \mathbb{R}$ and, for $z = (x, y)$, $\ell_f(x, y) = (y - f(x))^2$; logistic regression is obtained by taking $\mathcal{F}$ as before, $\mathcal{Y} = \{-1, 1\}$ and $\ell_f(x, y) = \log(1 + \exp(-yf(x))$. We get conditional density estimation by taking $\{p_f(Y \mid X) : f \in \mathcal{F}\}$ to be a family of conditional probability mass or density functions (defined relative to some measure μ), extended to n outcomes by the i.i.d. assumption, and taking conditional log-loss $\ell_f(x, y) \coloneqq -\log p_f(y \mid x)$.

We are given an i.i.d. sample $Z^n \coloneqq Z_1, Z_2, \ldots, Z_n \sim P$ where each Z_i takes values in $\mathcal{Z}$, and we consider, as our learning algorithm, the *generalized Bayesian posterior*, also known as the *Gibbs posterior*, Π_n on $\mathcal{F}$, defined by its density

$$\pi_n(f) \coloneqq \frac{\exp\left(-\eta \sum_{i=1}^n \ell_f(z_i)\right) \cdot \pi_0(f)}{\int_{\mathcal{F}} \exp\left(-\eta \sum_{i=1}^n \ell_f(z_i)\right) \cdot \pi_0(f) \mathrm{d}\rho(f)}, \quad (1)$$

where $\eta > 0$ is the *learning rate*, and π_0 is the density of some prior distribution Π_0 on $\mathcal{F}$ relative to an underlying measure ρ. Note that, in the conditional log-loss setting, we get that

$$\pi_n(f) \propto \prod_{i=1}^n (p_f(y_i \mid x_i))^\eta \pi_0(f), \quad (2)$$

which, if $\eta = 1$, reduces to standard Bayesian inference. While GM's result (quoted as Theorem 1 below) works for arbitrary loss functions, Theorem 2 and our empirical simulations (this paper's new results) revolve around (generalized) linear models. For these models, (1) can be equivalently interpreted either in terms of the original loss functions ℓ_f or in terms of the conditional likelihood p_f. For example, consider regression with $\ell_f(x, y) = (y - f(x))^2$ and fixed η. Then (1) induces the same posterior distribution $\pi_n(f)$ over $\mathcal{F}$ as does (2) with the conditional distributions $p_f(y|x) \propto \exp(-(y - f(x))^2$, which is again the same as (1) with ℓ_f replaced by the conditional log-loss $\ell'_f(x, y) \coloneqq -\log p_f(y|x)$, giving a likelihood corresponding to Gaussian errors with a particular fixed variance; an analogous statement holds for logistic regression. Thus, all our examples can be interpreted in terms of (2) for a model that is misspecified, i.e., the density of $P(Y|X)$ is not equal to p_f for any $f \in \mathcal{F}$. As is customary (see e.g. Bartlett et al. (2005)), we assume throughout that there exists an optimal $f^* \in \mathcal{F}$ that achieves the smallest *risk* (expected loss) $\mathbf{E}[\ell_{f^*}(Z)] = \inf_{f \in \mathcal{F}} \mathbf{E}[\ell_f(Z)]$. If $\mathcal{F}$ is a GLM, the risk minimizer again has additional interpretations: first, f^* minimizes, among all $f \in \mathcal{F}$, the conditional KL divergence $\mathbf{E}_{(X,Y) \sim P}[\log(p(Y|X)/p_f(Y|X))]$ to the true distribution P. Second, if there is an $f \in \mathcal{F}$ with $\mathbf{E}_{X,Y \sim P}[Y \mid X] = f(X)$ (i.e. $\mathcal{F}$ contains the *true regression function*, or equivalently, *true conditional mean*), then the risk minimizer satisfies $f^* = f$.

2.1 Bad Behavior of Standard Bayes

Example 1. We consider a Bayesian lasso regression setting (Park and Casella, 2008) with random design, with a Fourier basis. We sample data $Z_i = (X_i, Y_i)$ i.i.d. $\sim P$, where P is defined as follows: we first sample *preliminary* (X'_i, Y'_i) with $X'_i \overset{i.i.d.}{\sim} \text{Uniform}([-1, 1])$; the dependent variable Y'_i is set to $Y'_i = 0 + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for some fixed value of σ, independently of

X'_i. In other words: the true distribution for (X'_i, Y'_i) is 'zero with Gaussian noise'. Now we toss a fair coin for each i. If the coin lands heads, we set the actual $(X_i, Y_i) \coloneqq (X'_i, Y'_i)$, i.e. we keep the (X'_i, Y'_i) as they are, and if the coin lands tails, we put the pair to zero: $(X_i, Y_i) \coloneqq (0, 0)$.

