Article report: The Bayesian Learning Rule (2022)

NDIAYE Aminata
M2 MASH
Paris Dauphine-PSL
aminata.ndiaye@dauphine.eu

SONGO Edimah
M2 MASH
Paris Dauphine-PSL
edimah.songo@dauphine.eu

Abstract

In the article *The Bayesian Learning Rule* [1], Mohammad Emtiyaz Khan and H$\mathring{a}$vard Rue postulate that many well-known algorithms used in Machine Learning and Deep Learning are applications of a single Bayesian principle. The so-called "Bayesian Learning Rule" provides a unified framework from which we can derive both established algorithms (such as Ridge Regression) and new algorithms. Its aim is to find a posterior distribution for the parameter of interest, and it relies heavily on natural gradients, which are a way to take into account the geometry of the parameter space in the optimisation step of the algorithm.

1 Motivation & Introduction

Many current learning algorithms are created by mixing algorithms from different disciplines. For example, Stochastic Gradient Descent uses both classical Gradient Descent and Monte Carlo estimators.

The article was born out of this assessment, and the goal of the authors was to provide a way to unify, generalise and potentially improve many of these well-known algorithms through what they call the Bayesian Learning Rule.

The Bayesian Learning Rule in a nutshell: In this new framework, we look for a posterior distribution of the parameter we would typically update in classical optimisation. As we will see in the next section, this choice of posterior minimises both the empirical risk and the "distance" to the prior. Making sure we take previous information into account.

In this report, we first explain our understanding of the Bayesian Learning Rule and natural gradients. Then, we work through the application to Ridge Regression, a known learning algorithm, and explain the underlying structure of the solution. Finally, we conclude on the relevance of this new rule and share our opinion on the matter.

2 The Bayesian Learning Rule

We first have to formulate the problem. Our starting point is the variational formulation by Zellner [3] which allows use to write our bayesian objective as follow:

$$q_*(\theta) = \arg\min_{q(\theta)} \mathbb{E}_q \left[\sum_{i=1}^{N} \ell(y_i, f_\theta(x_i)) \right] + \mathbb{D}_{KL}[q(\theta) \parallel p(\theta)] \tag{1}$$

This equation is derived from Bayesian principles. Actually $q_*(\theta)$ defines a generalized posterior.

Remark 1. *For a prior distribution $p(\theta) \propto exp(-\mathbb{D}_{KL}[\cdot \parallel \cdot])$ and for a likelihood proportional to $\exp(-\ell(y_i, f_\theta(x_i)))$, then, q_* is the posterior distribution for θ.*

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

We assume that data are conditionally independent. The log posterior of θ is:

$$p(\theta|\mathcal{D}) = \frac{p(\theta)\prod_{i=1}^{N} p(y_i|f_\theta(x_i))}{\mathcal{Z}(\mathcal{D})}$$

We want to minimize the the following function with respect to q. Taking the negative log-likelihood as our loss function, we get :

$$\begin{aligned}
\mathcal{L}(q) &= -\mathbb{E}_{q(\theta)}\left[\sum_{i=1}^{N} \log p(y_i|f_\theta(x_i))\right] + \mathbb{D}_{KL}[q(\theta) \parallel p(\theta)] \\
&= \mathbb{E}_{q(\theta)}\left[\log 1/\prod_{i=1}^{N} p(y_i|f_\theta(x_i))\right] + \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)}\right] \\
&= \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta)\prod_{i=1}^{N} p(y_i|f_\theta(x_i))}\right] + \log \mathcal{Z}(\mathcal{D}) - \log \mathcal{Z}(\mathcal{D}) \\
&= \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{\frac{p(\theta)}{\mathcal{Z}(\mathcal{D})}\prod_{i=1}^{N} p(y_i|f_\theta(x_i))}\right] - \log \mathcal{Z}(\mathcal{D}) \\
&= \mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta|\mathcal{D})}\right] - \log \mathcal{Z}(\mathcal{D}) = \mathbb{D}_{KL}[q(\theta) \parallel p(\theta|\mathcal{D}) - \log \mathcal{Z}(\mathcal{D})
\end{aligned}$$

The right choice of $q(\theta)$ to minimize $\mathcal{L}(q)$ is $q(\theta) = p(\theta|\mathcal{D})$. We recover the posterior distribution of θ. We understand better know the bayesian set up.

The next step is to design the algorithm to resolve equation (1).

2.1 Sub-class of distributions

We have to choose a sub-class of distribution on which we will find the minimizer of our objective function. In the article, the autors worked with $\mathcal{Q}$ a set of regular and minimal exponential family, then:

$$q(\theta) = h(\theta)\exp\left[\langle\lambda, T(\theta)\rangle - A(\lambda)\right]$$

Where:

1. Ω is a non-empty open set such that $\lambda \in \Omega \in \mathbb{R}^M$,for which the cumulant function $A(\lambda)$ is finite, strictly convex and differentiable.
2. $T(\theta)$ is the natural sufficient statistics.
3. h is a non-negative function.
4. The expectation parameters are $\mu = \mathbb{E}_{\mathrm{q}}\left[\mathbb{T}(\theta)\right]$, it is also a bijective function of λ.

