Universiteit
Leiden
The Netherlands

Bayesian learning: Challenges, limitations and pragmatics

Heide, R. de

Citation
Heide, R. de. (2021, January 26). *Bayesian learning: Challenges, limitations and pragmatics*. Retrieved from https://hdl.handle.net/1887/3134738

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3134738

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle https://hdl.handle.net/1887/3134738 holds various files of this Leiden University dissertation.

Author: Heide, R. de
Title: Bayesian learning: Challenges, limitations and pragmatics
Issue Date: 2021-01-26

Chapter 6

Safe-Bayesian generalized linear regression

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.

6.1 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, Grünwald and Rooij, 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 (6.1) 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 6.5.1 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 $\overline{\eta}$-*central condition* (Definition 6.1 below), generalized Bayes with a specific finite learning rate $\overline{\eta}$ (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 6.2.1 below, we give (a) a toy example on which the Bayesian lasso and the Horseshoe estimator fail; later in the chapter, in Section 6.5 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 much better performance. In Section 6.3, 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 6.5. 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 6.1 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 6.4) MCMC algorithms (f) for generalized Bayes posterior sampling for Bayesian lasso and logistic regression. (a)-(f) are all novel contributions.

In Section 6.2 we first define our setting more precisely. Section 6.2.1) gives a first example of bad standard-Bayesian behavior and Section 6.2.2) recalls a theorem from GM indicating that under the $\overline{\eta}$-central condition, generalized Bayes for $\eta < \overline{\eta}$ should perform well. We

present our new theoretical results in Section 6.3. We next (Section 6.4), present our algorithms for generalized Bayesian posterior sampling, and we continue (Section 6.5) to empirically demonstrate how generalized Bayes outperforms standard Bayes under misspecification. All proofs are in Appendix 6.A.

6.2 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 $\ell_f(x, y) = \log(1 + \exp(-y f(x))$. 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

$$\pi_n(f) \coloneqq \frac{\exp\left(-\eta \sum_{i=1}^n \ell_f(z_i)\right) \cdot \pi_0(f)}{\int_{\mathcal{F}} \exp\left(-\eta \sum_{i=1}^n \ell_f(z_i)\right) \cdot \pi_0(f) \mathrm{d}\rho(f)}, \tag{6.1}$$

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

$$\pi_n(f) \propto \prod_{i=1}^n (p_f(y_i \mid x_i))^\eta \pi_0(f), \tag{6.2}$$

which, if $\eta = 1$, reduces to standard Bayesian inference. While GM's result (quoted as Theorem 6.1 below) works for arbitrary loss functions, Theorem 6.2 and our empirical simulations (this chapter's new results) revolve around (generalized) linear models. For these models, (6.1) 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 (6.1) induces the same posterior distribution $\pi_n(f)$ over $\mathcal{F}$ as does (6.2) with the conditional distributions $p_f(y|x) \propto \exp(-(y - f(x))^2$, which is again the same as (6.1) 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 (6.2) 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, Bousquet and Mendelson (2005)), we assume throughout that there exists an optimal $f^* \in \mathcal{F}$ that achieves the smallest *risk* (expected loss) $\mathbb{E}[\ell_{f^*}(Z)] = \inf_{f \in \mathcal{F}} \mathbb{E}[\ell_f(Z)]$. 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 $\mathbf{E}_{(X,Y)\sim P}[\log\big(p(Y|X)/p_f(Y|X)\big)]$ 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 $f^* = f$.

6.2.1 Bad Behavior of Standard Bayes

Example 6.1. 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 $X_i' \overset{i.i.d.}{\sim}$ Uniform$([-1,1])$; the dependent variable Y_i' is set to $Y_i' = 0 + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ 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 $(X_i, Y_i) := (X_i', Y_i')$, i.e. we keep the (X_i', Y_i') as they are, and if the coin lands tails, we put the pair to zero: $(X_i, Y_i) := (0, 0)$.

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 (6.1) with $\ell_{f_\beta}(x, y) = (y - f_\beta(x))^2$; the prior is the Bayesian lasso prior whose definition we recall in Section 6.4.1. 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 6.1, it overfits and learns the noise instead (later on (Figure 6.3 in Section 6.5.1) 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 $Y_i = 0 + \varepsilon_i$, 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 (6.1) 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 6.1. Implementation details are in Section 6.4.1 and Appendix 6.D; the details of the figure are in Appendix 6.E.

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

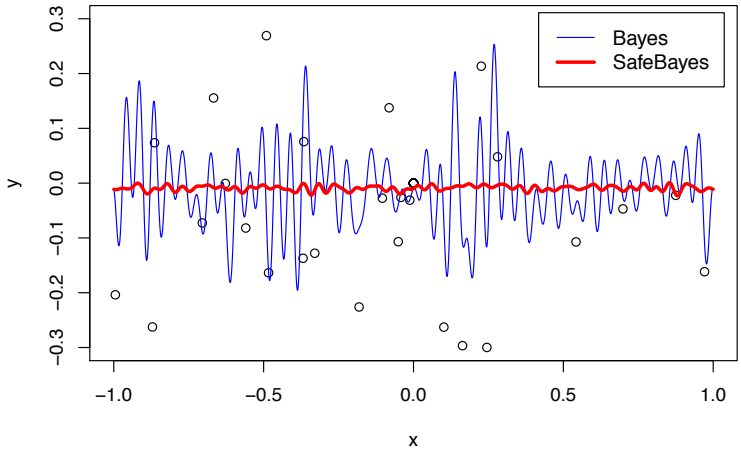

Figure 6.1: *Predictions of standard Bayes (blue) and SafeBayes (red),* $n = 50$, $p = 101$.

and Langford (2007) to properties of the Bayesian predictive distribution

$$\overline{p}(Y_{n+1} \mid X_{n+1}, Z^n) := \int_{\mathcal{F}} p_f(Y_{n+1} \mid X_{n+1}) \pi_n(f \mid Z^n) \mathrm{d}\rho(f).$$

Being a mixture of $f \in \mathcal{F}$, $\overline{p}(Y_{n+1} \mid X_{n+1})$, 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 $\overline{p}(Y_{n+1} \mid X_{n+1}, Z^n)$ 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 6.2 (details in Appendix 6.E). 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 $\overline{p}(Y_{n+1} \mid X_{n+1}, Z^n)$ 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 $\overline{p}$ is 'proper' (Shalev-Shwartz and Ben-David, 2014), and the overfitting behavior then does not occur anymore.

6.2.2 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* $d_{\overline{\eta}}$

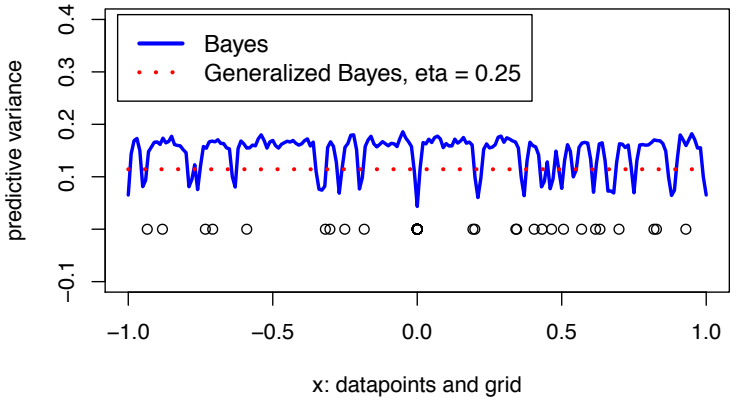

Figure 6.2: *Variance of Predictive Distribution* $\overline{p}(Y_{n+1} \mid X_{n+1}, Z^n)$ *for a single run with* $n = 50$.