We model the relationship between X and Y with a p^{th} order Fourier basis. Thus, $\mathcal{F} = \{f_\beta : \beta \in \mathbb{R}^{2p+1}\}$, with $f_\beta(x)$ given by

$$\left\langle \beta, \frac{1}{\pi} \cdot \left(2^{-1/2}, \cos(x), \sin(x), \cos(2x), \ldots, \sin(px)\right)\right\rangle,$$

and the η-posterior is defined by (1) with $\ell_{f_\beta}(x,y) = (y - f_\beta(x))^2$; the prior is the Bayesian lasso prior whose definition we recall in Section 4.1. Since our 'true' regression function $\mathbf{E}[Y_i \mid X_i]$ is 0, in an actual sample around 50% of points will be noiseless, *easy* points, lying on the true regression function. Since the actual sample of (X_i, Y_i) has less noise then the original sample (X'_i, Y'_i), we would expect Bayesian lasso regression to learn the correct regression function, but as we see in the blue line in Figure 1, it overfits and learns the noise instead (later on (Figure 3 in Section 5.1) we shall see that, not surprisingly, this results in terrible predictive behavior). By removing the noise in half the data points, we misspecified the model: we made the noise heteroscedastic, whereas the model assumes homoscedastic noise. Thus, in this experiment the *model is wrong*. Still, the distribution in $\mathcal{F}$ closest to the true P, both in KL divergence and in terms of minimizing the squared error risk, is given by the conditional distribution corresponding to $Y_i = 0 + \epsilon_i$, where ϵ_i is i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. While this element of $\mathcal{F}$ is in fact *favored* by the prior (the lasso prior prefers β with small $\|\beta\|_1$), nevertheless, for small samples, the standard Bayesian posterior puts most if its mass at f with many nonzero coefficients. In contrast, the generalized posterior (1) with $\eta = 0.25$ gives excellent results here. To learn this η from the data, we can use the Safe-Bayesian algorithm of Grünwald (2012). The result is depicted as the red line in Figure 1. Implementation details are in Section 4.1 and Appendix E; the details of the figure are in Appendix F.

The example is similar to that of Grünwald and Van Ommen (2017), who use multidimensional X and a ridge (normal) prior on $\|\beta\|$; in their example, standard Bayes succeeds when equipped with a lasso prior; by using a trigonometric basis we can make it 'fail' after all. Grünwald and Van Ommen (2017) relate the potential for the overfitting-type of behavior of standard Bayes, as well as the potential for full inconsistency (i.e. even holding as $n \to \infty$) as noted by Grünwald and Langford (2007) to properties of the Bayesian predictive distribution $\bar{p}(Y_{n+1} \mid X_{n+1}, Z^n) \coloneqq \int_{\mathcal{F}} p_f(Y_{n+1} \mid X_{n+1}) \pi_n(f \mid Z^n) \mathrm{d}\rho(f)$. Being a mixture of $f \in \mathcal{F}$, $\bar{p}(Y_{n+1} \mid X_{n+1})$, is a member of the convex hull of densities $\mathcal{F}$ but not necessarily of $\mathcal{F}$ itself. As explained by Grünwald and Van Ommen, severe overfitting may take place if $\bar{p}(Y_{n+1} \mid X_{n+1}, Z^n)$ is 'far' from any of the distributions in $\mathcal{F}$. It turns out that this is exactly what happens in the lasso example above, as we see from Figure 2 (details in Appendix F). This figure plots the data points as $(X_i, 0)$ to indicate their location; we see that the predictive variance of standard Bayes fluctuates, being small around the data points and large elsewhere. However, it is obvious that for every density p_f in our model $\mathcal{F}$, the variance is fixed independently of X, and thus $\bar{p}(Y_{n+1} \mid X_{n+1}, Z^n)$ is indeed very far from any particular p_f with $f \in \mathcal{F}$. In contrast, for the generalized Bayesian lasso with $\eta = 0.25$, the corresponding predictive variance is almost constant; thus, at the level $\eta = 0.25$ the predictive distribution is almost 'in-model' (in machine learning terminology, we may say that $\bar{p}$ is 'proper' (Shalev-Shwartz and Ben-David, 2014), and the overfitting behavior then does not occur anymore.

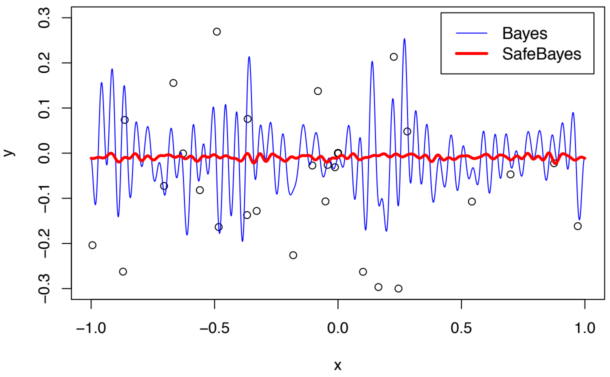

Figure 1: *Predictions of standard Bayes (blue) and SafeBayes (red), $n = 50$, $p = 101$.*

2.2 When Generalized Bayes Concentrates

Having just seen bad behavior for $\eta = 1$, we now recall some results from GM. Under some conditions, GM show that generalized Bayes, for appropriately chosen η, does concentrate at fast rates even under misspecification. We first recall (a very special case of) the asymptotic behavior under misspecification theorem of GM. GM bound (a) the *misspecification metric* $d_{\bar{\eta}}$ in terms of (b) the *information complexity*. The bound (c) holds under a simple condition on the learning problem that was termed the *central condition* by Van Erven et al. (2015). Before presenting the theorem we explain (a)–(c). As to (a), we define the *misspeci-*

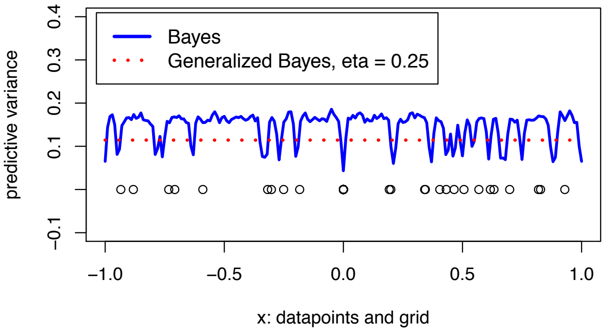

Figure 2: *Variance of Predictive Distribution* $\bar{p}(Y_{n+1} \mid X_{n+1}, Z^n)$ *for a single run with* $n = 50$.