Remark 2. *The advantage of using exponential families is the efficiency and the simplicity of computation in this particular setting. It is still a sufficiently general configuration to derive a large number of algorithm with the Bayesian learning rule.*

2.2 The optimizing algorithm

The Bayesian learning rule is an optimizing algorithm which works iteratively. The minimiser of the objective function q_* belongs to an exponential family and we have to recover its natural parameters λ_*. The update at each iteration is the following:

$$\lambda_{t+1} \leftarrow \lambda_t - \rho_t \tilde{\nabla}_\lambda\left[\mathbb{E}_{q_t}[\bar{\ell}(\theta)] - \mathcal{H}(q_t)\right]$$

where:

- $\mathcal{H}(q_t) = \mathbb{E}_q\left[-\log q(\theta)\right])$
- ρ_t is the learning rate at time t
- $\tilde{\nabla}_\lambda$ is the **natural gradient**

2.3 The natural gradient

Natural Gradients implement knowledge about the curvature of the parameter space into the algorithm, and consequently, maximise the amount of information used by the update step. Thanks to them, we are able to improve the convergence rate of the optimisation process. They are defined as follows for :

$$\tilde{\nabla}_\lambda \mathbb{E}_q(\cdot) = \mathbf{F}(\lambda)^{-1} \nabla_\lambda \mathbb{E}_q(\cdot) = \nabla_{\{\nabla_\lambda A(\lambda)\}} \mathbb{E}_q(\cdot)$$

For :

- λ the natural parameter,
- $\mathbf{F}(\lambda) = \nabla_\lambda^2 A(\lambda)$ the Fisher Information Matrix

Why is it relevant ? In the paper, the target natural parameter λ^* is equal to the natural gradient of the expected negative-loss :

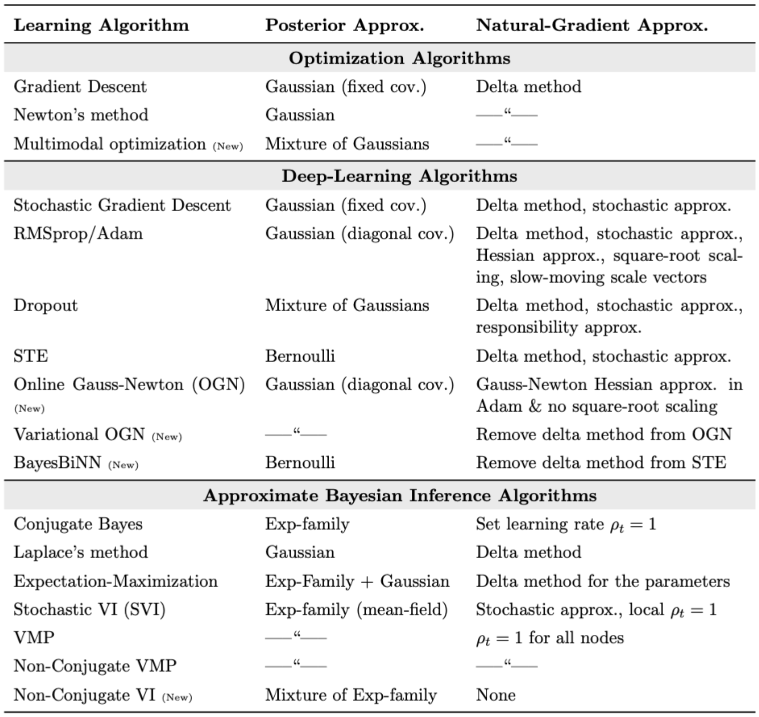
$$\lambda^* = \tilde{\nabla}_\lambda \mathbb{E}_{q^*}[-\bar{\ell}(\theta)] \tag{2}$$

3 Applications of the Bayesian Learning Rule

In the paper, the authors have worked through examples like Gradient Descent, Newton's method and Ridge Regression. For the purpose of our class, we will focus on the Ridge Regression.

3.1 Generalisation

The solution of the problem is characterised by the choice of $\mathcal{Q}$ and the natural gradient approximation method. The authors have drawn up a table of ways to obtain known algorithms through the rule.



Learning Algorithm	Posterior Approx.	Natural-Gradient Approx.
Optimization Algorithms		
Gradient Descent	Gaussian (fixed cov.)	Delta method
Newton's method	Gaussian	——“——
Multimodal optimization (New)	Mixture of Gaussians	——“——
Deep-Learning Algorithms		
Stochastic Gradient Descent	Gaussian (fixed cov.)	Delta method, stochastic approx.
RMSprop/Adam	Gaussian (diagonal cov.)	Delta method, stochastic approx., Hessian approx., square-root scaling, slow-moving scale vectors
Dropout	Mixture of Gaussians	Delta method, stochastic approx., responsibility approx.
STE	Bernoulli	Delta method, stochastic approx.
Online Gauss-Newton (OGN) (New)	Gaussian (diagonal cov.)	Gauss-Newton Hessian approx. in Adam & no square-root scaling
Variational OGN (New)	——“——	Remove delta method from OGN
BayesBiNN (New)	Bernoulli	Remove delta method from STE
Approximate Bayesian Inference Algorithms		
Conjugate Bayes	Exp-family	Set learning rate $\rho_t = 1$
Laplace's method	Gaussian	Delta method
Expectation-Maximization	Exp-Family + Gaussian	Delta method for the parameters
Stochastic VI (SVI)	Exp-family (mean-field)	Stochastic approx., local $\rho_t = 1$
VMP	——“——	$\rho_t = 1$ for all nodes
Non-Conjugate VMP	——“——	——“——
Non-Conjugate VI (New)	Mixture of Exp-family	None

