A OUTLINE

The appendix of this paper is organized as follows:

- Appendix B provides the proofs for Section 3.
- Appendix C shows how we can replace $d^2_{\bar{\eta}}$ in Theorem 1 by stronger notions.
- Appendix D provides an example in which taking a learning rate larger than results in faster learning under misspecification than $\eta = 1$.
- Appendix E provides (implementation) details on the η-generalized Bayesian lasso and logistic regression; and the Safe-Bayesian algorithm.
- Appendix F contains details for the experiments and figures in the main text, and provides additional figures.

B PROOFS

B.1 Proof of Theorem 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 $\bar{\Theta}$, in particular for $\Theta = [\underline{\theta}, \bar{\theta}]$ with $\underline{\theta} < \bar{\theta}$ both in the interior of $\overline{\Theta}$, we have:

$$\begin{aligned} -\infty < \inf_{\theta \in \Theta} F(\theta) &< \sup_{\theta \in \Theta} F(\theta) < \infty \\ -\infty < \inf_{\theta \in \Theta} F'(\theta) &< \sup_{\theta \in \Theta} F'(\theta) < \infty \\ 0 < \inf_{\theta \in \Theta} F''(\theta) &< \sup_{\theta \in \Theta} F''(\theta) < \infty. \end{aligned} \tag{7}$$

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

$$\mathbf{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] = \mathbf{E}_{Y \sim P}\left[\left(\frac{p_\theta(Y)}{p_{\theta^*}(Y)}\right)^\eta\right] = \exp\left(-G(\eta(\theta - \theta^*)) + \eta F(\theta^*) - \eta F(\theta)\right). \tag{8}$$

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

$$\limsup_{\epsilon \to 0} \left\{\eta : \text{for all } \theta \in [\theta^* - \epsilon, \theta^* + \epsilon],\ \mathbf{E}[\exp(\eta L_{p_\theta})] \le 1\right\} = \frac{(\sigma^*)^2}{\sigma^2}. \tag{9}$$

Clearly, this statement also implies (iii). To prove (i), (ii) and (iii), it is thus sufficient to prove (ii) and (9). We prove both by a second-order Taylor expansion (around θ^*) of the right-hand side of (8).

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

$$\mathbf{E}[\exp(2\eta(\theta - \theta^*) \cdot Y)] \le \mathbf{E}[\exp(2\eta|\theta - \theta^*| \cdot |Y|)] \le \mathbf{E}[\exp(\eta^\circ(|\theta - \theta^*|/|\bar{\theta} - \underline{\theta}|) \cdot |Y|)] \le \bar{C} < \infty. \tag{10}$$

The first derivative of the right of (8) is:

$$\eta\, \mathbf{E}\left[(Y - F'(\theta)) \exp\Big(\eta\big((\theta - \theta^*)Y + F(\theta^*) - F(\theta)\big)\Big)\right]. \tag{11}$$

The second derivative is:

$$\mathbf{E}\left[\left(-\eta F''(\theta) + \eta^2 (Y - F'(\theta))^2\right) \cdot \exp\Big(\eta\big((\theta - \theta^*)Y + F(\theta^*) - F(\theta)\big)\Big)\right]. \tag{12}$$

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

$$\mathbf{E}[Y] = \mathbf{E}_{Y \sim P_{\theta^*}}[Y] = \mu(\theta^*); \qquad \text{for all } \theta \in \bar{\Theta}\text{: } F'(\theta) = \mu(\theta); \qquad F''(\theta) = \mathbf{E}_{Y \sim P_\theta}(Y - E(Y))^2, \tag{13}$$

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 $0 < \eta \leq \eta'$, all $\theta \in \Theta$, $\mathbf{E}[\exp(2\eta(\theta - \theta^*)Y)] < \infty$. Fix any such η. For some $\theta' \in \{(1-\alpha)\theta + \alpha\theta^* \colon \alpha \in [0,1]\}$, the (exact) expansion is:

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

Defining $\Delta = \theta' - \theta$, and since $F'(\theta^*) = \mathbf{E}[Y]$ (see (13)), we see that the central condition is equivalent to the inequality:

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

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

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

which is equivalent to

$$\eta \leq \frac{F''(\theta')\,\mathbf{E}\left[e^{\eta\Delta Y}\right]}{\sqrt{\mathbf{E}\left[(Y - F'(\theta'))^4\right]\mathbf{E}\left[e^{2\eta\Delta Y}\right]}}. \tag{14}$$

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 (7). Next, by Jensen's inequality,

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

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 $\bar{C}$ in (10). The first factor is bounded by a fixed multiple of $\mathbf{E}\,|Y|^4 + \mathbf{E}[F'(\theta)^4]$. The second term is bounded by (7), so it remains to bound the first term. By assumption $\mathbf{E}[\exp(\eta^\circ|Y|)] \leq \bar{C}$ and this implies that $\mathbf{E}\,|Y^4| \leq a^4 + \bar{C}$ for any $a \geq e$ such that $a^4 \leq \exp(\eta^\circ a)$; such an a clearly exists and only depends on η°.