fication metric $d_{\bar{\eta}}$ in terms of its square by

$$d^2_{\bar{\eta}}(f, f') \coloneqq \frac{2}{\bar{\eta}}\left(1 - \int \sqrt{p_{f,\bar{\eta}}(z) p_{f',\bar{\eta}}(z)} \mathrm{d}\mu(z)\right)$$

which is the ($2/\bar{\eta}$-scaled) squared Hellinger distance between $p_{f,\bar{\eta}}$ and $p_{f',\bar{\eta}}$. Here, a density $p_{f,\bar{\eta}}$ is defined as

$$p_{f,\bar{\eta}}(z) \coloneqq p(z)\frac{\exp(-\bar{\eta} L_f(z))}{\mathbf{E}[\exp(-\bar{\eta} L_f(Z))]},$$

where $L_f = \ell_f - \ell_{f^*}$ is the *excess loss* of f. GM show that $d_{\bar{\eta}}$ defines a metric for all $\bar{\eta} > 0$. If $\bar{\eta} = 1$, ℓ is log-loss, and the model is well-specified, then it is straightforward to verify that $p_{f,\bar{\eta}} = p_f$, and so $(1/2) \cdot d_{\bar{\eta}}$ becomes the standard squared Hellinger distance.

As to (b), we denote by $\mathrm{IC}_{n,\eta}(\Pi_0)$ the information complexity, defined as:

$$\mathrm{IC}_{n,\eta}(\Pi_0) \coloneqq \mathbf{E}_{\underline{f} \sim \Pi_n}\left[\frac{1}{n}\sum_{i=1}^{n} L_{\underline{f}}(Z_i)\right] + \frac{\mathrm{KL}(\Pi_n \,\|\, \Pi_0)}{\eta \cdot n} =$$
$$-\frac{1}{\eta n}\log \int_{\mathcal{F}} \pi_0(f) e^{-\eta \sum_{i=1}^{n} \ell_f(Z_i)} \mathrm{d}\rho(f) - \sum_{i=1}^{n} \ell_{f^*}(Z_i), \quad (3)$$

where $\underline{f}$ denotes the predictor sampled from the posterior Π_n and KL denotes KL divergence; we suppress dependency of IC on f^* in the notation. The fact that both lines above are equal (noticed by, among others, Zhang (2006b); GM give an explicit proof) allows us to write the information complexity in terms of a generalized Bayesian predictive density which is also known as *extended stochastic complexity* (Yamanishi, 1998). It also plays a central role in the field of prediction with expert advice as the *mix-loss* (Van Erven et al., 2015; Cesa-Bianchi and Lugosi, 2006) and coincides with the minus log of the standard Bayesian predictive density if $\eta = 1$ and ℓ is log-loss. It can be thought of as a complexity measure analogous to VC dimension and Rademacher complexity.

As to (c), GM's result holds under the *central condition* ((Li, 1999); name due to Van Erven et al. (2015)) which expresses that, for some fixed $\bar{\eta} > 0$, for all fixed f, the probability that the loss of f exceeds that of the optimal f^* by $a/\bar{\eta}$ is exponentially small in a:

Definition 1 (Central Condition, Def. 7 of GM)**.** *Let* $\bar{\eta} > 0$*. We say that* $(P, \ell, \mathcal{F})$ *satisfies the* $\bar{\eta}$-strong *central condition if, for all* $f \in \mathcal{F}$: $\mathbf{E}\left[e^{-\bar{\eta} L_f}\right] \le 1$.

As straightforward rewriting shows, this condition holds *automatically*, for any $\bar{\eta} \le 1$ in the density estimation setting, if the model is correct; Van Erven et al. (2015) provide some other cases in which it holds, and show that many other conditions on ℓ and P that allow fast rate convergence that have been considered before in the statistical and on-line learning literature, such as *exp-concavity* (Cesa-Bianchi and Lugosi, 2006), the *Tsybakov* and *Bernstein* conditions (Bartlett et al., 2005; Tsybakov, 2004) and several others, can be viewed as special cases of the central condition; yet they don't discuss GLMs. Here is GM's result:

Theorem 1 (Theorem 10 from GM)**.** *Suppose that the* $\bar{\eta}$*-strong central condition holds. Then for any* $0 < \eta < \bar{\eta}$*, the metric* $d_{\bar{\eta}}$ *satisfies*

$$\mathbf{E}_{Z^n \sim P}\, \mathbf{E}_{\underline{f} \sim \Pi_n}\left[d^2_{\bar{\eta}}(f^*, \underline{f})\right] \le C_\eta \cdot \mathbf{E}_{Z^n \sim P}\left[\mathrm{IC}_{n,\eta}(\Pi_0)\right]$$

with $C_\eta = \eta/(\bar{\eta} - \eta)$*. In particular,* $C_\eta < \infty$ *for* $0 < \eta < \bar{\eta}$*, and* $C_\eta = 1$ *for* $\eta = \bar{\eta}/2$.

