Boosting Black-Box Variational Inference by Incorporating the Natural Gradient

Felix Trusheim
Robert Bosch GmbH
Stuttgart, Germany
Felix.Trusheim@de.bosch.com

Alexandru Paul Condurache
Robert Bosch GmbH
Stuttgart, Germany
AlexandruPaul.Condurache@de.bosch.com

Alfred Mertins
University of Lübeck
Lübeck, Germany
mertins@isip.uni-luebeck.de

***Abstract*—In this paper we present a modification of the popular Black-Box Variational Inference (BBVI) approach which significantly improves the computational efficiency of the inference. We achieve this performance boost by replacing the standard gradient in the stochastic gradient ascent framework of BBVI with the natural gradient. Our experimental results (e.g. training of neutral networks) show that the proposed method outperforms the original BBVI algorithm on both synthetic and real data.**

I. Introduction

Solving a probabilistic regression problem of the form $y_{out} = f(y_{in}, x)$ is a challenging task, because the identification of those latent parameters x which link the observable data y_{in} and y_{out} best, generally requires the estimation of the extremal point of the corresponding posterior $p(x|y)$. This estimation problem often cannot be solved with the help of differential calculus due to its structure. In these cases, it is appropriate to approximate the posterior by a traceable model $q(x)$ in order to perform the estimation indirectly. The achievable accuracy of this solution is related to the accuracy of the approximation model $q(x)$. An established class of methods to approximate the posterior density is *Variational Inference* proposed by Jordan et al. [5]. Variational-Inference (VI) based methods cast the approximation of $p(x|y)$ as a variational problem with the help of the so-called Evidence Lower Bound (ELBO). Thereby, the ELBO corresponds to a scalar objective function L which reflects the dissimilarity between the process's joint distribution $p(x, y)$ and $q(x)$. A very popular VI method is the Black-Box Variational Inference (BBVI) algorithm proposed by Ranganath et al. [10]. This method is characterized by its simple implementation and its independence from differential derivatives of the process model. This makes BBVI quite generic and therefore usable in many applications. BBVI solves a variational problem by using a parametric model $q(x|\lambda)$ in the framework of stochastic gradient ascent (SGA). If the parameter manifold of a variational problem corresponds to an Euclidean space, SGA evolves in the direction of largest increase of the ELBO by following the gradient. However, if the structure of the parameter manifold within a problem differs from that, the gradient does not reflect the largest increase of the ELBO and thereby misleads SGA. Therefore, SGA needs to be adapted accordingly. In some cases, this can be done with the help of the so-called natural gradient proposed by Amari [1]. In this contribution, we propose a modified version of the BBVI method that incorporates the natural gradient and thereby enables a faster and computational more efficient inference. We will analyze the effectiveness of our method empirically with both real and synthetic data.
The remainder of the paper is structured as follows: In *Sec. II* we briefly explain the basics of Black-Box Variational Inference algorithm. In *Sec. III* we outline our NG-BBVI algorithm. Within *Sec. IV* we relate our proposed idea to the state-of-the-art. Subsequently in *Sec. V*, we present experimental results and based on that discuss the effectiveness of our algorithm in comparison to the BBVI algorithm. In *Sec. VI*, we summarize our approach and give an outlook on planned advancements.

II. Solving Regression Problems with Black-Box Variational Inference

If the posterior $p(x|y)$ of a regression problem is not differentiable, then it is not possible to determine the maximum of the density with the help of a gradient-based analysis. An alternative for that is the following two step procedure:

1) Approximation of the posterior $p(x|y)$ by a simple, traceable density $q(x)$

$$p(x|y) \approx q(x) \tag{1}$$

2) Approximative estimate of the maximum of $p(x|y)$ based on the corresponding approximation $q(x)$

$$x_{opt} = \underset{x}{\operatorname{argmax}} \left(p(x|y)\right) \approx \underset{x}{\operatorname{argmax}} \left(q(x)\right) \ . \tag{2}$$

