On Theoretical Motivations in an Optimization-centric View on Bayes' Rule [1]

Henri Duprieu
henri.duprieu@polytechnique.edu

Thomas Robert
thomas.robert.x21@polytechnique.edu

1 From Bayes's Rule to the Generalized Variational Inference (GVI)

From Bayes' Rule to an optimization problem. Bayes' rule is classically conceptualized as a multiplicative update rule. A prior distribution π on parameter θ is updated by multiplication of the likelihood of the model for the available data $x_{1:n}$. Denoting $q_B(\theta)$ the resulting posterior distribution, and considering only probability distributions which admit a density on the parameter space, this update step can trivially be seen as finding the probability function which minimizes the Kulback-Leibler divergence with $q_B(\theta)$. Substituting $q_B(\theta)$ with its definition, this allows to see Bayes' rule as infinite-dimensional optimization problem. Notably, the objective functional exhibits separate terms for the attachment to the data, and regularization through prior.

$$q_B(\theta) = \arg\min_{q\in\mathcal{P}(\Theta)} \mathrm{KLD}\left(q||q_B\right) = \arg\min_{q\in\mathcal{P}(\Theta)} \left\{ \mathbb{E}_{q(\theta)}\left[-\sum_{i=1}^{n}\log\left(\theta, x_i\right)\right] + \mathrm{KLD}\left(q\|\pi\right)\right\}. \quad (1)$$

From generalized Bayes to the Rule of Three (RoT). From this optimization-centric point of view, some previous generalizations of Bayes' rule including Gibbs Bayes [2] and PAC-Bayes [3], which the authors generically refer to as *Generalized Bayes* [4], can be seen as the optimum of similar problem, replacing the -log function with different loss functions ℓ. This corresponds to adding a degree of freedom in the data-attachment term. Notably, reweighting the KLD term as $\frac{1}{w}$KLD in the objective also fits into generalized Bayes since it amounts to multiplying the loss term by the scaling factor w. The authors' proposition is to generalize this approach by also adding a degree of freedom in the prior attachment term, replacing KLD with a more general statistical divergence D. Considering the set of admissible posteriors Π as another degree of freedom, these added degrees of freedom yield a new - *post-Bayesian* - approach to calculating posteriors : the *Rule of Three* (RoT).

$$q^*(\theta) = P(\ell, D, \Pi) = \arg\min_{q\in\Pi}\left\{\mathbb{E}_{q(\theta)}\left[\sum_{i=1}^{n}\ell(\theta, x_i)\right] + D(q\|\pi)\right\}. \quad (2)$$

Variational inference for the Rule of Three : Generalized Variational inference. Variational inference, a tractable counterpart of Bayes' rule, is a particular case of the Rule of Three where Π is a variational family $\{q(\theta,\kappa), \kappa\in K\}$, $D = \mathrm{KLD}$ and $\ell = -\log$. It is extended as *Generalized Variational Inference* (GVI), the tractable counterpart of the Rule of Three (RoT) for which Π is a variational family. A previous line of work suggesting a generalization of variational inference, coined *Divegence Variational Inference* (DVI), consisted in directly substituting a more general divergence D for the KLD in the middle term of 1. DVI cannot be seen as a particular case of GVI, which only substitutes D for the KLD in the prior regularization term in the right term of 1.

Outline of this report and contributions. Our ambition in this report is to focus on the theoretical foundations of the Generalized Variational Inference (GVI). We first outline the conceptual, theoretical and practical aspects of the introduced method and briefly discuss its limitations in Section 2. We then present our main original contribution in Section 3, reviewing the modularity theorem, highlighting the need to weaken the authors' interpretation and showing the weaknesses of the experimental support provided by the authors. Additional contributions are presented in Section 4.

2 Critical Reading

2.1 Addressing Challenges for Bayesian Machine Learning

The *traditional Bayesian paradigm*, derived from fundamental rules of probability, relies on the practitioner's ability to provide a meaningful modeling of the learning problem at hand. In fact, the Bayesian posterior $q_B(\theta)$ results from conditional probability update rule assuming: **(P)** that the prior distribution $\pi(\theta)$ encapsulates information from domain expertise available prior to observing the data; and **(L)** that the data generating mechanism falls into the specified model (or at least that again it best describe knowledge available prior to observing the data). Recent advances in statistical learning from large-scale inference applications have led to a paradigm shift as machine learning relies on highly over-parameterized functions to fit the typical behavior of the data. As a result, in the contemporary *Bayesian machine learning paradigm*, the model is an uninterpretable black box, which prohibits the use of domain expertise to initialize the model in a meaningful way, and thus ultimately invalidates both assumptions **(P)** and **(L)**. Furthermore, scaling the model's dimension increases the computational burden and thus induce specific modeling choices not in accordance with task specific knowledge, further deceiving assumptions **(P)** and **(L)**.