Thus, we expect the posterior to concentrate at a rate dictated by $\mathbf{E}[\mathrm{IC}_{n,\eta}]$ in neighborhoods of the best (risk-minimizing, KL optimal, or even true regression function) f^*. The misspecification metric $d^2_{\bar{\eta}}$ on the left hand side is a weak metric, however, in Appendix C we show that we can replace it by stronger notions such as KL-divergence, squared error or logistic loss. Theorem 1 generalizes previous results (e.g. Zhang (2006a,b)) to the misspecified setting. In the well-specified case, Zhang, as well as several other authors (Walker and Hjort, 2002; Martin et al., 2017), state a result that holds for any $\eta < 1$ but not $\eta = 1$. This suggests that there is an advantage to taking η slightly smaller than one even when the model is well-specified (for more details see Zhang (2006a)).

To make the theorem work for GLMs under misspecification, we must verify (a) that the central condition still holds (which is in general not guaranteed) and that (b) the information complexity is sufficiently small. As to (a), in the following section we show that the central condition holds (with $\bar{\eta}$ usually $\neq 1$) for 1-dimensional exponential families and high-dimensional generalized linear models (GLMs) if the noise is misspecified, as long as P has exponentially small tails;

in particular, we relate $\bar{\eta}$ to the variance of P. As to (b), if the model is correct (the conditional distribution $P(Y \mid X)$ has density f equal to p_f with $f \in \mathcal{F}$), where $\mathcal{F}$ represents a d-dimensional GLM, then it is known (see e.g. Zhang (2006b)) that, for any prior Π_0 with continuous, strictly positive density on $\mathcal{F}$, the information complexity satisfies

$$\mathbf{E}_{Z^n \sim P}\left[\mathrm{IC}_{n,\eta}(\Pi_0)\right] = O\left(\frac{d}{n} \cdot \log n\right), \tag{4}$$

which leads to bounds within a log-factor of the minimax optimal rate (among all possible estimators, Bayesian or not), which is $O(d/n)$. While such results were only known for the well-specified case, in Proposition 1 below we show that, for GLMs, they continue to hold for the misspecified case.

3 GENERALIZED GLM BAYES

Below we first show that the central condition holds for natural univariate exponential families; we then extend this result to the GLM case, and establish bounds in information complexity of GLMs. Let the class $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ be a univariate natural exponential family of distributions on $\mathcal{Z} = \mathcal{Y}$, represented by their densities, indexed by natural parameter $\theta \in \Theta \subset \mathbb{R}$ (Barndorff-Nielsen, 1978). The elements of this restricted family have probability density functions

$$p_\theta(y) \coloneqq \exp(\theta y - F(\theta) + r(y)), \tag{5}$$

for log-normalizer F and carrier measure r. We denote the corresponding distribution as P_θ. In the first part of the theorem below we assume that Θ is restricted to an arbitrary closed interval $[\underline{\theta}, \bar{\theta}]$ with $\underline{\theta} < \bar{\theta}$ that resides in the interior of the natural parameter space $\bar{\Theta} = \{\theta : F(\theta) < \infty\}$. Such Θ allow for a simplified analysis because within Θ the log-normalizer F as well as all its derivatives are uniformly bounded from above and below; see (7) in Appendix B. As is well-known (see e.g. Barndorff-Nielsen (1978)), exponential families can equivalently be parameterized in terms of the mean-value parameterization: there exists a 1-to-1 strictly increasing function $\mu : \bar{\Theta} \to \mathbb{R}$ such that $\mathbf{E}_{Y \sim P_\theta}[Y] = \mu(\theta)$. As is also well-known, the density $p_{f^*} \equiv p_{\theta^*}$ within $\mathcal{F}$ minimizing KL divergence to the true distribution P satisfies $\mu(\theta^*) = \mathbf{E}_{Y \sim P}[Y]$, whenever the latter quantity is contained in $\mu(\Theta)$ (Grünwald, 2007). In words, the best approximation to P in $\mathcal{F}$ in terms of KL divergence has the same mean of Y as P.

Theorem 2. *Consider a learning problem $(P, \ell, \mathcal{F})$ with $\ell_\theta(y) = -\log p_\theta(y)$ the log loss and $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ a univariate exponential family as above.*
(1). Suppose that $\Theta = [\underline{\theta}, \bar{\theta}]$ is compact as above and that $\theta^* = \arg\min_{\theta \in \bar{\Theta}} D(P \| P_\theta)$ lies in Θ. Let $\sigma^2 > 0$ be the true variance $\mathbf{E}_{Y \sim P}(Y - \mathbf{E}[Y])^2$ and let $(\sigma^*)^2$ be the variance $\mathbf{E}_{Y \sim P_{\theta^*}}(Y - \mathbf{E}[Y])^2$ according to θ^*. Then

(i) for all $\bar{\eta} > (\sigma^*)^2/\sigma^2$, the $\bar{\eta}$-central condition does not *hold.*

(ii) Suppose there exists $\eta^\circ > 0$ such that $\bar{C} \coloneqq \mathbf{E}_P[\exp(\eta^\circ |Y|)] < \infty$. Then there exists $\bar{\eta} > 0$, depending only on η°, $\bar{C}$, $\underline{\theta}$ and $\bar{\theta}$ such that the $\bar{\eta}$-central condition holds. Moreover,

(iii), for all $\delta > 0$, there is an $\epsilon > 0$ such that, for all $\bar{\eta} \leq (\sigma^*)^2/\sigma^2 - \delta$, the $\bar{\eta}$-central condition holds relative to the restricted model $\mathcal{F}_\epsilon = \{p_\theta : \theta \in [\theta^* - \epsilon, \theta^* + \epsilon]\}$.