The essential part of this procedure is an accurate approximation of the posterior $p(x|y)$ by $q(x)$. An efficient method to compute such an approximation is the BBVI algorithm proposed by Ranganath [10]. This method is based on the principles of VI proposed by Jordan et al. [5], which formalizes the approximation of the posterior $p(x|y)$ by $q(x)$ as a variational problem over the complementary dissimilarity between the joint distribution $p(x, y)$ and $q(x)$. Therefore, BBVI uses the ELBO objective which relies on the negative Kullback-Leibler divergence. Beyond that, BBVI restricts $q(x)$ to a parametric model $q(x|\lambda)$. This regularization significantly reduces the complexity of the variational problem and thereby simplifies the mathematical handling. Based on that, the optimal parameter configuration λ_{opt} of $q(x|\lambda)$ is:

$$\begin{aligned}\lambda_{opt} &= \underset{\lambda}{\operatorname{argmin}}\left(KL\left(p(x|y)||q(x|\lambda)\right)\right) \qquad (3)\\ &= \underset{\lambda}{\operatorname{argmax}}(\underbrace{E_{q(x|\lambda)}\left\{\log\left(\frac{p(x,y)}{q(x|\lambda)}\right)\right\}}_{ELBO}) \ .\end{aligned}$$

Since in general, the maximum point of the ELBO cannot be determined analytically, BBVI applies SGA. Therefore, BBVI approximates the intractable gradient of the ELBO

$$\begin{aligned}\frac{dL}{d\lambda} &= E_{q(x|\lambda)}\left\{\frac{d}{d\lambda}\log(q\left(x|\lambda\right))\right. \qquad (4)\\ &\quad \cdot \quad \left.\left(\log\left(p\left(x,y\right)\right) - \log\left(q\left(x|\lambda\right)\right)\right)\right\}\end{aligned}$$

with the help of the score function $\frac{d}{d\lambda}\log(q\left(x|\lambda\right))$ as well as Monte-Carlo integration. The variance caused by Monte-Carlo integration is reduced with the help of a Rao-Blackwellization [2] of the expression and the introduction of control variates. Here, the Rao-Blackwellization corresponds to a decomposition of the global expectation into variable-wise expectations. This decomposition, which is enabled by a mean-field structured model $q(x|\lambda)$

$$q(x|\lambda) = \prod_{n=1}^{N} q_n(x_n|\lambda_n) \ , \qquad (5)$$

reduces the variance of the ELBO gradient significantly and thereby improves its usage within SGA. In doing so, BBVI achieves, in contrast to the other popular VI method the Auto-Encoding Variational-Bayes (AEVB) algorithm proposed by Kingma et alt. [7], an approximation of the ELBO gradient without relying on differential derivatives of the joint distribution $p(x,y)$. This makes BBVI quite generic.

Within its applied SGA framework, BBVI assumes that the gradient of the ELBO points in the direction of the largest increase of the ELBO and therefore represents the optimal search direction. But this assumption only holds if the parameter space corresponds to an Euclidean space. Here, the Euclidean space represents a space that is defined by an orthonormal base system and therefore allows one to calculate the distance between a point w and the origin 0 with the help of the Euclidean metric $||w|| = \sqrt{\sum_{i=1}^{N} w_i^2}$. However, this characterization of the parameter space is not guaranteed within an arbitrary variational problem. Thus, the usage of the standard gradient in SGA can lead to a slow, a suboptimal or even a failing inference. One solution to address this problem, is the natural gradient proposed by Amari [1].

III. Natural-Gradient based Black-Box Variational Inference

Amari showed that for a general scalar function $f(x)$ with an arbitrarily characterized parameter space, the direction of the largest increase of $f(x)$ is described by:

$$p = G(x)^{-1} \cdot \frac{df}{dx} \ . \qquad (6)$$

The author denotes this direction as the *natural gradient*. The matrix $G(x)$ in this expression corresponds to Riemannian metric tensor (RMT), which encodes the structural shift of parameter space of $f(x)$ around x relative to an Euclidean space. Hence, in the special case, if the local parameter space of $f(x)$ is equivalent to an Euclidean space, $G(x)$ corresponds to an identity matrix. As a side note, it can be mentioned that the definition of p offers a structural analogy to the search direction of $f(x)$ in case of an Newton-method based optimization scheme:

$$p = -H(x)^{-1} \cdot \frac{df}{dx}. \qquad (7)$$

However, different from Eq. 6, $H(x)$ does not correspond to the RMT, but instead to the Hessian matrix of $f(x)$. Similarly to the RMT, the Hessian incorporates local curvature information to define of a suitable search direction, but effectively uses a different formalization.