Figure 1: Summary table of learning algorithms derived by the Bayesian Learning Rule

The only required steps are then :

- Choosing a posterior approximation in the exponential family
- Choosing a way to approximate the natural gradient

3.2 Ridge Regression

With a quadratic loss $\bar{\ell}(\theta) = \frac{1}{2}(y - X\theta)^T(y - X\theta) + \frac{1}{2}\delta\theta^T\theta$ for some penalty term $\delta > 0$, we know the closed form solution to be $\theta^* = (X^TX + \delta I)^{^-1}X^Ty$.

Natural gradients in Ridge Regression We take $\mathcal{N}(m, S)$ as a candidate posterior. Recall that $\mu = \mathbb{E}[T(\theta)]$. The expected loss is $\mathbb{E}[\bar{\ell}(\theta)] = -y^TX\mu^{(1)} + \text{Tr}[\frac{1}{2}(X^TX + \delta I)\mu^{(2)}]$, and through its linearity we derive the natural gradients :

$$\tilde{\nabla}_{\mu^{(1)}}\mathbb{E}_q[\bar{\ell}(\theta)] = -X^Ty \text{ and } \tilde{\nabla}_{\mu^{(2)}}\mathbb{E}_q[\bar{\ell}(\theta)] = \frac{1}{2}(X^TX + \delta I)$$

Finding the classical formula in practice Unlike other algorithms, we derive the original we can find the closed-form solution above using natural gradients with no approximation.

Using equation (2), we find that :

$$S^*m^* = X^Ty \text{ and } S^* = X^TX + \delta I$$

Consequently :

$$\theta^* = m^* = (S^*)^{^-1}X^Ty = (X^TX + \delta I)^{-1}X^Ty$$

4 Discussion

Potential of the Bayesian Learning Rule: Finding a Bayesian link between many well-known learning algorithm has opened a whole field of research possibilities and could potentially contribute to advance the state-of-the-art in many fields such as optimisation, Deep Learning, and graphical models.

Drawbacks and criticism: The exponential family is a very convenient functional space to work with thanks to its useful properties, and it contains important probability distributions.

But the constraint that the posterior distribution must be part of that family could potentially restrict the amount of algorithms we can derive from the rule.

Furthermore, natural gradients are not always easy to compute for more complex distributions.

Lie-Group Bayesian Learning Rule: Mohammad Emtiyaz Khan and others have identified a third drawback that is out of the scope of our curriculum : the updated parameters might not stay on the manifold. They have recently come out with a new article to address all of these issues through Group Theory. In *The Lie-Group Bayesian Learning Rule* **?**, they build an extension to the rule to algebraic structures that do not have this problem.

Although the knowledge of topology to fully understand this work eludes us for now, this seems like a great way to address the problem.

5 Conclusion

Most known learning algorithms learn by using provided data and revising prior beliefs.

The paper explains that building such algorithms can be equivalent to optimising a Bayesian objective. This is intuitively sound, as focusing on finding a candidate posterior distribution on the target parameter means that we use prior information. To that end, the choice of exponential family distributions for such candidates makes it easier to implement.

Furthermore, we use Fisher scaling to modify the gradients. The use of information geometry through natural gradients is another great way to conserve information : it takes into account the geometry of the parameter space. We have seen through our Ridge Regression application that high order information even appears in the solution, and the optimal natural parameter can be found with natural gradients.

Although the implementation of algorithms under the BLR framework isn't necessarily optimal (computationally challenging gradients), it is a good way to generalise learning and potentially derive new algorithms.

Our opinion on the article: We have found the main subject matter very compelling : being able to break down famous algorithms to a singular bayesian rule is both simple and fascinating. However the article was at times difficult to understand : some notations weren't clearly defined, its structure was not always easy to follow, and many incomplete proofs were not trivial. But the article was very clear about its purpose and full of examples and applications to many different fields, and very exhaustive.

Bibliography

[1] Mohammad Emtiyaz Khan and Håvard Rue. *The Bayesian Learning Rule*. 2022. arXiv: 2107.04562 [stat.ML].

[2] Eren Mehmet Kıral, Thomas Möllenhoff, and Mohammad Emtiyaz Khan. *The Lie-Group Bayesian Learning Rule*. 2023. arXiv: 2303.04397 [cs.LG].

[3] Arnold Zellner. "Optimal Information Processing and Bayes's Theorem". In: *The American Statistician* 42.4 (1988), pp. 278–280. ISSN: 00031305. URL: http://www.jstor.org/stable/2685143 (visited on 03/23/2023).