(2). Suppose that P is Gaussian with variance $\sigma^2 > 0$ and that $\mathcal{F}$ indexes a full Gaussian location family. Then the $\bar{\eta}$-central condition holds iff $\bar{\eta} \leq (\sigma^*)^2/\sigma^2$.

We provide (iii) just to give insight — 'locally', i.e. in restricted models that are small neighborhoods around the best-approximating θ^*, the smallest $\bar{\eta}$ for which the central condition holds is determined by a ratio of variances. The final part shows that for the Gaussian family, the same holds not just locally but globally (note that we do not make the compactness assumption on Θ there); we warn the reader though that the standard posterior ($\eta = 1$) based on a model with fixed variance σ^* is quite different from the generalized posterior with $\eta = (\sigma^*)^2/\sigma^2$ and a model with variance σ^2 (Grünwald and Van Ommen, 2017). Finally, while in practical cases we often find $\bar{\eta} < 1$ (suggesting that Bayes may only succeed if we learn 'slower' than with the standard $\eta = 1$, i.e. the prior becomes more important), the result shows that we can also very well have $\bar{\eta} > 1$; we give a practical example at the end of Section 5. Theorem 2 is new and supplements Van Erven et al.'s (2015) various examples of $\mathcal{F}$ which satisfy the central condition. In the theorem we require that both tails of Y have exponentially small probability.

Central Condition: GLMs Let $\mathcal{F}$ be the generalized linear model (McCullagh and Nelder, 1989) (GLM) indexed by parameter $\beta \in \mathcal{B} \subset \mathbb{R}^d$ with link function $g : \mathbb{R} \to \mathbb{R}$. By definition this means that there exists a set $\mathcal{X} \subset \mathbb{R}^d$ and a univariate exponential family $\mathcal{Q} = \{p_\theta : \theta \in \bar{\Theta}\}$ on $\mathcal{Y}$ of the form (5) such that the conditional distribution of Y given $X = x$ is, for all possible values of $x \in \mathcal{X}$, a member of the family $\mathcal{Q}$, with mean-value parameter $g^{-1}(\langle \beta, x \rangle)$. Then the class $\mathcal{F}$ can be written as $\mathcal{F} = \{p_\beta : \beta \in \mathcal{B}\}$, a set of conditional probability density functions such that

$$p_\beta(y \mid x) \coloneqq \exp\big(\theta_x(\beta) y - F(\theta_x(\beta)) + r(y)\big), \tag{6}$$

where $\theta_x(\beta) \coloneqq \mu^{-1}(g^{-1}(\langle \beta, x\rangle))$, and μ^{-1}, the inverse of μ defined above, sends mean parameters to natural parameters. We then have $\mathbf{E}_{P_\beta}[Y \mid X] = g^{-1}(\langle \beta, X\rangle)$, as required.

Proposition 1. *Under the following three assumptions, the learning problem $(P, \ell, \mathcal{F})$ with $\mathcal{F}$ as above satisfies the $\bar{\eta}$-central condition for some $\bar{\eta} > 0$ depending only on the parameters of the problem:*

1. *(Conditions on g): the inverse link function g^{-1} has bounded derivative on the domain $\mathcal{B} \times \mathcal{X}$, and the image of the inverse link on the same domain is a bounded interval in the interior of the mean-value parameter space $\{\mu \in \mathbb{R} : \mu = \mathbf{E}_{Y \sim q}[Y] : q \in \mathcal{Q}\}$ (for all standard link functions, this can be enforced by restricting $\mathcal{B}$ and $\mathcal{X}$ to an (arbitrarily large but still) compact domain).*
2. *(Condition on 'true' P): for some $\eta > 0$ we have $\sup_{x \in \mathcal{X}} \mathbf{E}_{Y \sim P}[\exp(\eta |Y|) \mid X = x] < \infty$.*
3. *(Well-specification of conditional mean): there exists $\beta^\circ \in \mathcal{B}$ such that $\mathbf{E}[Y \mid X] = g^{-1}(\langle \beta^\circ, X\rangle)$.*

A simple argument (differentiation with respect to β) shows that under the third condition, it must be the case that $\beta^\circ = \beta^*$, where $\beta^* \in \mathcal{B}$ is the index corresponding to the density $p_{f^*} \equiv p_{\beta^*}$ within $\mathcal{F}$ that minimizes KL divergence to the true distribution P. Thus, our conditions imply that $\mathcal{F}$ contains a β^* which correctly captures the conditional mean (and this will then be the risk minimizer); thus, as is indeed the case in Example 1, the regression function must be well-specified but the noise can be severely misspecified.

We stress that the three conditions have very different statuses. The first is mathematically convenient; it can be enforced by truncating parameters and data, which is awkward but may not lead to substantial deterioration in practice. Whether it is even really needed or not is not clear (and may in fact depend on the chosen exponential family). The second condition is really necessary — as can immediately be seen from Definition 1, the strong central condition cannot hold if Y has polynomial tails and for some f and x, $\ell_f(x, Y)$ increases polynomially in Y (in Section 6 of their paper, GM consider weakenings of the central condition that still work in such situations). For the third condition, however, we suspect that there are many cases in which it does not hold yet still the strong central condition holds; so then the GM convergence result would still be applicable under 'full misspecification'; investigating this will be the subject of future work.