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 *misspecification metric* $d_{\overline{\eta}}$ in terms of its square by

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

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

$$p_{f,\overline{\eta}}(z) := p(z) \frac{\exp(-\overline{\eta} L_f(z))}{\mathbb{E}[\exp(-\overline{\eta} L_f(Z))]},$$

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

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

$$\mathrm{IC}_{n,\eta}(\Pi_0) := \mathbf{E}_{\underline{f} \sim \Pi_n} \left[\frac{1}{n} \sum_{i=1}^n L_{\underline{f}}(Z_i)\right] + \frac{KL(\Pi_n \,\|\, \Pi_0)}{\eta \cdot n} =$$
$$-\frac{1}{\eta n} \log \int_{\mathcal{F}} \pi_0(f) e^{-\eta \sum_{i=1}^n \ell_f(Z_i)} \mathrm{d}\rho(f) - \sum_{i=1}^n \ell_{f^*}(Z_i), \tag{6.3}$$

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 $\overline{\eta} > 0$, for all fixed f, the probability that the loss of f exceeds that of the optimal f^* by $a/\overline{\eta}$ is exponentially small in a:

Definition 6.1 (Central Condition, Def. 7 of GM)**.** Let $\overline{\eta} > 0$. We say that $(P, \ell, \mathcal{F})$ satisfies the *$\overline{\eta}$-strong central condition* if, for all $f \in \mathcal{F}$: $\mathbb{E}\left[e^{-\overline{\eta} L_f}\right] \leq 1$.

As straightforward rewriting shows, this condition holds *automatically*, for any $\overline{\eta} \leq 1$ 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, Bousquet and Mendelson, 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 6.1 (Theorem 10 from GM)**.** *Suppose that the $\overline{\eta}$-strong central condition holds. Then for any $0 < \eta < \overline{\eta}$, the metric $d_{\overline{\eta}}$ satisfies*

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

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

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 $d^2_{\overline{\eta}}$ on the left hand side is a weak metric, however, in Appendix 6.B we show that we can replace it by stronger notions such as KL-divergence, squared error or logistic loss. Theorem 6.1 generalizes previous results (e.g. Zhang (2006a) and Zhang (2006b)) to the misspecified setting. In the well-specified case, Zhang, as well as several other authors (Walker and Hjort, 2002; Martin, Mess and Walker, 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 $\overline{\eta}$ 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 $\overline{\eta}$ 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

$$\mathbf{E}_{Z^n \sim P}\left[\mathrm{IC}_{n,\eta}(\Pi_0)\right] = O\left(\frac{d}{n} \cdot \log n\right), \tag{6.4}$$

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 3 below we show that, for GLMs, they continue to hold for the misspecified case.

6.3 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

$$p_\theta(y) := \exp(\theta y - F(\theta) + r(y)), \tag{6.5}$$

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 $[\underline{\theta}, \overline{\theta}]$ with $\underline{\theta} < \overline{\theta}$ that resides in the interior of the natural parameter space $\overline{\Theta} = \{\theta : F(\theta) < \infty\}$. 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 (6.7) in Appendix 4.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 $\mu : \overline{\Theta} \to \mathbb{R}$ such that $\mathbb{E}_{Y\sim P_\theta}[Y] = \mu(\theta)$. 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 6.2. *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 $\Theta = [\underline{\theta}, \overline{\theta}]$ is compact as above and that $\theta^* = \arg\min_{\theta\in\overline{\Theta}} D(P\|P_\theta)$ lies in Θ. Let $\sigma^2 > 0$ be the true variance $\mathbb{E}_{Y\sim P}(Y - \mathbb{E}[Y])^2$ and let $(\sigma^*)^2$ be the variance $\mathbb{E}_{Y\sim P_{\theta^*}}(Y - \mathbb{E}[Y])^2$ according to θ^*. Then

- *(i) for all $\overline{\eta} > (\sigma^*)^2/\sigma^2$, the $\overline{\eta}$-central condition does* not *hold.*
- *(ii) Suppose there exists $\eta^\circ > 0$ such that $\overline{C} := \mathbb{E}_P[\exp(\eta^\circ|Y|)] < \infty$. Then there exists $\overline{\eta} > 0$, depending only on η°, $\overline{C}$, $\underline{\theta}$ and $\overline{\theta}$ such that the $\overline{\eta}$-central condition holds. Moreover,*
- *(iii), for all $\delta > 0$, there is an $\varepsilon > 0$ such that, for all $\overline{\eta} \le (\sigma^*)^2/\sigma^2 - \delta$, the $\overline{\eta}$-central condition holds relative to the restricted model $\mathcal{F}_\varepsilon = \{p_\theta : \theta \in [\theta^* - \varepsilon, \theta^* + \varepsilon]\}$.*

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

We provide (iii) just to give insight — 'locally', i.e. in restricted models that are small neighborhoods around the best-approximating θ^*, the smallest $\overline{\eta}$ 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 $\overline{\eta} < 1$ (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 $\overline{\eta} > 1$; we give a practical example at the end of Section 6.5. Theorem 6.2 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 $\mathcal{Q} = \{p_\theta : \theta \in \overline{\Theta}\}$ on $\mathcal{Y}$ of the form (6.5) 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

$$p_\beta(y \mid x) := \exp\big(\theta_x(\beta) y - F(\theta_x(\beta)) + r(y)\big), \tag{6.6}$$

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

Proposition 3. *Under the following three assumptions, the learning problem $(P, \ell, \mathcal{F})$ with $\mathcal{F}$ as above satisfies the $\overline{\eta}$-central condition for some $\overline{\eta} > 0$ 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 $\{\mu \in \mathbb{R} : \mu = \mathbb{E}_{Y \sim q}[Y] \ : \ q \in \mathcal{Q}\}$ (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 $\sup_{x \in \mathcal{X}} \mathbb{E}_{Y \sim P}[\exp(\eta |Y|) \mid X = x] < \infty$.*
3. *(Well-specification of conditional mean): there exists $\beta^\circ \in \mathcal{B}$ such that $\mathbb{E}[Y \mid X] = g^{-1}(\langle \beta^\circ, X \rangle)$.*

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 6.1, 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 6.1, 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 6.1 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 (6.4) 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 (6.3):

Proposition 4. *Let $(P, \ell, \mathcal{F})$ be a learning problem with $\mathcal{F}$ a GLM satisfying Conditions 1–3 above. Then for all $f \in \mathcal{F}$, $\mathbb{E}_{X,Y\sim P}[L_f] = \mathbb{E}_{X,Y\sim P_{f^*}}[L_f]$.*

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 4.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 (6.4), immediately transfers to the same bound for the misspecified case, as long as our regularity conditions hold, allowing us to apply Theorem 6.1 to obtain the parametric rate for GLMs under misspecification.

6.4 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 6.5, we compare the performance of generalized Bayesian lasso with Horseshoe regression (Carvalho, Polson and Scott, 2010). The derivations of samplers are given in Appendix 6.D.

6.4.1 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 $\widehat{\beta}_{\text{lasso}} \coloneqq \arg\min_\beta \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$, 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 $D_\tau \coloneqq \operatorname{diag}(\tau_1, \ldots, \tau_n)$. 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.

$$\begin{aligned}
\beta &\sim \mathcal{N}\left(\eta M_\tau X^T Y, \sigma^2 M_\tau\right), \\
\sigma^2 &\sim \text{Inv-Gamma}\left(a, b_\tau\right), \\
\tau_j^{-2} &\sim \text{IG}\left(\sqrt{\lambda^2\sigma^2/\beta_j^2}, \lambda^2\right),
\end{aligned}$$

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 6.D). We will use the Safe-Bayesian algorithm for choosing the optimal η developed by Grünwald and Van Ommen (2017) (see Appendix 6.D.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, Polson and Scott (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).

6.4.2 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, Scott and Windle (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, Scott and Windle (2013)). Let

$$\begin{aligned}\Omega &\coloneqq \operatorname{diag}\{\omega_1, \dots, \omega_n\},\\ \kappa &\coloneqq (Y_1 - 1/2, \dots, Y_n - 1/2)^T,\\ V_\omega &\coloneqq (X^T \Omega X + B^{-1})^{-1}, \text{ and}\\ m_\omega &\coloneqq V_\omega(\eta X^T \kappa + B^{-1} b).\end{aligned}$$

Then the Gibbs sampler for η-generalized posterior is given by

$$\omega_i \sim \mathrm{PG}(\eta, X_i^T \beta), \quad \beta \sim \mathcal{N}(m_\omega, V_\omega),$$

where PG is the Pòlya-Gamma distribution.

6.5 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 6.E.