Adopting an *optimization-centric view* allows for a relaxation that shifts design requirements from strong modeling assumptions to flexible optimization objectives. The latter program gives root to the Rule of Three (RoT) as derived from the Representation Axiom, that is as defined by an optimization problem in two arguments: (i) a data-attachment term given by a loss function ℓ; and (ii) a prior regularization term given by a statistical divergence D. The RoT theory is a refined version of (i)-(ii) that comes with theoretical foundations that advocate going beyond reweighting the Kullback-Leibler divergence to address prior misspecification by considering a wider range of divergence as a regularizer.

2.2 Theoretical Guarantees - Strong and Weak Points

The motivation and exposition of the RoT and it's tractable counterpart, GVI, are complemented with theoretical results. The first two such results correspond to that would be expected for a generalization of Generalized Bayes : (1) fequentist consistency (Theorem 12 [1]); and (2) link between lower-bounds of the limited generalized and fully generalized versions (Theorem 14 [1]). Discussing (1), under mild assumptions, and in particular assuming that a single optimal parameter θ^* does exist, Theorem 12 shows that the GVI posterior converges to the dirac measure on this optimal parameter, regardless of the chosen divergence (D in the GVI). Previous work from the authors introduce and prove this result in it's most general case. This is a strong point of the framework that was introduced in previous work [5]. Regarding (2), if the RoT does actually generalize *Generalized Bayes*, then its corresponding tractable version, the GVI should generalize *generalized Bayes Variational Inference*, including the well known frameworks of *Gibbs Variational Inference* and *PAC-Bayes*. In particular, for the same loss and parameter spaces, a good choice of divergence should produce a lower bound on the Gibbs evidence lower bound. Theorem 12 [1] provides this desired property for a wide range of divergences.

Considering the central role played by the divergence parameter in the new framework of the RoT, the authors conduct a thorough study of different classes of divergence functions. First based on toy problems in annex A and section 5.2 [1], structural observations are complemented with theoretical reasoning in section 5.2.2 [1]. While results described above confirm the validity of the new framework as a generalization of generalized Bayes, the author's main theoretical claim concerns the modularity of GVI. Theorem 10 [1] proves that each degree of freedom of the RoT can be used to independently address shortcomings of Bayesian posteriors in machine learning. For example, changing the prior π should allow to care for prior misspecification (**P**) independently from mode misspecification (**L**). The proof provided by the authors is not as general as their claim for modularity. As a consequence, we argue in section 3 that modularity of the RoT is only proved in a weaker sense than that claimed by the authors.

The authors do not focus on theoretical optimization guarantees of their framework, wether it be on the theoretical part, or for the algorithm used in practice. Only brief justification of existence and uniqueness of a minimizer is given in section 4.2 [1]. It is not clear which line of thought, and more precisely which assumptions are required for the well posedness of the RoT. Clues for the reasoning are present in previous work, although the most complete formulation of a proof of existence of a

minimizer can be found in appendix A of [6]. Additionally, we provide comments on convergence guarantees for the algorithm used in practice (Black Box GVI) in paragraph 4.1. In a similar fashion, the authors only provide heuristic to justify the following claim: working assumptions of the RoT imply that coherence is not a desirable property of the optimal solution. We question the general reasoning behind this heuristic, and independently identify a weak point in the heuristic in paragraph 4.2.

2.3 Practical Implication - Strong and Weak Points

From a practical point of view, the RoT $P(\ell, D, \Pi)$ extends naturally to a handy Bayesian inference strategy by restricting the set Π to a variational family Q. The latter comes with a tractable algorithmic counterpart, referred to as *black-box GVI* (BBGVI), which allows one to leverage sampling techniques to perform GVI on specific tasks. Comparison of standard divergences for prior regularization provides critical practical insight. The authors exhibit on toy examples how one can leverage: (i) Scoring rule derived from the β-divergence to address model misspecification in the specific case of model contamination by outliers; and (ii) Rényi's α divergence to address prior misspecification in the specific case of label switching applications, mitigating the zero-forcing behavior of the Kullback-Leibler divergence used for standard variational inference (VI). Further, on a larger scale Bayesian Neural Network application [7], the authors recommend using Rényi's α divergence with $\alpha > 1$ to neglect the increasing effect of misspecified prior regularization on the posterior variance. Beyond these examples, designing the right GVI formulation for a given task remains a challenge for practitioners.