We have thus shown that the RHS of (14) is upper bounded by a quantity that only depends on $\bar{C}, \eta^\circ$ and the values of the extrema in (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 (8) is finite for all θ' in between θ and θ^*; such a $\theta \neq \theta^*$ must exist since for any $\delta > 0$, if $|\theta - \theta^*| \leq \delta$, then by assumption (8) must be finite for all $\eta \leq \eta^\circ/\delta$. Evaluating the first and second derivative (11) and (12) at $\theta = \theta^*$ gives:

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

where $h(\theta)$ is a function satisfying $\lim_{\theta\to\theta^*} h(\theta) = 0$, where we again used (13), i.e. that $F'(\theta^*) = \mathbf{E}[Y]$. Using further that $\sigma^2 = \mathbf{E}\left[(Y - F'(\theta^*))^2\right]$ and $F''(\theta^*) = (\sigma^*)^2$, we find that $\mathbf{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] \leq 1$ 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 $\epsilon > 0$ such that for all $\theta \in [\theta^* - \epsilon, \theta^* + \epsilon]$, all $\eta > 0$,

$$\frac{\eta^2}{2}\sigma^2 \leq \frac{\eta}{2}(\sigma^*)^2 - \delta \Rightarrow \mathbf{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] \leq 1 \tag{15}$$

$$\frac{\eta^2}{2}\sigma^2 \geq \frac{\eta}{2}(\sigma^*)^2 + \delta \Rightarrow \mathbf{E}\left[e^{-\eta(\ell_\theta - \ell_{\theta^*})}\right] \geq 1 \tag{16}$$

The condition in (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 (15) is satisfied for η_δ. Thus, for all small enough δ, both the premise and the conclusion in (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 $\epsilon_{(j)} > 0$ such that for all $\theta \in [\theta^* - \epsilon_{(j)}, \theta^* + \epsilon_{(j)}]$, $\mathbf{E}\left[e^{-\eta_{(j)}(\ell_\theta - \ell_{\theta^*})}\right] \leq 1$. It follows that the lim sup in (9) is at least $(\sigma^*)^2/\sigma^2$. A similar argument (details omitted) using (16) shows that the lim sup is at most this value; the result follows.

B.2 Proof of Proposition 2

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

$$\mathbf{E}_{P'}[Y|X] = g^{-1}(\langle\beta, X), \tag{17}$$

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

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

where the latter equation follows by (17). The result now follows because (17) both holds for the 'true' P and for P_{f^*}.

B.3 Proof of Proposition 1

The fact that under the three imposed conditions the $\bar{\eta}$-central condition holds for some $\bar{\eta} > 0$ is a simple consequence of Theorem 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 $\bar{\eta}$ in Theorem 2 can be chosen uniformly for all $x \in \mathcal{X}$.

C EXCESS RISK AND KL DIVERGENCE INSTEAD OF GENERALIZED HELLINGER DISTANCE

The misspecification metric/generalized Hellinger distance $d_{\bar{\eta}}$ appearing in Theorem 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 $d_{\bar{\eta}}$ 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 et al., 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 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* $f \in \mathcal{F}$

$$\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u\}}\right] \geq c\,\mathbf{E}[\ell_f - \ell_{f^*}].$$

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* $f \in \mathcal{F}$, $\mathbf{E}[\ell_f - \ell_{f^*}] < \infty$ *and*

$$\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq \tau(\mathbf{E}[\ell_f - \ell_{f^*}])\}}\right] \geq c\,\mathbf{E}[\ell_f - \ell_{f^*}].$$

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 1 but with the misspecification metric replaced by the excess risk.