6.5.1 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: $\mathbb{E}_{(X,Y)\sim P}(Y - X\beta)^2$, 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 6.1. 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 6.3 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*: $\mathbb{E}_{(X,Y)\sim P}\left[-\log \mathrm{Li}_\beta(Y \mid X)\right]$, where $\mathrm{Li}_\beta(Y \mid X) \coloneqq e^{YX^T\beta}/(1 + e^{X^T\beta})$. 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 $Y_i \sim \mathrm{Binom}(\exp(Z_i)/(1 + \exp(Z_i))$. We compare standard Bayesian logistic regression ($\eta = 1$) to a generalized version ($\eta = 0.125$). In Figure 6.4 we plot

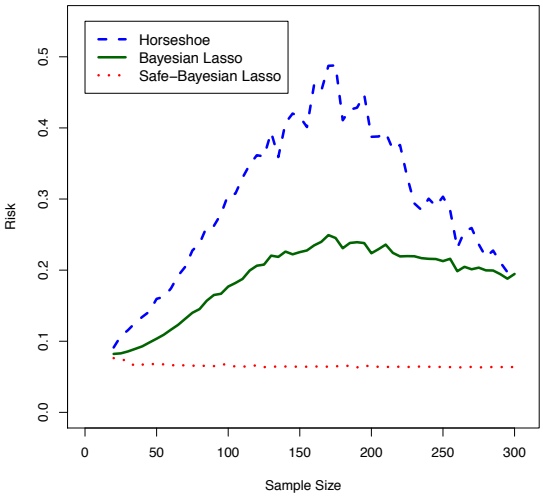

Figure 6.3: *Simulated squared error risk (test error) with respect to P as function of sample size for the* wrong-model *experiments of Section 6.5.1 using the posterior predictive distribution of the standard Bayesian lasso (green, solid), the Safe-Bayesian lasso (red, dotted), both with standard improper priors, and the Horseshoe (blue, dashed); and* 201 *Fourier basis functions.*

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 $\eta > 1$.

6.5.2 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

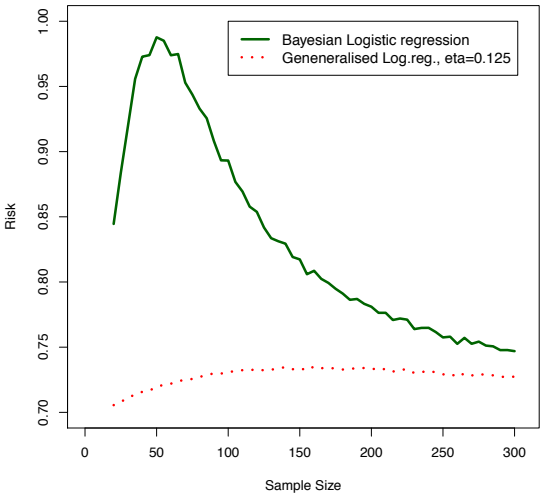

Figure 6.4: *Simulated logistic risk as function of sample size for* wrong-model *experiments of Section 6.5.1 using posterior predictive distribution of standard Bayesian logistic regression (green, solid), and generalized Bayes ($\eta = 0.125$, red, dotted) with 25 noise dimensions.*

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

Table 6.1: Mean square errors for predictions on the Seattle and London data sets of Section 6.5.2

London air pollution data (source: `R`-package `Openair`, for more details see Carslaw and Ropkins (2012) and 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 $\widehat{\eta}$ 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 6.1 (details in Appendix 6.E).

6.6 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.

6.A Proofs

6.A.1 Proof of Theorem 6.2

The second part of the theorem about the Gaussian location family is a straightforward calculation, which we omit. As to the first part (Part (i)—(iii)), we will repeatedly use the following fact: for every Θ that is a nonempty compact subset of the interior of $\overline{\Theta}$, in particular for $\Theta = [\underline{\theta}, \overline{\theta}]$ with $\underline{\theta} < \overline{\theta}$ both in the interior of $\overline{\Theta}$, we have:

$$\begin{aligned} -\infty < \inf_{\theta\in\Theta} F(\theta) < \sup_{\theta\in\Theta} F(\theta) < \infty \\ -\infty < \inf_{\theta\in\Theta} F'(\theta) < \sup_{\theta\in\Theta} F'(\theta) < \infty \\ 0 < \inf_{\theta\in\Theta} F''(\theta) < \sup_{\theta\in\Theta} F''(\theta) < \infty. \end{aligned} \tag{6.7}$$

Now, let $\theta, \theta^* \in \Theta$. We can write

$$\mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] = \mathbb{E}_{Y\sim P}\left[\left(\frac{p_\theta(Y)}{p_{\theta^*}(Y)}\right)^\eta\right] = \exp\left(-G(\eta(\theta-\theta^*)) + \eta F(\theta^*) - \eta F(\theta)\right). \tag{6.8}$$

where $G(\lambda) = -\log \mathbb{E}_{Y\sim P}\left[\exp(\lambda Y)\right]$. If this quantity is $-\infty$ for all $\eta > 0$, then (i) holds trivially. If not, then (i) is implied by the following statement:

$$\limsup_{\varepsilon\to 0}\left\{\eta : \text{for all } \theta \in [\theta^* - \varepsilon, \theta^* + \varepsilon],\ \mathbb{E}[\exp(\eta L_{p_\theta})] \le 1\right\} = \frac{(\sigma^*)^2}{\sigma^2}. \tag{6.9}$$

Clearly, this statement also implies (iii). To prove (i), (ii) and (iii), it is thus sufficient to prove (ii) and (6.9). We prove both by a second-order Taylor expansion (around θ^*) of the right-hand side of (6.8).

Preliminary Facts. By our assumption there is a $\eta^\circ > 0$ such that $\mathbb{E}[\exp(\eta^\circ|Y|)] = \overline{C} < \infty$. Since $\theta^* \in \Theta = [\underline{\theta}, \overline{\theta}]$ we must have for every $0 < \eta < \eta^\circ/(2|\overline{\theta} - \underline{\theta}|)$, every $\theta \in \Theta$,

$$\begin{aligned} \mathbb{E}[\exp(2\eta(\theta-\theta^*)\cdot Y)] &\le \mathbb{E}[\exp(2\eta|\theta-\theta^*|\cdot|Y|)] \\ &\le \mathbb{E}[\exp(\eta^\circ(|\theta-\theta^*|/|\overline{\theta}-\underline{\theta}|)\cdot|Y|)] \\ &\le \overline{C} \\ &< \infty. \end{aligned} \tag{6.10}$$

The first derivative of the right of (6.8) is:

$$\eta\mathbb{E}\left[(Y - F'(\theta))\exp\Big(\eta\big((\theta-\theta^*)Y + F(\theta^*) - F(\theta)\big)\Big)\right]. \tag{6.11}$$

The second derivative is:

$$\mathbb{E}\left[\left(-\eta F''(\theta) + \eta^2(Y - F'(\theta))^2\right)\cdot\exp\Big(\eta\big((\theta-\theta^*)Y + F(\theta^*) - F(\theta)\big)\Big)\right]. \tag{6.12}$$

We will also use the standard result (Grünwald, 2007; Barndorff-Nielsen, 1978) that, since we assume $\theta^* \in \Theta$,

$$\mathbb{E}[Y] = \mathbb{E}_{Y\sim P_{\theta^*}}[Y] = \mu(\theta^*); \quad \text{for all } \theta \in \overline{\Theta}\colon F'(\theta) = \mu(\theta); \quad F''(\theta) = \mathbb{E}_{Y\sim P_\theta}(Y - E(Y))^2, \tag{6.13}$$

the latter two following because F is the cumulant generating function.

