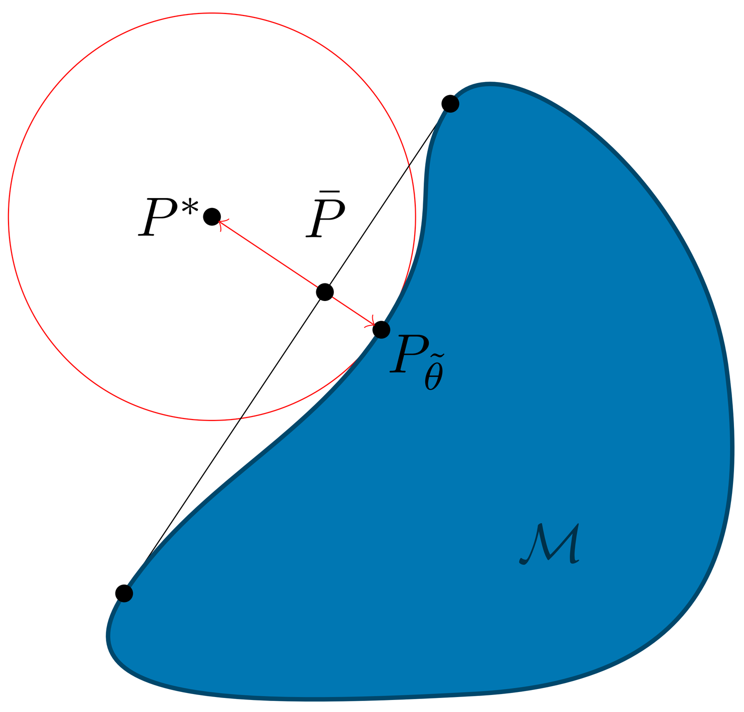
Interpreting Generalized Bayesian Inference by Generalized Bayesian Inference

Julian Rodemann, Thomas Augustin, Rianne de Heide
julian@stat.uni-muenchen.de

Generalized Bayes I – Learning Rate

Problem: *Bad* misspecification

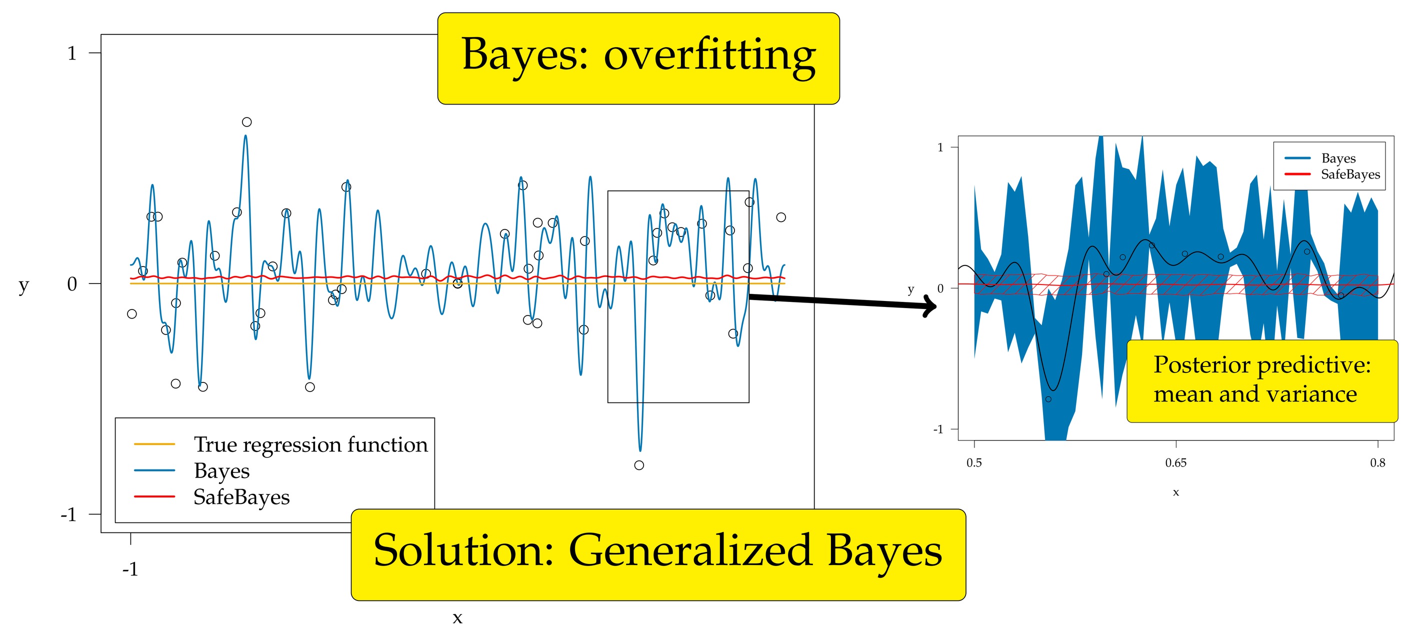
$P_{\tilde{\theta}}$ is the closest distribution in the model $\mathcal{M}$ to the true P^* in KL-divergence. When the model is not convex, the posterior predictive distribution $\overline{P}$ might be a mixture of *bad* distributions in the model that ends up outside $\mathcal{M}$. We get:



- Bad square-risk behaviour
- Good log-risk behaviour

This discrepancy implies that the posterior is not concentrated.