3 Discussing the Modularity Theorem

Far beyond proposing a third degree of freedom in the prior regularization term, the authors claim to introduce a novel strategy to separately address both prior and model misspecification in a flexible and decoupled manner. This claim takes the form of the *Modularity theorem* and aims to justify a conceptual shift to overcome the limitations of the Bayesian paradigm for modern machine learning applications, ultimately leading to a so-called *post-Bayesian* era. In fact, modularity supports the groundbreaking idea that careful attention to the regularization term, *beyond reweighting with respect to the data-attachment term*, is necessary to address prior misspecification. The Modularity theorem formally stated and proved for the Rule of Three (RoT) is further extended with a strong interpretation for Generalized Variational Inference (GVI) and resolves, in the author's words, in: (1) GVI can address model misspecification by changing ℓ; (2) GVI can address prior misspecification by changing D; and (3) GVI can address undesirable uncertainty quantification by changing D. As a contribution, we will first discuss the theoretical foundations of the Modularity theorem, and then show the limitations of the empirical support provided by the author.

3.1 On the Theoretical Foundations of the Modularity Theorem

The authors' frequentist definition of Robustness. By construction, the author characterizes addressing model misspecification as satisfying a *Robustness condition* ultimately defined by the parameter $\theta^* = \arg\min_\theta\{\mathbb{E}_X[l(\theta, X)]\}$, also refered to as the *population-optimal* parameter. This chosen definition is not anodyne, it chooses to focus on the infinite data regime, and it values frequentist consistency above all else. The limitations of this definition for Bayesian machine learning are threefold. First, focusing on the infinite data regime annihilates the motivation behind Bayesian learning, as it views the posterior distribution as the Dirac mass δ_{θ^*}, thus invalidating the interest of posterior sampling both for improving inference upon regularization and for allowing localized uncertainty quantification. Second, focusing on a population-optimal parameter leads to either the objectivist or the subjectivist interpretation of the assumption **(L)** that the authors claim to overcome in formulating the optimization-centric view. Third, recent advances in machine learning for statistical learning suggest a weak interpretation of fitting the data in terms of typical behavior, thus placing regularization at the core of the inference strategy.

Point (1) of the Modularity theorem is a trivial direct consequence of the Robustness definition, and is thus subject to its limitations. Robustness characterization also affects points (2) and (3) of the Modularity theorem, since, contrary to the authors' reformulation and interpretation of the theorem,

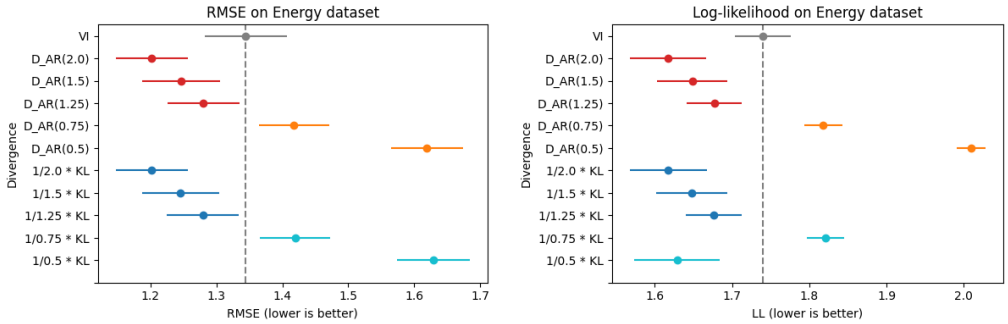

Figure 1: Results of Bayesian Neural Network Regression.

preserving the parameter of interest θ^* is set as a requirement before addressing prior misspecification and poor uncertainty quantification.

A reformulation of the Modularity theorem. In view of the Representation theorem, enforcing the authors' notion of robustness is equivalent to restricting the freedom on adjusting ℓ to affine transformation of the form $l' = w \cdot l + C$. The Representation theorem itself is a consequence of the *axiomatic design choice* to enforce recovering Bayesian posterior for $D = \text{KLD}$ and $\Pi = \mathcal{P}(\theta)$. The latter amounts to enforcing a balanced additive relation of the data-attachment and the prior regularization term, which leads to a crucial simplification for the development of the theory, but at the expense of the inference problems that admits an RoT formulation. Given these preliminary assumptions, the core of the proof lies in the observation that $P(w \cdot l + C, D, \Pi) = P(\ell, \frac{1}{w} D, \Pi)$; and the theorem reduces to: given the robustness definition and design choice, any change in the loss ℓ that doesn't change the parameter θ^* can be rewritten as an adjustment to the divergence D.