Part (ii). We use an exact second-order Taylor expansion via the Lagrange form of the remainder. We already showed there exist $\eta' > 0$ such that, for all $0 < \eta \le \eta'$, all $\theta \in \Theta$, $\mathbb{E}[\exp(2\eta(\theta - \theta^*)Y)] < \infty$. Fix any such η. For some $\theta' \in \{(1-\alpha)\theta + \alpha\theta^* \colon \alpha \in [0,1]\}$, the (exact) expansion is:

$$\begin{aligned}\mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] = 1 + \eta(\theta - \theta^*)\mathbb{E}\left[Y - F'(\theta^*)\right] - \frac{\eta}{2}(\theta - \theta^*)^2 F''(\theta') \ldots \\ \ldots \cdot \mathbb{E}\left[\exp\Big(\eta\big((\theta' - \theta^*)Y + F(\theta^*) - F(\theta')\big)\Big)\right] \ldots \\ \ldots + \frac{\eta^2}{2}(\theta - \theta^*)^2 \mathbb{E}\left[(Y - F'(\theta'))^2 \cdot \exp\Big(\eta\big((\theta' - \theta^*)Y + F(\theta^*) - F(\theta')\big)\Big)\right].\end{aligned}$$

Defining $\Delta = \theta' - \theta$, and since $F'(\theta^*) = \mathbb{E}[Y]$ (see (6.13)), we see that the central condition is equivalent to the inequality:

$$\eta\mathbb{E}\left[(Y - F'(\theta'))^2 e^{\eta\Delta Y}\right] \le F''(\theta')\mathbb{E}\left[e^{\eta\Delta Y}\right].$$

From Cauchy-Schwarz, to show that the η-central condition holds it is sufficient to show that

$$\eta \left\|(Y - F'(\theta'))^2\right\|_{L_2(P)} \left\|e^{\eta\Delta Y}\right\|_{L_2(P)} \le F''(\theta')\mathbb{E}\left[e^{\eta\Delta Y}\right],$$

which is equivalent to

$$\eta \le \frac{F''(\theta')\mathbb{E}\left[e^{\eta\Delta Y}\right]}{\sqrt{\mathbb{E}\left[(Y - F'(\theta'))^4\right]\mathbb{E}\left[e^{2\eta\Delta Y}\right]}}. \tag{6.14}$$

We proceed to lower bound the RHS by lower bounding each of the terms in the numerator and upper bounding each of the terms in the denominator. We begin with the numerator. $F'(\theta)$ is bounded by (6.7). Next, by Jensen's inequality,

$$\mathbb{E}\left[\exp(\eta\Delta Y)\right] \ge \exp(\mathbb{E}[\eta\Delta \cdot Y]) \ge \exp(-\eta^\circ|\overline{\theta} - \underline{\theta}||\mu(\theta^*)|)$$

is lower bounded by a positive constant. It remains to upper bound the denominator. Note that the second factor is upper bounded by the constant $\overline{C}$ in (6.10). The first factor is bounded by a fixed multiple of $\mathbb{E}|Y|^4 + \mathbb{E}[F'(\theta)^4]$. The second term is bounded by (6.7), so it remains to bound the first term. By assumption $\mathbb{E}[\exp(\eta^\circ|Y|)] \le \overline{C}$ and this implies that $\mathbb{E}|Y^4| \le a^4 + \overline{C}$ for any $a \ge e$ such that $a^4 \le \exp(\eta^\circ a)$; such an a clearly exists and only depends on η°.

We have thus shown that the RHS of (6.14) is upper bounded by a quantity that only depends on $\overline{C}, \eta^\circ$ and the values of the extrema in (6.7), which is what we had to show.

Proof of (iii). We now use the asymptotic form of Taylor's theorem. Fix any $\eta > 0$, and pick any θ close enough to θ^* so that (6.8) is finite for all θ' in between θ and θ^*; such a $\theta \ne \theta^*$ must

exist since for any $\delta > 0$, if $|\theta - \theta^*| \leq \delta$, then by assumption (6.8) must be finite for all $\eta \leq \eta^\circ/\delta$. Evaluating the first and second derivative (6.11) and (6.12) at $\theta = \theta^*$ gives:

$$\begin{aligned}\mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] &= 1 + \eta(\theta - \theta^*)\mathbb{E}\left[Y - F'(\theta^*)\right] \ldots \\ \ldots &- \left(\frac{\eta}{2}(\theta - \theta^*)^2 F''(\theta^*) - \frac{\eta^2}{2}(\theta - \theta^*)^2 \cdot \mathbb{E}\left[(Y - F'(\theta^*))^2\right]\right) + h(\theta)(\theta - \theta^*)^2 \\ &= 1 - \frac{\eta}{2}(\theta - \theta^*)^2 F''(\theta^*) + \frac{\eta^2}{2}(\theta - \theta^*)^2 \mathbb{E}\left[(Y - F'(\theta^*))^2\right] + h(\theta)(\theta - \theta^*)^2,\end{aligned}$$

where $h(\theta)$ is a function satisfying $\lim_{\theta \to \theta^*} h(\theta) = 0$, where we again used (6.13), i.e. that $F'(\theta^*) = \mathbb{E}[Y]$. Using further that $\sigma^2 = \mathbb{E}\left[(Y - F'(\theta^*))^2\right]$ and $F''(\theta^*) = (\sigma^*)^2$, we find that $\mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] \leq 1$ iff

$$-\frac{\eta}{2}(\theta - \theta^*)^2(\sigma^*)^2 + \frac{\eta^2}{2}(\theta - \theta^*)^2\sigma^2 + h(\theta)(\theta - \theta^*)^2 \leq 0.$$

It follows that for all $\delta > 0$, there is an $\varepsilon > 0$ such that for all $\theta \in [\theta^* - \varepsilon, \theta^* + \varepsilon]$, all $\eta > 0$,

$$\frac{\eta^2}{2}\sigma^2 \leq \frac{\eta}{2}(\sigma^*)^2 - \delta \Rightarrow \mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] \leq 1 \tag{6.15}$$

$$\frac{\eta^2}{2}\sigma^2 \geq \frac{\eta}{2}(\sigma^*)^2 + \delta \Rightarrow \mathbb{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] \geq 1 \tag{6.16}$$

The condition in (6.15) is implied if:

$$0 < \eta \leq \frac{(\sigma^*)^2}{\sigma^2} - \frac{2\delta}{\eta\sigma^2}.$$

Setting $C = 4\sigma^2/(\sigma^*)^4$ and $\eta_\delta = (1 - C\delta)(\sigma^*)^2/\sigma^2$ we find that for any $\delta < (\sigma^*)^4/(8\sigma^2)$, we have $1 - C\delta \geq 1/2$ and thus $\eta_\delta > 0$ so that in particular the premise in (6.15) is satisfied for η_δ. Thus, for all small enough δ, both the premise and the conclusion in (6.15) hold for $\eta_\delta > 0$; since $\lim_{\delta \downarrow 0} \eta_\delta = (\sigma^*)^2/\sigma^2$, it follows that there is an increasing sequence $\eta_{(1)}, \eta_{(2)}, \ldots$ converging to $(\sigma^*)^2/\sigma^2$ such that for each $\eta_{(j)}$, there is $\varepsilon_{(j)} > 0$ such that for all $\theta \in [\theta^* - \varepsilon_{(j)}, \theta^* + \varepsilon_{(j)}]$, $\mathbb{E}\left[e^{-\eta_{(j)}(\ell_\theta - \ell_{\theta^*})}\right] \leq 1$. It follows that the lim sup in (6.9) is at least $(\sigma^*)^2/\sigma^2$. A similar argument (details omitted) using (6.16) shows that the lim sup is at most this value; the result follows.

6.A.2 Proof of Proposition 4

For arbitrary conditional densities $p'(y \mid x)$ with corresponding distribution $P' \mid X$ for which

$$\mathbb{E}_{P'}[Y|X] = g^{-1}(\langle \beta, X \rangle), \tag{6.17}$$

and densities $p_{f^*} = p_{\beta^*}$ and p_β with $\beta^*, \beta \in \mathcal{B}$, we can write:

$$\begin{aligned}\mathbb{E}_{X \sim P}\mathbb{E}_{Y \sim P'|X}\left[\log \frac{p_{\beta^*}(Y \mid X)}{p_\beta(Y|X)}\right] &= \mathbb{E}\mathbb{E}\left[(\theta_X(\beta^*) - \theta_X(\beta))Y - \log \frac{F(\theta_X(\beta^*))}{F(\theta_X(\beta))} \mid X\right] \\ &= \mathbb{E}_{X \sim P}\left[(\theta_X(\beta^*) - \theta_X(\beta))g^{-1}(\langle \beta, X \rangle_d) \ldots\right. \\ &\quad \left. \ldots - \log F(\theta_X(\beta^*)) + \log F(\theta_X(\beta)) \mid X\right],\end{aligned}$$

where the latter equation follows by (6.17). The result now follows because (6.17) both holds for the 'true' P and for P_{f^*}.