GLM Information Complexity To apply Theorem 1 to get convergence bounds for exponential families and GLMs, we need to verify that the central condition holds (which we just did) and we need to bound the information complexity, which we proceed to do now. It turns out that the bound on $\mathrm{IC}_{n,\eta}$ of $O((d/n)\log n)$ of (4) continues to hold unchanged under misspecification, as is an immediate corollary of applying the following proposition to the definition of $\mathrm{IC}_{n,\eta}$ given above (3):

Proposition 2. *Let $(P, \ell, \mathcal{F})$ be a learning problem with $\mathcal{F}$ a GLM satisfying Conditions 1–3 above. Then for all $f \in \mathcal{F}$, $\mathbf{E}_{X,Y \sim P}[L_f] = \mathbf{E}_{X,Y \sim P_{f^*}}[L_f]$.*

This result follows almost immediately from the 'robustness property of exponential families' (Chapter 19 of Grünwald (2007)); for convenience we provide a proof in Appendix B. The result implies that any bound in $\mathrm{IC}_{n,\eta}(\Pi_0)$ for a particular prior in the well-specified GLM case, in particular (4), immediately transfers to the same bound for the misspecified case, as long as our regularity conditions hold, allowing us to apply Theorem 1 to obtain the parametric rate for GLMs under misspecification.

4 MCMC SAMPLING

Below we devise MCMC algorithms for obtaining samples from the η-generalized posterior distribution for two problems: regression and classification. In the regression context we consider one of the most commonly used sparse parameter estimation techniques, the lasso. For classification we use the logistic regression model. In our experiments in Section 5, we compare the performance of generalized Bayesian lasso with Horseshoe regression (Carvalho et al., 2010). The derivations of samplers are given in Appendix E.

4.1 Bayesian lasso regression

Consider the regression model $Y = X\beta + \varepsilon$, where $\beta \in \mathbb{R}^p$ is the vector of parameters of interest, $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is a noise vector. The Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani (1996) is a regularization method used in regression problems for shrinkage and selection of features. The lasso estimator is defined as $\hat{\beta}_{\text{lasso}} \coloneqq \arg\min_\beta \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$, where $\|\cdot\|_1, \|\cdot\|_2$ are l_1 and l_2 norms correspondingly. It can be interpreted as a Bayesian posterior mode (MAP) estimate when the priors on β are given by independent Laplace distributions. As discovered by Park and Casella (2008), the same posterior on β is also obtained by the following Gibbs sampling scheme: set $\eta = 1$ and denote $D_\tau \coloneqq \mathrm{diag}(\tau_1, \dots, \tau_n)$. Also, let $a \coloneqq \frac{\eta}{2}(n-1) + \frac{p}{2} + \alpha$ and $b_\tau \coloneqq \frac{\eta}{2}(Y - X\beta)^T(Y - X\beta) + \frac{1}{2}\beta^T {D_\tau}^{-1}\beta + \gamma$, where $\alpha, \gamma > 0$ are hyperparameters. Then the Gibbs sampler is constructed as follows.

$$\beta \sim \mathcal{N}\left(\eta M_\tau X^T Y, \sigma^2 M_\tau\right),$$

$$\sigma^2 \sim \text{Inv-Gamma}\left(a, b_\tau\right), \quad \tau_j^{-2} \sim \text{IG}\left(\sqrt{\lambda^2\sigma^2/\beta_j^2}, \lambda^2\right),$$

where IG is the inverse Gaussian distribution and $M_\tau \coloneqq (\eta X^T X + D_\tau{}^{-1})^{-1}$. Following Park and Casella (2008), we put a Gamma prior on the shrinkage parameter λ. Now, in their paper Park and Casella only give the scheme for $\eta = 1$, but, as is straightforward to derive from their paper, the scheme above actually gives the η-*generalized* posterior corresponding to the lasso prior for general η (more details in Appendix E). We will use the Safe-Bayesian algorithm for choosing the optimal η developed by Grünwald and Van Ommen (2017) (see Appendix E.3). The code for Generalized- and Safe-Bayesian lasso regression can be found in the CRAN R-package 'SafeBayes' (De Heide, 2016).

Horseshoe estimator The Horseshoe prior is the state-of-the-art global-local shrinkage prior for tackling high-dimensional regularization, introduced by Carvalho et al. (2010). Unlike the Bayesian lasso, it has flat Cauchy-like tails, which allow strong signals to remain unshrunk a posteriori. For completeness we include the horseshoe in our regression comparison, using the implementation of Van der Pas et al. (2016).

4.2 Bayesian logistic regression

Consider the standard logistic regression model $\{f_\beta : \beta \in \mathbb{R}^p\}$, the data $Y_1, \ldots, Y_n \in \{0, 1\}$ are independent binary random variables observed at the points $X \coloneqq (X_1, \ldots, X_n) \in \mathbb{R}^{n \times p}$ with $P_{f_\beta}(Y_i = 1 \mid X_i) \coloneqq p_{f_\beta}(1 \mid X_i) \coloneqq \frac{e^{X_i^T \beta}}{1+e^{X_i^T \beta}}$. The standard Bayesian approach involves putting a Gaussian prior on the parameter $\beta \sim \mathcal{N}(b, B)$ with mean $b \in \mathbb{R}^p$ and the covariance matrix $B \in \mathbb{R}^{p \times p}$. To sample from the η-generalized posterior we modify a Pólya–Gamma latent variable scheme described in Polson et al. (2013). We first introduce latent variables $\omega_1, \ldots, \omega_n \in \mathbb{R}$, which will be sampled from Pòlya-Gamma distribution (constructed to yield a simple Gibbs sampler for Bayesian logistic regression, for more details see Polson et al. (2013)). Let $\Omega \coloneqq \text{diag}\{\omega_1, \ldots, \omega_n\}$, $\kappa \coloneqq (Y_1 - 1/2, \ldots, Y_n - 1/2)^T$, $V_\omega \coloneqq (X^T \Omega X + B^{-1})^{-1}$, and $m_\omega \coloneqq V_\omega(\eta X^T \kappa + B^{-1} b)$. Then the Gibbs sampler for η-generalized posterior is given by $\omega_i \sim \text{PG}(\eta, X_i^T \beta)$, $\beta \sim \mathcal{N}(m_\omega, V_\omega)$, where PG is the Pòlya-Gamma distribution.