The above discussion undermines the theoretical foundations of the specific modular representation of the Rule of Three beyond reconceptualizing the Bayesian machine learning inference task as solving an optimization problem that balances data-attachment and regularization.

3.2 On the Empirical Support of the Modularity theorem

The practical implication of the RoT Modularity lies in defining specific strategies for dealing with either model misspecification or prior misspecification. In particular, the modularity theorem suggests adjusting the divergence D when one expects prior misspecification. In the specific case of Bayesian Neural Network regression, the design of an expert-informed prior is out of sight, and the fully factorized prior (namely the isotropic normal distribution) remains the de facto choice for initializing the network parameters' distribution. In this situation, the optimization-centric thesis recommends a weaker prior regularization in the form of Rényi's α divergence with $\alpha > 1$. The reproducible experiment provided by the authors provides empirical support: for $\alpha \in (1, 2)$, $P(\ell, D^{\alpha}_{AR}, \mathcal{Q})$ consistently outperforms $P(\ell, \text{KLD}, \mathcal{Q})$, and further regularization with $\alpha > 2$ leads to overconfidence in the fitted model and ultimately worsens inference.

According to the Modularity theorem, and contrary to previous work in the field, Bayesian Neural Network prior misspecification must be addressed by adjusting the form of the regularizing divergence D. However, a weaker interpretation of the supporting experiments using Rényi's α divergence with $\alpha > 1$ also suggests that shrinking the posterior variance is the step toward improved inference. To compare the authors' strong claim of Modularity with the latter previously known interpretation, we reproduce the authors' experiments (from their implementation) and compare the results with adjusting the balance between log-likelihood penalization and Kullback-Leibler regularization by a factor of $\frac{1}{w}$ – we compare $P(\ell, D^{\alpha}_{AR}, \mathcal{Q})$ and $P(\ell, \frac{1}{w}\text{KLD}, \mathcal{Q}) = P(w \cdot l, \text{KLD}, \mathcal{Q})$. The results are given in Figure 1 (code is made available at this https URL), and support our interpretation: in this particular case, reweighting the loss effectively addresses the prior misspecification. That is, the Modularity theorem's frequentist view of Robustness, which de facto enforces a new form for the regularizer, may not be the key to designing new variational formulations to overcome challenges in Bayesian machine learning.

4 Additional Contributions : Detailing Theoretical Considerations

4.1 An Optimisation-centric View on the Optimisation-centric View on Bayes' Rule

Convergence of Black Box GVI, the practical algorithm for posterior estimation, is not studied theoretically by the authors. BBGVI is introduced with a practical lense by referencing to *Black Box Variational Inference* [8]. At a high level, BBVI reduces to stochastic gradient ascent on the Evidence Lower Bound (ELBO), which is assumed to be regular enough in practice. BBGVI extends the stochastic gradient algorithm approach from the ELBO to the more general GVI objective function, with a similar practical perspective. Notably, they generalize certain practical variance reduction techniques underlying BBVI (II) to the GVI case in Appendix H.3 [8], under additional assumptions. Authors of the BBVI article restrain their theoretical work to deriving an unbiased estimate of the ELBO and refer to the classical stochastic optimization framework, notably [9] for the practical choice of learning rates, and to [10] for convergence guarantees. However, these results rely on strong assumptions on the general objective function (generalized smoothness and convexity, see section 4., in particular assumption 4.16 [10]). More precisely, we wonder under which assumptions on the loss ℓ, and the variational family Q (resp. ℓ, Q and D), the SGD underlying BBVI (resp. BBGVI) converges ? What speed guarantees can be achieved ? Answering these questions for BBVI is still an active area of research. For example, more subtle results relying on milder assumptions on ℓ and the variational family can be found in [11] and [12]. As a consequence, no consensual theoretical framework can be applied to BBGVI.

Notably, the authors of [1] obtain unbiased estimates of the gradient of the objective by using the *score-trick*, also called *REINFORCE* ($\nabla_\kappa \mathbb{E}_{\theta\sim p(-|\kappa)}[f(\theta)] = \mathbb{E}_{\theta\sim p(-|\kappa)}[f(\theta)\nabla_\kappa \log p(\theta|\kappa)]$). This allows the framework to remain general (or "blackbox"), with no assumption on the variational family. However, existing theoretical results for BBVI rely on unbiased estimates derived through the *reparameterization trick* ($\nabla_\kappa \mathbb{E}_{\theta\sim p(\theta|\kappa)}[f(\theta)] = \mathbb{E}_{\varepsilon\sim q(\varepsilon)}[\nabla_\kappa f(g_\kappa(\varepsilon))]$) for a certain class of variational families [11] and [12]. As a consequence, BBGVI should be considered in a more general sense than presented in the paper to apply existing results, all the while restricting the variational family. Moreover, as existing results already consider generalized loss ℓ, work to extend them to BBGVI would mostly revolve around the substitution of a more general statistical divergence D for KLD.