6.A.3 Proof of Proposition 3

The fact that under the three imposed conditions the $\overline{\eta}$-central condition holds for some $\overline{\eta} > 0$ is a simple consequence of Theorem 6.2: Condition 1 implies that there is some compact Θ such that for all $x \in \mathcal{X}$, $\beta \in \mathcal{B}$, $\theta_x(\beta) \in \Theta$. Condition 3 then ensures that $\theta_x(\beta)$ lies in the interior of this Θ. And Condition 2 implies that $\overline{\eta}$ in Theorem 6.2 can be chosen uniformly for all $x \in \mathcal{X}$.

6.B Excess risk and KL divergence instead of generalized Hellinger distance

The misspecification metric/generalized Hellinger distance $d_{\overline{\eta}}$ appearing in Theorem 6.1 is rather weak (it is 'easy' for two distributions to be close) and lacks a clear interpretation for general, non-logarithmic loss functions. Motivated by these facts, GM study in depth under what additional conditions the (square of this) metric can be replaced by a stronger and more readily interpretable divergence measure. They come up with a new, surprisingly weak condition, the *witness condition*, under which $d_{\overline{\eta}}$ can be replaced by the *excess risk* $\mathbf{E}_P[L_f]$, which is the additional risk incurred by f as compared to the optimal f^*. For example, with the squared error loss, this is the additional mean square error of f compared to f^*; and with (conditional) log-loss, it is the well-known *generalized KL divergence* $\mathbf{E}_{X,Y\sim P}[\log \frac{p_{f^*}(Y|X)}{p_f(Y|X)}]$, coinciding with standard KL divergence if the model is correctly specified. Bounding the excess risk is a standard goal in statistical learning theory; see for example (Bartlett, Bousquet and Mendelson, 2005; Van Erven et al., 2015).

The following definition appears (with substantial explanation including the reason for its name) as Definition 12 in GM:

Definition 6.2 (Empirical Witness of Badness)**.** We say that $(P, \ell, \mathcal{F})$ satisfies the (u, c)-*empirical witness of badness condition* (or *witness condition*) for constants $u > 0$ and $c \in (0, 1]$ if for all $f \in \mathcal{F}$

$$\mathbb{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbb{1}_{\{\cdot\}}\ell_f - \ell_{f^*} \le u\right] \ge c\mathbb{E}[\ell_f - \ell_{f^*}].$$

More generally, for a function $\tau : \mathbb{R}^+ \to [1, \infty)$ and constant $c \in (0, 1)$ we say that $(P, \ell, \mathcal{F})$ satisfies the (τ, c)-*witness condition* if for all $f \in \mathcal{F}$, $\mathbb{E}[\ell_f - \ell_{f^*}] < \infty$ and

$$\mathbb{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbb{1}_{\{\cdot\}}\ell_f - \ell_{f^*} \le \tau(\mathbb{E}[\ell_f - \ell_{f^*}])\right] \ge c\mathbb{E}[\ell_f - \ell_{f^*}].$$

It turns out that the (τ, c)-witness condition holds in many practical situations, including our GLM-under-misspecification setting. Before elaborating on this, let us review (a special case of) Theorem 12 of GM, which is the analogue of Theorem 6.1 but with the misspecification metric replaced by the excess risk.

First, let, for arbitrary $0 < \eta < \overline{\eta}$, $c_u := \frac{1}{c}\frac{\eta u+1}{1-\frac{\eta}{\overline{\eta}}}$. Note that for large u, c_u is approximately linear in u/c.

Theorem 6.5. [Specialization of Theorem 12 of GM] *Consider a learning problem $(P, \ell, \mathcal{F})$. Suppose that the $\overline{\eta}$-strong central condition holds. If the (u, c)-witness condition holds, then for any $\eta \in (0, \overline{\eta})$,*

$$\mathbf{E}_{Z^n\sim P}\mathbb{E}_{\underline{f}\sim\Pi_n}\left[\mathbb{E}[L_f]\right] \le c_u \cdot \mathbf{E}_{Z^n\sim P}\left[\mathrm{IC}_{n,\eta}\left(\Pi_0\right)\right], \tag{6.18}$$

with c_u as above. If instead the (τ, c)-witness condition holds for some nonincreasing function τ as above, then for any $\lambda > 0$,

$$\mathbf{E}_{Z^n\sim P}\mathbb{E}_{\underline{f}\sim\Pi_n}\left[\mathbb{E}[L_f]\right] \;\le\; \lambda + c_{\tau(\lambda)} \cdot \mathbf{E}_{Z^n\sim P}\left[\mathrm{IC}_{n,\eta}\left(\Pi_0\right)\right].$$

The actual theorem given by GM generalizes this to an in-probability statement for general (not just generalized Bayesian) learning methods. If the (u, c)-witness condition holds, then, as is obvious from (6.18) and Theorem 6.1, the same rates can be obtained for the excess risk as for the squared misspecification metric. For the (τ, c)-witness condition things are a bit more complicated; the following lemma (Lemma 16 of GM) says that, under an exponential tail condition, (τ, c)-witness holds for a sufficiently 'nice' function τ, for which we loose at most a logarithmic factor:

Lemma 6. *Define $M_\kappa := \sup_{f\in\mathcal{F}} \mathbb{E}\left[e^{\kappa L_f}\right]$ and assume that the excess loss L_f has a uniformly exponential upper tail, i.e. $M_\kappa < \infty$. Then, for the map $\tau : x \mapsto 1 \vee \kappa^{-1}\log\frac{2M_\kappa}{\kappa x} = O(1 \vee \log(1/x))$, the (τ, c)-witness condition holds with $c = 1/2$.*

As an immediate consequence of this lemma, GM's theorem above gives that for any $\eta \in (0, \overline{\eta})$, (using $\lambda = 1/n$), there is $C_\eta < \infty$ such that

$$\mathbf{E}_{Z^n\sim P}\mathbb{E}_{\underline{f}\sim\Pi_n}\left[\mathbb{E}[L_{\underline{f}}]\right] \le \frac{1}{n} + C_\eta \cdot (\log n) \cdot \mathbf{E}_{Z^n\sim P}\left[\mathrm{IC}_{\eta,n}\left(f^* \parallel \Pi_|\right)\right], \tag{6.19}$$

so our excess risk bound is only a log factor worse than the bound that can be obtained for the squared misspecification metric in Theorem 6.1. We now apply this to the misspecified GLM setting:

Generalized Linear Models and Witness Recall that the central condition holds for generalized linear models under the three assumptions made in Proposition 3. Let $\ell_\beta := \ell_\beta(X, Y) = -\log p_\beta(Y \mid X)$ be the loss of action $\beta \in \mathcal{B}$ on random outcome $(X, Y) \sim P$, and let β^* denote the risk minimizer over $\mathcal{B}$. The first two assumptions taken together imply, via (6.7), that there is a $\kappa > 0$ such that

$$\begin{aligned}\sup_{\beta\in B}\mathbb{E}_{X,Y\sim P}\left[e^{\kappa(\ell_\beta-\ell_{\beta^*})}\right] &\le \sup_{\beta\in\mathcal{B},x\in\mathcal{X}}\mathbb{E}_{Y\sim P|X=x}\left[e^{\kappa(\ell_\beta-\ell_{\beta^*})}\right] \\ &= \sup_{\beta\in\mathcal{B},x\in\mathcal{X}}\left(\frac{F_{\theta_x(\beta)}}{F_{\theta_x(\beta^*)}}\right)^\kappa \cdot \mathbb{E}_{Y\sim P|X=x}\left[e^{\kappa|Y|}\right] < \infty.\end{aligned}$$

The conditions of Lemma 6 are thus satisfied, and so the (τ, c)-witness condition holds for the τ and c in that lemma. From (6.19) we now see that we get an $O((\log n)^2/n)$ bound on the expected excess risk, which is equal to the parametric (minimax) rate up to a $(\log n)^2$ factor. Thus, fast learning rates in terms of excess risks and KL divergence under misspecification with GLMs are possible under the conditions of Proposition 3.