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

Theorem 3. [Specialization of Theorem 12 of GM] *Consider a learning problem $(P, \ell, \mathcal{F})$. Suppose that the $\bar{\eta}$-strong central condition holds. If the (u,c)-witness condition holds, then for any $\eta \in (0, \bar{\eta})$,*

$$\mathbf{E}_{Z^n \sim P}\,\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{E}[L_f]\right] \leq c_u \cdot \mathbf{E}_{Z^n \sim P}\left[\mathrm{IC}_{n,\eta}\left(\Pi_0\right)\right], \tag{18}$$

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

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

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 (18) and Theorem 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 1. *Define $M_\kappa \coloneqq \sup_{f \in \mathcal{F}} \mathbf{E}\left[e^{\kappa L_f}\right]$ 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 $\eta \in (0, \bar{\eta})$, (using $\lambda = 1/n$), there is $C_\eta < \infty$ such that

$$\mathbf{E}_{Z^n \sim P}\,\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{E}[L_{\underline{f}}]\right] \leq \frac{1}{n} + C_\eta \cdot (\log n) \cdot \mathbf{E}_{Z^n \sim P}\left[\mathrm{IC}_{\eta,n}\left(f^* \parallel \Pi_|\right)\right], \tag{19}$$

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 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 1. Let $\ell_\beta \coloneqq \ell_\beta(X,Y) = -\log p_\beta(Y \mid X)$ 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 (7), that there is a $\kappa > 0$ such that

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

The conditions of Lemma 1 are thus satisfied, and so the (τ,c)-witness condition holds for the τ and c in that lemma. From (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 1.

D 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 $Y_1, \ldots, Y_n$ be sampled i.i.d. from P_0, i.e. $Y_i = 0$ for all $i = 1, \ldots, n$. In this case the log-likelihood function is given by

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

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 $\eta >> 1$

E MCMC SAMPLING

E.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.

$$\begin{aligned} Y|\mu, X, \beta, \sigma^2 &\sim N(\mu + X\beta, \sigma^2 I_n)\,, \\ \beta|\tau_1^2, \dots, \tau_p^2, \sigma^2 &\sim N(0, \sigma^2 D_\tau), \quad D_\tau = \mathrm{diag}(\tau_1^2, \dots, \tau_p^2)\,, \\ \tau_1^2, \dots, \tau_p^2 &\sim \prod_{j=1}^p \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2, \quad \tau_1^2, \dots, \tau_p^2 > 0\,, \\ \sigma^2 &\sim \pi(\sigma^2)\, d\sigma^2\,. \end{aligned} \tag{20}$$

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 (20) the joint density for the posterior with the likelihood to the power η becomes

$$\begin{aligned} &(f(Y|\mu, \beta, \sigma^2))^\eta\, \pi(\sigma^2)\, \pi(\mu) \prod_{j=1}^p \pi(\beta_j|\tau_j^2, \sigma^2)\, \pi(\tau_j^2) = \\ &= \left(\frac{1}{(2\pi\sigma^2)^{n/2}}\, e^{\frac{1}{2\sigma^2}(Y - \mu 1_n - X\beta)^T (Y - \mu 1_n - X\beta)} \right)^\eta \frac{\gamma^\alpha}{\Gamma(\alpha)} \sigma^{-2\alpha-2} e^{-\frac{\gamma}{\sigma^2}} \prod_{j=1}^p \frac{1}{(2\sigma^2\tau_j^2)^{1/2}}\, e^{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2}\, \frac{\lambda^2}{2}\, e^{-\lambda^2\tau_j^2/2}\,. \end{aligned} \tag{21}$$

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

$$\sigma^{-\eta(n-1)}\, e^{-\frac{\eta}{2\sigma^2}(\tilde{Y} - X\beta)^T(\tilde{Y} - X\beta)}\, \sigma^{-2\alpha-2}\, e^{-\frac{\gamma}{\sigma^2}} \prod_{j=1}^p \frac{1}{(\sigma^2\tau_j^2)^{1/2}}\, e^{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2}\, e^{-\lambda^2\tau_j^2/2}. \tag{22}$$

First, observe that the full conditional for β is multivariate normal: the exponent terms involving β in (22) are

$$-\frac{\eta}{2\sigma^2}(\tilde{Y} - X\beta)^T(\tilde{Y} - X\beta) - \frac{1}{2\sigma^2}\beta^T D_\tau{}^{-1}\beta = -\frac{1}{2\sigma^2}\left\{(\beta^T(\eta X^T X + D_\tau{}^{-1})\beta - 2\eta\tilde{Y} X\beta + \eta\tilde{Y}^T\tilde{Y})\right\}.$$

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

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

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

The terms in (22) that involve σ^2 are:

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

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 $\dfrac{\eta}{2}(\tilde{Y} - X\beta)^T(\tilde{Y} - X\beta) + \beta^T D_\tau{}^{-1}\beta/2 + \gamma$.