Extreme example: Model $y_i = \mathbf{f}(\mathbf{x_i}) + \epsilon_i, \epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \frac{1}{4})$, Fourier basis, and a simple model misspecification: $y_i = \mathbf{0} + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \frac{1}{4})$, $x_i \overset{iid}{\sim} \mathrm{U}(-1, 1)$, **but then** set half of the data to $(0, 0)$.



Solution: Generalized (Safe-) Bayes with learning rate η [5]

$$\pi(\theta \mid x, \eta) := \frac{\ell_\theta(x)^\eta \pi(\theta)}{\int_\Theta \ell_\theta(x)^\eta \pi(\theta) d\theta}, \tag{1}$$

with $\ell_\theta(x)$ the likelihood and $\pi(\theta)$ the prior in case of parametric models and log loss.

- learn optimal η^* with the **Safe-Bayesian algorithm** [3]
- posterior concentrates on $P_{\tilde{\theta}}$ with fast rates and mild condition ($\bar{\eta}$ -central condition) if η taken *small enough*, e.g. for GLMs [4]

Generalized Bayes II – Credal Sets

Credal Sets

In the IP literature, "generalized Bayes" typically refers to defining a set of priors

$$\Pi \subseteq \{\pi(\theta) \mid \pi(\cdot) \text{ a probability measure on } (\Theta, \sigma(\Theta))\} \tag{2}$$

with $\sigma(\cdot)$ an appropriate (σ-)algebra and Θ a (compact) parameter space. Inference then basically consists of updating Π to a set of poteriors.

Reformulation of Generalized Bayes I

Consider the numerator in the Bayes rule with learning rate η (equation 1). Note that $\ell_\theta(x)^\eta \pi(\theta) = \ell_\theta(x) \left[\pi(\theta)\, \ell_\theta(x)^{\eta-1}\right]$. This allows us to specify a set of priors given some base prior $\pi(\theta)$ as follows

$$\Pi_{\pi(\theta)} = \left\{\pi_\nu(\theta) \mid \tilde{\pi}(\theta) = \pi(\theta) \cdot \ell_\theta(x)^{\eta-1}, \eta \in (0, 1)\right\} \tag{3}$$

with normalization by $\pi_\nu(\theta) = \tilde{\pi}(\theta)/C_\nu, C_\nu = \int_\Theta \tilde{\pi}(\theta) d\theta$.

Bayes Theorem for Unbounded Priors [2]

Unnormalized versions of the prior functions in $\Pi_{\pi(\theta)}$ can be charaterized by a point-wise upper and a (trivial) point-wise lower bound, namely $\pi(\theta) \cdot \ell_\theta(x)^{-1}$ and $\pi(\theta)$

–> Sets of priors with these characteristics satisfy the requirements for *Bayes theorem for unbounded priors* [2, page 238]

–> So a posterior credal set exists, see [1, chapter 2.3]

Thus, we can identify the posteriors with learning rates η (equation 1) with a posterior credal set.

Generalized Bayes I Through the Lens of Generalized Bayes II

Why does the learning rate allow concentration of the posterior under model misspecification?

- Assume the Safe-Bayesian algorithm finds the optimal η^*
- Consider our unnormalized prior for η^*, i.e. $\tilde{\pi}(\theta)^* = \pi(\theta) \cdot \ell_\theta(x)^{\eta^*-1}$
- $\tilde{\pi}(\theta)^*$ gives a **counterfactual Bayesian explanation** of Safe-Bayesian learning: *If we 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.

Can this interpretation improve the Safe-Bayesian algorithm?
The Safe-Bayesian algorithm iterates both over a grid of η's, and over each datapoint, i.e. the posterior has to be computed for each combination of η and each datapoint anew. Can we use our representation of the learning rates by a credal set somehow to speed up the search for the optimal η^*?

References

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

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

[3] P. Grünwald. The safe Bayesian: learning the learning rate via the mixability gap. In *Algorithmic Learning Theory: 23rd International Conference (ALT), Proceedings 23*, pages 169–183. Springer, 2012.

[4] R. de Heide, A. Kirichenko, N. Mehta, and P. Grünwald. Safe-Bayesian generalized linear regression. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2623–2633. PMLR, 2020.

[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.