6.C Learning rate > 1 for misspecified models

In what follows we give an example of a misspecified setting, where the best performance is achieved with the learning rate $\eta > 1$. Consider a model $\{P_\beta, \beta \in [0.2, 0.8]\}$, where P_β is a Bernoulli distribution with $\mathbb{P}_\beta(Y = 1) = \beta$. Let the data $Y_1, \dots, Y_n$ be sampled i.i.d. from P_0, i.e. $Y_i = 0$ for all $i = 1, \dots, n$. In this case the log-likelihood function is given by

$$\log p(Y_1, \dots, Y_n \,|\, \beta) = n \log(1 - \beta).$$

Observe that in this setting $\beta^\star = 0.2$. Now assume that the model is correct and data $Y_1', \dots, Y_n'$ is sampled i.i.d. from P_β with $\beta = 0.2$. Then the log-likelihood is

$$\log p(Y_1', \dots, Y_n' \,|\, \beta = 0.2) \approx 0.2n \log 0.2 + 0.8n \log 0.8 \ll n \log 0.8 = \log p(Y_1, \dots, Y_n \,|\, \beta = 0.2).$$

Thus, the data are more informative about the best distribution than they would be if the model were correct. Therefore, we can afford to learn 'faster': let the data be more important and the (regularizing) prior be less important. This is realized by taking $\eta >> 1$

6.D MCMC sampling

6.D.1 The η-generalized Bayesian lasso

Here, following Park and Casella (2008) we consider a slightly more general version of the regression problem:

$$Y = \mu + X\beta + \varepsilon,$$

where $\mu \in \mathbb{R}^n$ is the overall mean, $\beta \in \mathbb{R}^p$ is the vector of parameters of interest, $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\varepsilon \sim N(0, \sigma^2 I_n)$ is a noise vector. For a given shrinkage parameter $\lambda > 0$ the Bayesian lasso of Park and Casella (2008) can be represented as follows.

$$\begin{aligned} Y|\mu, X, \beta, \sigma^2 &\sim N(\mu + X\beta, \sigma^2 I_n)\,, \\ \beta|\tau_1^2, \dots, \tau_p^2, \sigma^2 &\sim N(0, \sigma^2 D_\tau), \quad D_\tau = \operatorname{diag}(\tau_1^2, \dots, \tau_p^2)\,, \\ \tau_1^2, \dots, \tau_p^2 &\sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2, \quad \tau_1^2, \dots, \tau_p^2 > 0\,, \\ \sigma^2 &\sim \pi(\sigma^2)\, d\sigma^2\,. \end{aligned} \tag{6.20}$$

In this model formulation the μ on which the outcome variables Y depend, is the overall mean, from which $X\beta$ are deviations. The parameter μ can be given a flat prior and subsequently integrated out, as we do in the coming sections.

We will use the typical inverse gamma prior distribution on σ^2, i.e. for $\sigma^2 > 0$

$$\pi(\sigma^2) = \frac{\gamma^\alpha}{\Gamma(\alpha)} \sigma^{-2\alpha-2} e^{-\gamma/\sigma^2},$$

where $\alpha, \gamma > 0$ are hyperparameters. With the hierarchy of (6.20) the joint density for the posterior with the likelihood to the power η becomes

$$\begin{aligned}
&(f(Y|\mu,\beta,\sigma^2))^\eta \, \pi(\sigma^2)\, \pi(\mu) \prod_{j=1}^{p} \pi(\beta_j|\tau_j^2,\sigma^2)\, \pi(\tau_j^2) \\
&\qquad = \left(\frac{1}{(2\pi\sigma^2)^{n/2}} \, e^{\frac{1}{2\sigma^2}(Y-\mu 1_n - X\beta)^T (Y-\mu 1_n - X\beta)} \right)^\eta \cdots \\
&\qquad\qquad \cdots \frac{\gamma^\alpha}{\Gamma(\alpha)} \sigma^{-2\alpha-2} e^{-\frac{\gamma}{\sigma^2}} \prod_{j=1}^{p} \frac{1}{(2\sigma^2\tau_j^2)^{1/2}} \, e^{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2} \frac{\lambda^2}{2} \, e^{-\lambda^2\tau_j^2/2}. \qquad (6.21)
\end{aligned}$$

Let $\widetilde{Y}$ be $Y - \overline{Y}$. If we integrate out μ, the joint density marginal over μ is proportional to

$$\sigma^{-\eta(n-1)} \, e^{-\frac{\eta}{2\sigma^2}(\widetilde{Y}-X\beta)^T(\widetilde{Y}-X\beta)} \, \sigma^{-2\alpha-2} \, e^{-\frac{\gamma}{\sigma^2}} \prod_{j=1}^{p} \frac{1}{(\sigma^2\tau_j^2)^{1/2}} \, e^{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2} \, e^{-\lambda^2\tau_j^2/2}. \qquad (6.22)$$

First, observe that the full conditional for β is multivariate normal: the exponent terms involving β in (6.22) are

$$\begin{aligned}
&-\frac{\eta}{2\sigma^2}(\widetilde{Y}-X\beta)^T(\widetilde{Y}-X\beta) - \frac{1}{2\sigma^2}\beta^T {D_\tau}^{-1}\beta \\
&\qquad = -\frac{1}{2\sigma^2}\left\{ (\beta^T(\eta X^T X + {D_\tau}^{-1})\beta - 2\eta\widetilde{Y}X\beta + \eta\widetilde{Y}^T\widetilde{Y}) \right\}. \qquad (6.23)
\end{aligned}$$

If we now write $M_\tau = (\eta X^T X + {D_\tau}^{-1})^{-1}$ and complete the square, we arrive at

$$-\frac{1}{2\sigma^2}\left\{ (\beta - \eta M_\tau X^T\widetilde{Y})^T M_\tau^{-1} (\beta - \eta M_\tau X^T\widetilde{Y}) + \widetilde{Y}^T(\eta I_n - \eta^2 X^{-1} M_\tau X^T)\widetilde{Y} \right\}.$$

Accordingly we can see that β is conditionally multivariate normal with mean $\eta M_\tau X^T\widetilde{Y}$ and variance $\sigma^2 M_\tau$.

The terms in (6.22) that involve σ^2 are:

$$(\sigma^2)^{\{-\eta(n-1)/2-p/2-\alpha-1\}} \exp\Big\{-\frac{\eta}{2\sigma^2}(\widetilde{Y}-X\beta)^T(\widetilde{Y}-X\beta)-\frac{1}{2\sigma^2}\beta^T {D_\tau}^{-1}\beta-\frac{\gamma}{\sigma^2}\Big\}.$$

We can conclude that σ^2 is conditionally inverse gamma with shape parameter $\eta\,\dfrac{n-1}{2}+\dfrac{p}{2}+\alpha$ and scale parameter $\dfrac{\eta}{2}(\widetilde{Y}-X\beta)^T(\widetilde{Y}-X\beta)+\beta^T {D_\tau}^{-1}\beta/2+\gamma$.

Since τ_j^2 is not involved in the likelihood, we need not modify the implementation of it and follow Park and Casella (2008):

$$\frac{1}{\tau_j^2}\sim \mathrm{IG}\Big(\sqrt{\lambda^2\sigma^2/\beta_j^2},\,\lambda^2\Big).$$

Summarizing, we can implement a Gibbs sampler with the following distributions:

$$\beta\sim \mathrm{N}\left(\eta(\eta X^TX+{D_\tau}^{-1})^{-1}X^T\widetilde{Y},\,\sigma^2(\eta X^TX+{D_\tau}^{-1})^{-1}\right), \tag{6.24}$$

$$\sigma^2\sim \text{Inv-Gamma}\Big(\frac{\eta}{2}(n-1)+p/2+\alpha,\,\frac{\eta}{2}(\widetilde{Y}-X\beta)^T(\widetilde{Y}-X\beta)+\beta^T {D_\tau}^{-1}\beta/2+\gamma\Big), \tag{6.25}$$

$$\frac{1}{\tau_j^2}\sim \mathrm{IG}\Big(\sqrt{\lambda^2\sigma^2/\beta_j^2},\,\lambda^2\Big). \tag{6.26}$$

There are several ways to deal with the shrinkage parameter λ. We follow the hierarchical Bayesian approach and place a hyperprior on the parameter. In our implementation we provide three ways to do so: a point mass (resulting in a fixed λ), a gamma prior on λ^2 following Park and Casella (2008) and a beta prior following De los Campos et al. (2009), details about the implementation of the latter two priors can be found in those papers respectively.

