Interpreting Generalized Bayesian Inference By Generalized Bayesian Inference

Julian Rodemann JULIAN@STAT.UNI-MUENCHEN.DE
Thomas Augustin THOMAS.AUGUSTIN@STAT.UNI-MUENCHEN.DE
Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany
Rianne de Heide R.DE.HEIDE@VU.NL
Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands

Introduction The concept of safe Bayesian inference [4] with learning rates [5] has recently sparked a lot of research, e.g. in the context of generalized linear models [2]. It is occasionally also referred to as generalized Bayesian inference, e.g. in [2, page 1] – a fact that should let IP advocates sit up straight and take notice, as this term is commonly used to describe Bayesian updating of credal sets. On this poster, we demonstrate that this reminiscence extends beyond terminology.

Generalized Bayesian Inference (1) Several papers brought forward the idea of equipping Bayesian updating with an exponent on the likelihood, the *learning rate*, which we denote with η. [4] suggested its use as a method to deal with *bad misspecification* and proposed the Safe-Bayesian algorithm to learn the 'right' η from the data, which was instantiated and experimented with for (generalized) linear models in [2]. The η-generalized posterior is often [4, 2] defined for general learning problems and loss functions, and specialized to parametric models and log loss we obtain the following expression for it: $\pi(\theta \mid x, \eta) := \frac{\ell_x(\theta)^\eta \pi(\theta)}{\int_\Theta \ell_x(\theta)^\eta \pi(\theta) d\theta}$, with $\ell_x(\theta)$ the likelihood and $\pi(\theta)$ the prior. It has become clear that Bayesian inference can behave badly under misspecification of the model. It is well-known that when the model is well-specified, i.e. the true data generating process $P^\star$ is in the model $\mathcal{F}$, the posterior converges fast with high $P^\star$-probability to the true distribution in terms of Hellinger distance, under weak conditions on model and prior [5]. For misspecified models this only holds under very strong conditions. However, it was shown that under what is called the $\bar{\eta}$-central condition (which is much milder), generalized Bayes with a specific, finite, (often < 1) learning rate $\bar{\eta}$ also concentrates with high probability in the neighborhood of the distribution $\tilde{P} \in \mathcal{F}$ that is closest to $P^\star$ in KL-divergence, with fast rates.

Generalized Bayesian Inference (2) We observe $\ell_x(\theta)^\eta \pi(\theta) = \ell_x(\theta) \left[\pi(\theta)\, \ell_x(\theta)^{\eta-1}\right]$. This allows us to specify a set of priors given some base prior $\pi(\theta)$ as follows

$$\Pi_{\pi(\theta)} = \left\{ \pi_\nu(\theta) \mid \tilde{\pi}(\theta) = \pi(\theta) \cdot \ell_x(\theta)^{\eta-1}, \eta \in (0,1); \pi_\nu(\theta) = \tilde{\pi}(\theta)/C_\nu, C_\nu = \int_\Theta \tilde{\pi}(\theta) d\theta \right\}. \tag{1}$$

The (unnormalized versions of the) prior functions in this credal set can be charaterized by a point-wise upper and a (trivial) point-wise lower bound, namely $\pi(\theta) \cdot \ell_x(\theta)^{-1}$ and $\pi(\theta)$. Sets of priors with these characteristics satisfy the requirements for *Bayes theorem for unbounded priors* [3, page 238], so a posterior credal set exists, see [1, chapter 2.3].

This reformulation of safe Bayesian inference offers exciting insights into why safe Bayesian allows concentration of the posterior under model misspecification. Specifically, the unnormalized prior $\tilde{\pi}(\theta) = \pi(\theta) \cdot \ell_x(\theta)^{\eta^*-1}$ for optimal η^* can help. It provides us with a counterfactual Bayesian interpretation of safe Bayesian learning: If we have had specified the prior proportional to $\tilde{\pi}(\theta)$, we would have achieved concentration under model misspecification with regular Bayesian learning. In this way, it conveys information on which parts of Θ are relevant to the non-concentration under misspecification.

References

[1] F. P. A. Coolen. *Statistical modeling of expert opinions using imprecise probabilities.* PhD thesis, 1995.

[2] R. de Heide, A. Kirichenko, P. Grunwald, and N. Mehta. Safe-bayesian generalized linear regression. In *AISTATS*, volume 108 of *PMLR*, pages 2623–2633, 2020. URL https://proceedings.mlr.press/v108/heide20a.html.

[3] L. DeRoberts and J.A. Hartigan. Bayesian inference using intervals of measures. *The Annals of Statistics*, 9(2):235–244, 1981. doi:10.1214/aos/1176345391.

[4] P. Grünwald. The safe bayesian: learning the learning rate via the mixability gap. In *Algorithmic Learning Theory (ALT 2012)*, volume 7568 of *LNAI*, pages 169–183. Springer, 2012. doi:10.1007/978-3-642-34106-9_16.

[5] T. Zhang. From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006. doi:10.1214/009053606000000704.