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

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

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

$$\beta \sim \mathrm{N}\left(\eta(\eta X^T X + D_\tau{}^{-1})^{-1} X^T \tilde{Y},\, \sigma^2(\eta X^T X + D_\tau{}^{-1})^{-1}\right), \tag{23}$$

$$\sigma^2 \sim \text{Inv-Gamma}\Big(\frac{\eta}{2}(n-1) + p/2 + \alpha,\, \frac{\eta}{2}(\tilde{Y} - X\beta)^T(\tilde{Y} - X\beta) + \beta^T D_\tau{}^{-1}\beta/2 + \gamma\Big), \tag{24}$$

$$\frac{1}{\tau_j^2} \sim \mathrm{IG}\left(\sqrt{\lambda^2\sigma^2/\beta_j^2},\, \lambda^2\right). \tag{25}$$

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.

E.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 et al. (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

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

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

To derive our Gibbs sampler we use the following result from Polson et al. (2013).

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

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

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

Let

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

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

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

where $z \coloneqq \eta(\frac{\kappa_1}{\omega_1}, \ldots, \frac{\kappa_n}{\omega_n})$. 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

$$\omega_i \,|\, \beta \sim PG(\eta, X_i^T \beta), \tag{26}$$
$$\beta \,|\, Y, \omega \sim \mathcal{N}(m_\omega, V_\omega), \tag{27}$$

where

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

To sample from the Pólya-Gamma distribution $PG(b, c)$ we adopt a method from (Windle et al., 2014), which is based on the following representation result. According to Polson et al. (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

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

Therefore, we approximate the PG random variable by a truncated sum of weighted Gamma random variables. (Windle et al., 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 $\eta = 1$).

E.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 $\hat{\eta}$ is chosen from a grid of learning rates η that minimizes the *cumulative Posterior-Expected Posterior-Randomized log-loss*:

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

Minimizing this comes down to minimizing

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

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 $z_1, \dots z_n$, model $\mathcal{M} = \{f(\cdot|\theta)|\theta \in \Theta\}$, prior $\Pi$ on $\Theta$, step-size $\mathcal{K}_{\text{STEP}}$, max. exponent $\mathcal{K}_{\text{MAX}}$, loss
         function $\ell_\theta(z)$
Output: Learning rate $\hat{\eta}$
$\mathcal{S}_n := \{1, 2^{-\mathcal{K}_{\text{STEP}}}, 2^{-2\mathcal{K}_{\text{STEP}}}, 2^{-3\mathcal{K}_{\text{STEP}}}, \dots, 2^{-\mathcal{K}_{\text{MAX}}}, \}$ ;
for all $\eta \in \mathcal{S}_n$ do
    $s_\eta := 0$ ;
    for $i = 1 \dots n$ do
        Determine generalized posterior $\Pi(\cdot|z^{i-1}, \eta)$ of Bayes with learning rate $\eta$.
        Calculate posterior-expected posterior-randomized loss of predicting actual next outcome:
            $r := \ell_{\Pi|z^{i-1},\eta}(z_i) = \mathbf{E}_{\theta \sim \Pi|z^{i-1},\eta}[\ell_\theta(z_i)]$        (28)
        $s_\eta := s_\eta + r$ ;
    end
end
Choose $\hat{\eta} := \arg\min_{\eta \in \mathcal{S}_n}\{s_\eta\}$ (if min achieved for several $\eta \in \mathcal{S}_n$, pick largest) ;
```

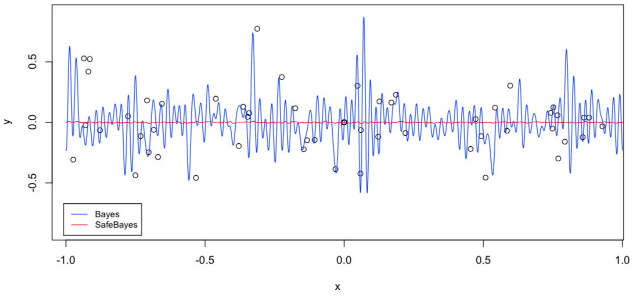

Figure 5: *Prediction of standard Bayesian lasso (blue) and Safe-Bayesian lasso (red, $\eta = 0.5$) with* $n = 200$, $p = 100$.

F DETAILS FOR THE EXPERIMENTS AND FIGURES

Below we present the results of additional simulation experiments for Section 5.1 (Appendix F.1) and the description of experiments with real-world data (Appendix F.2). We also give details for Figure 2 in Appendix F.3.

F.1 Additional Figures for Section 5.1

Consider the regression context described in Section 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 5 and 6 we see conditional expectations $\mathbf{E}[Y \mid X]$ according to the posteriors of the standard Bayesian lasso (blue) and the Safe-Bayesian lasso (red, $\hat{\eta} = 0.5$) for the *wrong-model* experiment described in Section 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 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 5.1). Here $p = 1$ and $\beta = 4$.

F.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

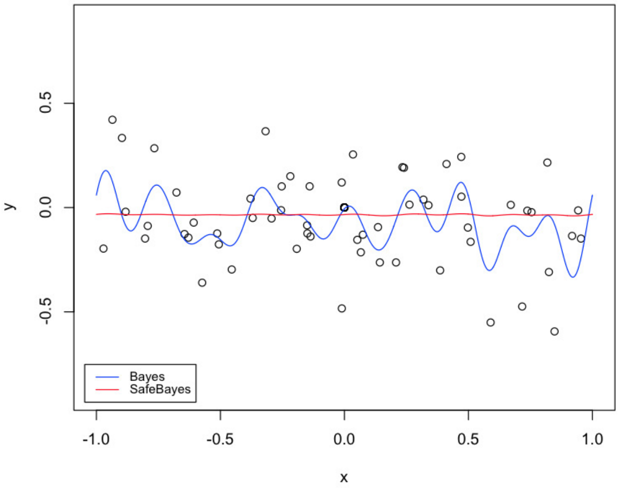

Figure 6: *Prediction of standard Bayesian lasso (blue) and Safe-Bayesian lasso (red, $\eta = 0.5$) with $n = 200$, $p = 12$.*

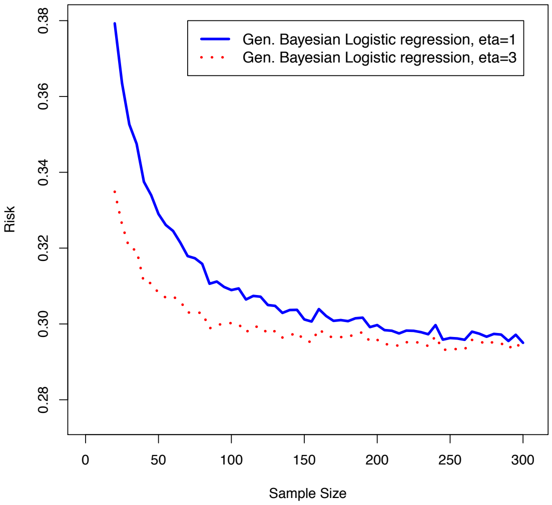

Figure 7: *Simulated logistic risk as a function of the sample size for the* correct-model *experiments described in Section 5.1 according to the posterior predictive distribution of standard Bayesian logistic regression ($\eta = 1$), and generalized Bayes ($\eta = 3$).*

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.

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 $\hat{\eta}$ 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 $\hat{\eta} = 1$. 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.

F.3 Details for Figure 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:

$$\widehat{\text{VAR}}(Y_{\text{new}} \mid X_{\text{new}}, Z_{\text{old}}) = \mathbf{E}_{\theta \mid Z_{\text{old}}} \left[\text{VAR}(Y_{\text{new}} \mid \theta) \right] + \widehat{\text{VAR}} \left[\mathbf{E}(Y_{\text{new}} \mid \theta) \right], \tag{29}$$

where

$$\begin{aligned} \mathbf{E}_{\theta \mid Z_{\text{old}}} \left[\text{VAR}(Y_{\text{new}} \mid \theta) \right] &= \frac{1}{m} \sum_{k=1}^{m} \sigma^{2[k]} = \overline{\sigma^2}, \\ \widehat{\text{VAR}} \left[\mathbf{E}(Y_{\text{new}} \mid \theta) \right] = \widehat{\text{VAR}} \left[X_{\text{new}} \beta \right] &= \frac{1}{m} \sum_{k=1}^{m} \left(X_{\text{new}} \beta^{[k]} \right)^2 - \left(X_{\text{new}} \overline{\beta} \right)^2. \end{aligned}$$

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