6.D.2 The η-generalized Bayesian logistic regression

We follow the construction of the Pólya–Gamma latent variable scheme for constructing a Bayesian estimator in the logistic regression context described in Polson, Scott and Windle, 2013.

First, for $b>0$ consider the density function of a Pólya-Gamma random variable $PG(b,0)$

$$p(x\,|\,b,0)=\frac{2^{b-1}}{\Gamma(b)}\sum_{n=1}^{\infty}(-1)^n\frac{\Gamma(n+b)}{\Gamma(n+1)}\frac{(2n+b)}{\sqrt{2\pi x^3}}e^{-\frac{(2n+b)^2}{8x}}.$$

The general class $PG(b,c)$ $(b,c>0)$ is defined through an exponential tilting of the $PG(b,0)$ and has the density function

$$p(x\,|\,b,c)=\frac{e^{-\frac{c^2x}{2}}p(x|b,0)}{\mathbb{E}\left[e\right]^{-\frac{c^2\omega}{2}}},$$

where $\omega\sim PG(b,0)$.

To derive our Gibbs sampler we use the following result from Polson, Scott and Windle, 2013.

Theorem 6.D.1. *Let $p_{b,0}(\omega)$ denote the density of $PG(b,0)$. Then for all $a \in \mathbb{R}$*

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2} p_{b,0}(\omega) d\omega,$$

where $\kappa = a - b/2$.

According to Theorem 6.D.1 the likelihood contribution of the observation i taken to the power η can be written as

$$L_{i,\eta}(\beta) = \left[\frac{(e^{X_i^T\beta})^{y_i}}{1+e^{X_i^T\beta}}\right]^\eta \propto e^{\eta\kappa_i X_i^T\beta} \int_0^\infty e^{-\omega_i \frac{(X_i^T\beta)^2}{2}} p(\omega_i \mid \eta, 0),$$

where $\kappa_i := y_i - 1/2$ and $p(\omega_i \mid \eta, 0)$ is the density function of $PG(\eta, 0)$.

Let

$$X := (X_1, \ldots, X_n)^T, \quad Y := (Y_1, \ldots, Y_n)^T, \quad \kappa := (\kappa_1, \ldots, \kappa_n)^T,$$
$$\omega := (\omega_1, \ldots, \omega_n)^T, \quad \Omega := \operatorname{diag}(\omega_1, \ldots, \omega_n).$$

Also, denote the density of the prior on β by $\pi(\beta)$. Then the conditional posterior of β given ω is

$$p(\beta \mid \omega, Y) \propto \pi(\beta) \prod_{i=1}^n L_{i,\eta}(\beta \mid \omega_i) = \pi(\beta) \prod_{i=1}^n e^{\eta\kappa_i X_i^T\beta - \omega_i \frac{(X_i^T\beta)^2}{2}} \propto \pi(\beta) e^{-\frac{1}{2}(z-X\beta)^T\Omega(z-X\beta)},$$

where $z := \eta(\frac{\kappa_1}{\omega_1}, \ldots, \frac{\kappa_n}{\omega_n})$. Observe that the likelihood part is conditionally Gaussian in β. Since the prior on β is Gaussian, a simple linear-model calculation leads to the following Gibbs sampler. To sample from the the η-generalized posterior one has to iterate these two steps

$$\omega_i \mid \beta \sim PG(\eta, X_i^T\beta), \tag{6.27}$$
$$\beta \mid Y, \omega \sim \mathcal{N}(m_\omega, V_\omega), \tag{6.28}$$

where

$$V_\omega := (X^T\Omega X + B^{-1})^{-1},$$
$$m_\omega := V_\omega(\eta X^T\kappa + B^{-1}b).$$

To sample from the Pólya-Gamma distribution $PG(b,c)$ we adopt a method from (Windle, Polson and Scott, 2014), which is based on the following representation result. According to Polson, Scott and Windle, 2013 a random variable $\omega \sim PG(b,c)$ admits the following representation

$$\omega \overset{\mathrm{d}}{=} \sum_{n=0}^\infty \frac{g_n}{d_n},$$

where $g_n \sim Ga(b,1)$ are independent Gamma distributed random variables, and

$$d_n := 2\pi^2(n+\frac{1}{2})^2 + 2c^2.$$

Therefore, we approximate the PG random variable by a truncated sum of weighted Gamma random variables. (Windle, Polson and Scott, 2014) shows that the approximation method performs well with the truncation level $N = 300$. Furthermore, we performed our own comparison of the sampler with the STAN implementation for Bayesian logistic regression, which showed no difference between the methods (for $\eta = 1$).

6.D.3 The Safe-Bayesian Algorithms

The version of the Safe-Bayesian algorithm we are using for the experiments is called *R-log-SafeBayes*, more details and other versions can be found in Grünwald and Van Ommen (2017). The $\widehat{\eta}$ is chosen from a grid of learning rates η that minimizes the *cumulative Posterior-Expected Posterior-Randomized log-loss*:

$$\sum_{i=1}^{n} \mathbb{E}_{\beta,\sigma^2\sim\Pi|z^{i-1},\eta}\left[-\log f(Y_i|X_i,\beta,\sigma^2)\right].$$

Minimizing this comes down to minimizing

$$\sum_{i=1}^{n-1} \text{AV}\left[\frac{1}{2}\log 2\pi\sigma_{i,\eta}^2 + \frac{1}{2}\frac{(Y_{i+1}-X_{i+1}\beta_{i,\eta})^2}{\sigma_{i,\eta}^2}\right].$$

The loss between the brackets is averaged over many draws of $(\beta_{i,\eta}, \sigma_{i,\eta}^2)$ from the posterior, where $\beta_{i,\eta}$ (or $\sigma_{i,\eta}^2$) denotes one random draw from the conditional η-generalized posterior based on data points z^i. For the sake of completeness we present the algorithm below.

Algorithm 1 The R-Safe-Bayesian algorithm

Input: data $z_1, \ldots z_n$, model $\mathcal{M} = \{f(\cdot|\theta)|\theta \in \Theta\}$, prior Π on Θ, step-size $\mathcal{K}_{\text{STEP}}$, max. exponent $\mathcal{K}_{\text{MAX}}$, loss function $\ell_\theta(z)$
$\mathcal{S}_n := \{1, 2^{-\mathcal{K}_{\text{STEP}}}, 2^{-2\mathcal{K}_{\text{STEP}}}, 2^{-3\mathcal{K}_{\text{STEP}}}, \ldots, 2^{-\mathcal{K}_{\text{MAX}}}, \}$
for all $\eta \in \mathcal{S}_n$ **do**
    $s_\eta := 0$
    for $i = 1 \ldots n$ **do**
        Determine generalized posterior $\Pi(\cdot|z^{i-1},\eta)$ of Bayes with learning rate η.
        Calculate posterior-expected posterior-randomized loss of predicting actual next outcome:

$$r := \ell_{\Pi|z^{i-1},\eta}(z_i) = \mathbb{E}_{\theta\sim\Pi|z^{i-1},\eta}\left[\ell_\theta(z_i)\right] \tag{6.29}$$

        $s_\eta := s_\eta + r$
    end for
end for
Ouput: Learning rate $\widehat{\eta}$

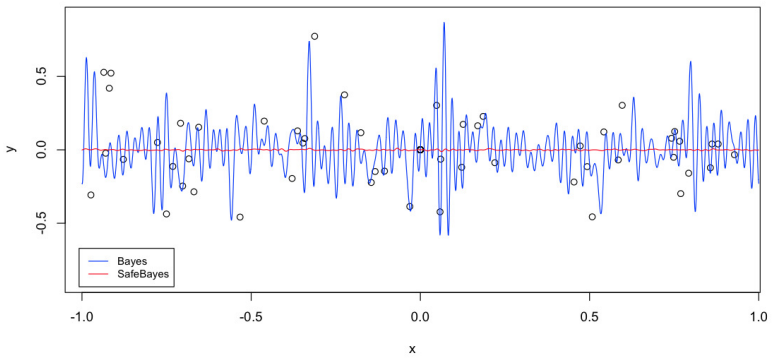

Figure 6.5: *Prediction of standard Bayesian lasso (blue) and Safe-Bayesian lasso (red, $\eta = 0.5$) with $n = 200$, $p = 100$.*

6.E Details for the experiments and figures

Below we present the results of additional simulation experiments for Section 6.5.1 (Appendix 6.E.1) and the description of experiments with real-world data (Appendix 6.E.2). We also give details for Figure 6.2 in Appendix 6.E.3.