4.2 On the Proof of the Irrelevance of Coherence

The authors argue that supposing **(P)** and **(C)**, namely a good specification of the prior as well as infinite compute power, implies that *coherence* is a desirable quality of the posterior. A proof sketch for this assumption is provided in section 4.4.3 [1]. This is a possible explanation of the coherence of Bayes' (or generalized Bayes') rule. Using this implication, the authors argue that, since in the RoT framework **(P)** and **(C)** are not assumed to be deceived, coherence is not desirable anymore. This heuristic reasoning is not viable as, contrary to what the authors finally state, they only have proven that enforcing coherence is reasonable *if* **(P)** and **(C)** are assumed to be true, not *only if*. In fact, there could be other reasons to desire coherence without assuming **(P)** and **(C)**. Moreover, coming back on the proof sketch, the argument along which supposing **(P)** implies setting $D = \text{KLD}$ is not detailed, and we have not been able to reproduce it. The remaining of the proof is explicit.

Conclusion

In the above report, we discussed the theoretical foundations that lead to the practical derivation of the Rule of Three in the form of Generalized Variational Inference; and put its application into perspective to address concrete limitations of applying the Bayesian paradigm to machine learning. Although *we have shown that some formal justifications remain unsatisfactory*, the conceptual shift introduced by adopting an optimization-centric view of Bayesian inference is a step toward a unified framework for machine learning methods [6], and has important practical implications.

References

[1] Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. "An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference". In: *Journal of Machine Learning Research* 23.132 (2022), pp. 1–109. URL: http://jmlr.org/papers/v23/19-1047.html.

[2] Pierre Alquier, James Ridgway, and Nicolas Chopin. "On the properties of variational approximations of Gibbs posteriors". In: *Journal of Machine Learning Research* 17.236 (2016), pp. 1–41. URL: http://jmlr.org/papers/v17/15-290.html.

[3] Pascal Germain et al. *PAC-Bayesian Theory Meets Bayesian Inference*. 2017. arXiv: 1605.08636 [stat.ML]. URL: https://arxiv.org/abs/1605.08636.

[4] P. G. Bissiri, C. C. Holmes, and S. G. Walker. "A General Framework for Updating Belief Distributions". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78.5 (Feb. 2016), pp. 1103–1130. ISSN: 1467-9868. DOI: 10.1111/rssb.12158. URL: http://dx.doi.org/10.1111/rssb.12158.

[5] Jeremias Knoblauch. *Frequentist Consistency of Generalized Variational Inference*. 2019. arXiv: 1912.04946 [math.ST]. URL: https://arxiv.org/abs/1912.04946.

[6] Veit David Wild et al. *A Rigorous Link between Deep Ensembles and (Variational) Bayesian Methods*. 2023. arXiv: 2305.15027 [stat.ML]. URL: https://arxiv.org/abs/2305.15027.

[7] Charles Blundell et al. *Weight Uncertainty in Neural Networks*. 2015. arXiv: 1505.05424 [stat.ML]. URL: https://arxiv.org/abs/1505.05424.

[8] Rajesh Ranganath, Sean Gerrish, and David Blei. "Black Box Variational Inference". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. Ed. by Samuel Kaski and Jukka Corander. Vol. 33. Proceedings of Machine Learning Research. Reykjavik, Iceland: PMLR, 22–25 Apr 2014, pp. 814–822. URL: https://proceedings.mlr.press/v33/ranganath14.html.

[9] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407. ISSN: 00034851. URL: http://www.jstor.org/stable/2236626 (visited on 03/20/2025).

[10] Léon Bottou. "On-line learning and stochastic approximations". In: *On-Line Learning in Neural Networks*. USA: Cambridge University Press, 1999, pp. 9–42. ISBN: 0521652634.

[11] Kyurae Kim et al. "On the Convergence of Black-Box Variational Inference". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 44615–44657. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/8bea36ac39e11ebe49e9eddbd4b8bd3a-Paper-Conference.pdf.

[12] Justin Domke, Robert Gower, and Guillaume Garrigos. "Provable convergence guarantees for black-box variational inference". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 66289–66327. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/d0bcff6425bbf850ec87d5327a965db9-Paper-Conference.pdf.