5 EXPERIMENTS

Below we present the results of experiments that compare the performance of the derived Gibbs samplers with their standard counterparts. More details/experiments are in Appendix F.

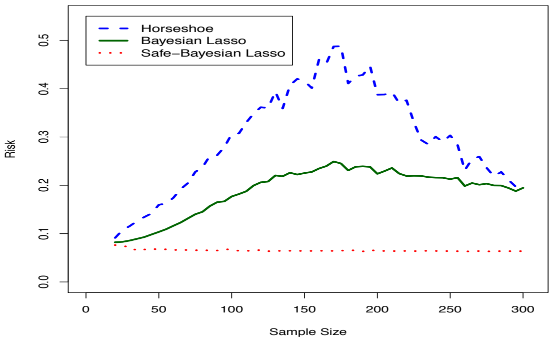

Figure 3: *Simulated squared error risk (test error) with respect to P as function of sample size for the* wrong-model *experiments of Section 5.1 using the posterior predictive distribution of the standard Bayesian lasso (green, solid), the Safe-Bayesian lasso (red, dotted), both with standard improper priors, and the Horseshoe (blue, dashed); and* 201 *Fourier basis functions.*

5.1 Simulated data

Regression In our experiments we focus on prediction, and we run simulations to determine the *square-risk* (expected squared error loss) of our estimate relative to the underlying distribution P: $\mathbf{E}_{(X,Y)\sim P}(Y - X\beta)^2$, where $X\beta$ would be the conditional expectation, and thus the square-risk minimizer, if β would be the true parameter (vector).

Consider the data generated as described in Example 1. We study the performance of the η-generalized Bayesian lasso with η chosen by the Safe-Bayesian algorithm (we call it the Safe-Bayesian lasso) in comparison with two popular estimation procedures for this context: the Bayesian lasso (which corresponds to η=1), and the Horseshoe method. In Figure 3 the simulated square-risk is plotted as a function of the sample size for all three methods. We average over enough samples so that the graph appears to be smooth (25 iterations for SafeBayes, 1000 for the two standard Bayesian methods). It shows that both the standard Bayesian lasso and the Horseshoe perform significantly worse than the Safe-Bayesian lasso. Moreover we see that the risks for the standard methods initially grows with the sample size (additional experiments not reported here suggest that Bayes will 'recover' at very large n).

Classification We focus on finding coefficients β for prediction, and our error measure is the expected logarithmic loss, which we call *log-risk*: $\mathbf{E}_{(X,Y)\sim P}\left[-\log \text{Li}_\beta(Y \mid X)\right]$, where $\text{Li}_\beta(Y \mid X) \coloneqq$

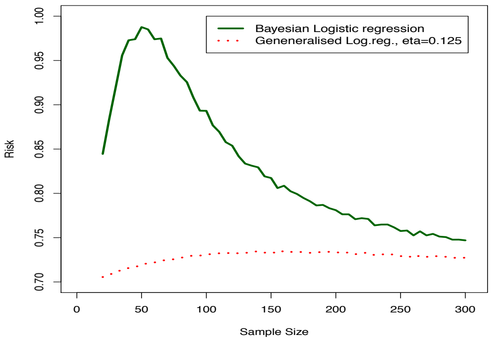

Figure 4: *Simulated logistic risk as function of sample size for* wrong-model *experiments of Section 5.1 using posterior predictive distribution of standard Bayesian logistic regression (green, solid), and generalized Bayes (*$\eta = 0.125$*, red, dotted) with* 25 *noise dimensions.*

$e^{YX^T\beta}/(1 + e^{X^T\beta})$. We start with an example that is very similar to the previous one. We generate a $n \times p$ matrix of independent standard normal random variables with $p = 25$. For every feature vector X_i we sample a corresponding $Z_i \sim \mathcal{N}(0, \sigma^2)$, as before, and we misspecify the model by putting approximately half of the Z_i and the corresponding $X_{i,1}$ to zero. Next, we sample the labels $Y_i \sim \text{Binom}(\exp(Z_i)/(1 + \exp(Z_i))$. We compare standard Bayesian logistic regression ($\eta = 1$) to a generalized version ($\eta = 0.125$). In Figure 4 we plot the log-risk as a function of the sample size. As in the regression case, the risk for standard Bayesian logistic regression ($\eta = 1$) is substantially worse than the one for generalized Bayes ($\eta = 0.125$). Even for generalized Bayes, the risk initially goes up a little bit, the reason being that the prior is *too good*: it is strongly concentrated around the risk-optimal $\beta^* = 0$. Thus, the first prediction made by the Bayesian predictive distribution coincides with the optimal ($\beta = 0$) prediction, and in the beginning, due to noise in the data, predictions will first get slightly worse. This is a phenomenon that also applies to standard Bayes with well-specified models; see for example (Grünwald and Halpern, 2004, Example 3.1).