6.E.1 Additional Figures for Section 6.5.1

Consider the regression context described in Section 6.5.1. Here, we explore different choices of the number of Fourier basis functions, showing that regardless of the choice Safe-Baysian lasso outperforms its standard counterpart. In Figures 6.5 and 6.6 we see conditional expectations $\mathbb{E}[Y \mid X]$ according to the posteriors of the standard Bayesian lasso (blue) and the Safe-Bayesian lasso (red, $\widehat{\eta} = 0.5$) for the *wrong-model* experiment described in Section 6.5.1, with 100 data points. We take 201 and 25 Fourier basis functions respectively.

Now we consider logistic regression setting and show that even for some well-specified problems it is beneficial to choose $\eta \neq 1$. In Figure 6.7 we see a comparison of the log-risk for $\eta = 1$ and $\eta = 3$ in the well-specified logistic regression case (described in Section 6.5.1). Here $p = 1$ and $\beta = 4$.

6.E.2 Real-world data

Seattle Weather Data The R-package `weatherData` (Narasimhan, 2014) loads weather data available online from `www.wunderground.com`. Besides data from many thousands of personal weather stations and government agencies, the website provides access to data from Automated Surface Observation Systems (ASOS) stations located at airports in the US, owned and maintained by the Federal Aviation Administration. Among them is a weather station at Seattle Tacoma International Airport, Washington (WMO ID 72793). From this station we collected the data for this experiment.

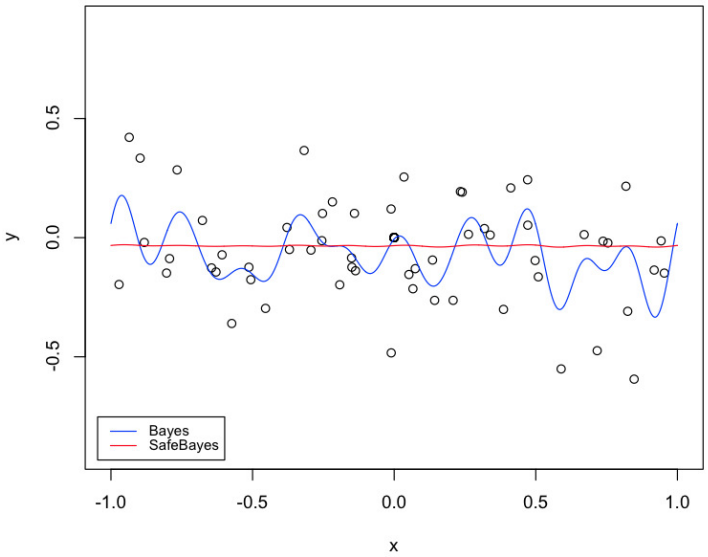

Figure 6.6: *Prediction of standard Bayesian lasso (blue) and Safe-Bayesian lasso (red, $\eta = 0.5$) with $n = 200$, $p = 12$.*

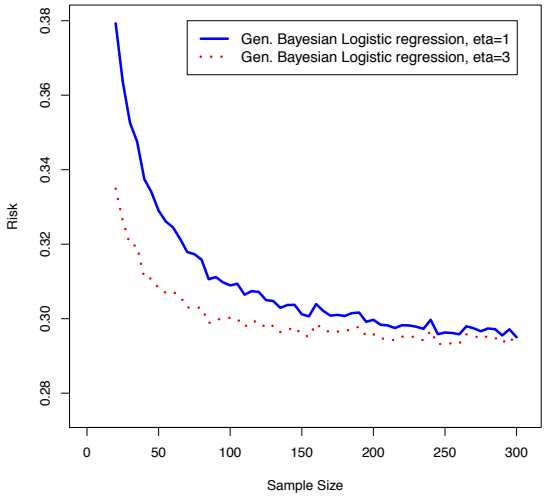

Figure 6.7: *Simulated logistic risk as a function of the sample size for the* correct-model *experiments described in Section 6.5.1 according to the posterior predictive distribution of standard Bayesian logistic regression ($\eta = 1$), and generalized Bayes ($\eta = 3$).*

The training data are the maximum temperatures for each day of the year 2011 at Seattle airport. We divided the data randomly in a training set (300 measurements) and a test set (65 measurements). First, we sampled the posterior of the standard Bayesian lasso with a 201-dimensional Fourier basis and standard improper priors on the training set, and we did the same for the Horseshoe. Next, we sampled the generalized posterior with the learning rate $\widehat{\eta}$ learned by the Safe-Bayesian algorithm, with the same model and priors on the same training set. The grid of η's we used was $1, 0.9, 0.8, 0.7, 0.6, 0.5$. We compare the performance of the standard Bayesian lasso and Horseshoe and the Safe-Bayesian versions of the lasso (SB) in terms of mean square error. In all experiments performed with different partitions, priors and number of iterations, SafeBayes never picked $\widehat{\eta} = 1$. We averaged over 10 runs. Moreover, whichever learning rate was chosen by SafeBayes, it always outperformed standard Bayes (with $\eta = 1$) in an unchanged set-up. Experiments with different priors for λ yielded similar results.

London Air Pollution Data As training set we use the following data. We start with the first four weeks of the year 2013, starting at Monday January 7 at midnight. We have a measurement for (almost) every hour until Sunday February 3rd, 23.00. We also have data for the first four weeks of 2014, starting at Monday January 6 at midnight, until Sunday February 2nd, 23.00. For each hour in the four weeks we randomly pick a data point from either 2013 or 2014. We remove the missing values. We predict for the same time of year in 2015: starting at Monday January 5 at midnight, until Sunday February 1st at 23.00. We do this with a (Safe-)Bayesian lasso and Horseshoe with a 201-dimensional Fourier basis and standard improper priors. The grid of η's we used for the Safe-Bayesian algorithm was again $1, 0.9, 0.8, 0.7, 0.6, 0.5$. We look at the mean square prediction errors, and average the errors over 20 runs of the generalized Bayesian lasso with the η learned by SafeBayes, and the standard Bayesian lasso and Horseshoe. Again we find that SafeBayes clearly performs better than standard Bayes.

6.E.3 Details for Figure 6.2

Here we sampled the posteriors of the standard and generalized Bayesian lasso ($\eta = 0.25$) on 50 model-wrong data points (approximately half easy points) with 101 Fourier basis functions, and estimated the predictive variance on a grid of new data points $X_{\text{new}} = \{-1.00, -0.99, \ldots, 1.00\}$ with the Monte Carlo estimate:

$$\widehat{\text{VAR}}\big(Y_{\text{new}} \mid X_{\text{new}}, Z_{\text{old}}\big) = \mathbb{E}_{\theta|Z_{\text{old}}}\big[\text{VAR}\big(Y_{\text{new}} \mid \theta\big)\big] + \widehat{\text{VAR}}\big[\mathbb{E}\big(Y_{\text{new}} \mid \theta\big)\big], \tag{6.30}$$

where

$$\mathbb{E}_{\theta|Z_{\text{old}}}\big[\text{VAR}\big(Y_{\text{new}} \mid \theta\big)\big] = \frac{1}{m}\sum_{k=1}^{m} \sigma^{2[k]} = \overline{\sigma^2},$$

$$\widehat{\text{VAR}}\big[\mathbb{E}\big(Y_{\text{new}} \mid \theta\big)\big] = \widehat{\text{VAR}}\big[X_{\text{new}}\beta\big] = \frac{1}{m}\sum_{k=1}^{m}\Big(X_{\text{new}}\beta^{[k]}\Big)^2 - \Big(X_{\text{new}}\overline{\beta}\Big)^2.$$

Here $\overline{\beta}$ is the posterior mean of the parameter for the coefficients and $\overline{\sigma^2}$ is the posterior mean of the variance.