Even for the well-specified case it can be beneficial to use $\eta \neq 1$. It is easy to see that the maximum *a posteriori* estimate for generalized logistic regression corresponds to the ridge logistic regression method (which penalizes large $\|\beta\|_2$) with the shrinkage parameter $\lambda = \eta^{-1}$. However, when the the prior mean is zero but the risk minimizer β^* is far from zero, penalizing large norms of β is inefficient, and we find that the best performance is achieved with $\eta > 1$.

	Horse-shoe	Bayesian lasso	SafeBayes lasso
MSE ($(°C)^2$)	6.53	6.16	6.04
MSE ($(ppm)^2$)	1169	1201	1142

Table 1: *Mean square errors for predictions on the Seattle and London data sets of Section 5.2.*

5.2 Real World Data

We present two examples with real world data to demonstrate that bad behavior under misspecification also occurs in practice. For these data sets, we compare the performance of Safe-Bayesian lasso and standard Bayesian lasso. As the first example we consider the data of the daily maximum temperatures at Seattle Airport as a function of the time and date (source: R-package `weatherData`, also available at `www.wunderground.com`). A second example is London air pollution data (source: R-package `Openair`, for more details see Carslaw and Ropkins (2012); Carslaw (2015)). Here the quantity of interest is the concentration of nitrogen dioxide (NO_2), again as a function of time and date. In both settings we divide the data into a training set and a test set and focus on the prediction error. In both examples, SafeBayes picks an $\hat{\eta}$ strictly smaller than one. Also, for both data sets the Safe-Bayesian lasso clearly outperforms the standard Bayesian lasso and the Horseshoe in terms of mean square prediction error, as seen from Table 1 (details in Appendix F).

6 FUTURE WORK

We provided both theoretical and empirical evidence that η-generalized Bayes can significantly outperform standard Bayes for GLMs. However, the empirical examples are only given for Bayesian lasso linear regression and logistic regression. In future work we would like to devise generalized posterior samplers for other GLMs and speed up the sampler for generalized Bayesian logistic regression, since our current implementation is slow and (unlike our linear regression implementation) cannot deal with high-dimensional (and thus, real-world) data yet. Furthermore, the Safe-Bayesian algorithm of Grünwald (2012), used to learn η, enjoys good theoretical performance but is computationally very slow. Since learning η for which the central condition holds (preferably the largest possible value, since small values of η mean slower learning) is essential for using generalized Bayes in practice, there is a necessity for speeding up SafeBayes or finding an alternative. A potential solution might be using cross-validation to learn η, but its theoretical properties (e.g. satisfying the central condition) are yet to be established.

Bibliography

O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* Wiley, Chichester, UK, 1978.

A. R. Barron and T. M. Cover. Minimum complexity density estimation. *Information Theory, IEEE Transactions on*, 37(4):1034–1054, 1991.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

G. de los Campos, D. Naya, J. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. Cotes. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182:375–385, 2009.

D. C. Carslaw. *The openair manual - open-source tools for analysing air pollution data. Manual for version 1.1-4.* King's College London, 2015.

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

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games.* Cambridge University Press, Cambridge, UK, 2006.

T. van. Erven, P. Grünwald, and S. de. Rooij. Catching up faster in Bayesian model selection and model averaging. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

T. van Erven, P. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

P. Grünwald and J. Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2-3):119–149, 2007. DOI 10.1007/s10994-007-0716-7.

P. Grünwald and T. van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

P. D. Grünwald. *The Minimum Description Length Principle.* MIT Press, Cambridge, MA, 2007.

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

P. D. Grünwald and J. Y. Halpern. When ignorance is bliss. In *Proceedings of the Twentieth Annual Conference on Uncertainty in Artificial Intelligence (UAI 2004)*, Banff, Canada, July 2004.

P. D. Grünwald and N. A. Mehta. Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *Journal of Machine Learning Research*, 2019. Accepted pending minor modifications; available as arXiv preprint arXiv:1605.00252.

R. de Heide. *SafeBayes: Generalized and Safe-Bayesian Ridge and Lasso Regression*, 2016. URL `https://CRAN.R-project.org/package=SafeBayes`. R package version 1.1.

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

Q. J. Li. *Estimation of mixture models.* PhD thesis, Yale University, 1999.

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

P. McCullagh and J. Nelder. *Generalized Linear Models.* Chapman and Hall/CRC, Boca Raton, second edition, 1989.

U. K. Müller. Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849, 2013.

R. Narasimhan. Package weatherdata, get weather data from the web., 06 2014. URL: http://ram-n.github.io/weatherData/.

T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103:369–412, 2008.

S. van der Pas, J. Scott, A. Chakraborty, and A. Bhattacharya. horseshoe: Implementation of the Horseshoe prior. *R package version 0.1. 0*, 2016.

N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

N. Syring and R. Martin. Calibrating general posterior credible regions. *arXiv preprint arXiv:1509.00922*, 2017.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. ISSN 0035-9246.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

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

J. Windle, N. G. Polson, and J. G. Scott. Sampling pólya-gamma random variates: alternate and approximate techniques. *arXiv preprint arXiv:1405.0506*, 2014.

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

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*, 2018. Advance publication; Number and pages to be announced.

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

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006b.