Now, in order to transfer the idea of the natural gradient to ELBO problem, a suitable RMT definition is needed which corresponds to the ELBO objective. For this purpose, Honkela et al. [4] propose the Fisher information matrix (FIM). The FIM corresponds to the second derivative of the Kullback-Leibler (KL) divergence of the approximation model $q(x|\lambda)$ within the following setup:

$$\begin{aligned}F(\lambda) &= \left.\frac{d^2\mathrm{KL}(q(x|\lambda)||q(x|\hat{\lambda}))}{(d\hat{\lambda})^2}\right|_{\hat{\lambda}=\lambda} \qquad (8)\\ &= E_{q(x|\lambda)}\left\{\left(\frac{d}{d\lambda}\log(q(x|\lambda))\right)\left(\frac{d}{d\lambda}\log(q(x|\lambda)\right)^T\right\}.\end{aligned}$$

According to this expression, the FIM can be characterized as the expectation of the outer product of the score function with itself. This equation encodes the local curvature information of the parameter space for the model around the parameter λ and thereby meets the idea of the RMT. Therefore, we use the FIM to incorporate the natural gradient idea into the SGA framework of the BBVI algorithm. For the practical implementation, we first apply the mean-field restrictions of the approximation model $q(x|\lambda)$ to the computation of the FIM. By doing so, the FIM simplifies to:

$$\hat{F}(\lambda) = \begin{cases} E_{q_i(x_i|\lambda_i)}\left\{\left(\frac{d}{d\lambda_i}\log(q_i(x_i|\lambda_i))\right)\right. & \\ \quad \left.\left(\frac{d}{d\lambda_i}\log(q_i(x_i|\lambda_i))\right)^T\right\} & , i = j \\ 0 & , i \neq j \end{cases} \ . \qquad (9)$$

This decomposition of the global expectation of Eq. 8 into variable-wise local expectations reduces the variance of the FIM significantly and thereby corresponds to a Rao-Blackwellization [2]. Aside of a few special approximation models, like a Gaussian-distribution, it is generally hard to calculate these local expectations analytically. Therefore, similar to Ranganath's considerations relating to the ELBO and the ELBO gradient, we suggest to approximate these expectations with the help of Monte Carlo integration. By applying this, the FIM becomes:

$$\hat{F}(\lambda) \approx \begin{cases} \frac{1}{S}\sum_{s=1}^{S}\left(\frac{d}{d\lambda_i}\log(q_i(x_i^{(s)}|\lambda_i))\right) \\ \quad \left(\frac{d}{d\lambda_i}\log(q_i(x_i^{(s)}|\lambda_i))\right)^T & , i = j \\ \qquad 0 & , i \neq j \end{cases} \tag{10}$$

with

$$x_i^{(s)} \sim q_i(x_i|\lambda_i) \ . \tag{11}$$

The approximation of the local expectations requires the evaluation of the score function at the samples $x_i^{(s)}$. However, these calculations are also required for the approximation of the ELBO or the ELBO gradient. This implies that these calculations can be reused in this context without any additional numerical effort. The block-diagonal structure of the FIM allows for an efficient variable-wise inversion of the matrix. This enables a decomposition of the calculation of the search direction p into variable-wise calculations. By embedding these ideas in the framework of the original BBVI algorithm, our proposed variant, NG-BBVI, is given by Alg. 1.

In our algorithm, we apply an Adam [6] scheme for scaling and modifying the search directions and thereby replace the AdaGrad [3] scheme, which is proposed in the original BBVI algorithm [10]. We explain this decision in detail in Sec. V.

Algorithm 1 Natural-Gradient based Black-Box Variational Inference

```
Input : data $y$, joint distribution $p(x,y)$, mean-field
  structured model $q(x|\lambda)$
Initialization : $\lambda, S, t, \beta_1, \beta_2, \eta, c$
repeat
    // Draw Samples
    for $s = 1$ to $S$ do
        $x[s] \sim q(x|\lambda)$
    end for
    $X[1] \leftarrow$ subset of $x$; $X[2] \leftarrow x$ // Group sets
    // Estimation of ELBO Natural Gradient
    for $n = 1$ to $N$ do // Loop over latent variable
        $a = 0$
        for $p = 1$ to $2$ do // Loop over sample sets
            for all sample $s$ of $X[p]$ do
                $h[s] = \frac{d}{d\lambda_n}\log\left(q_n\left(X^{[n]}[p][s], \lambda_n\right)\right)$
                $f[s] = h[s] \cdot (\log\left(p_n\left(X[p][s], y\right)\right) -$
                $\log\left(q_n\left(X^{[n]}[p][s], \lambda_n\right)\right) - a)$
            end for
            switch $p$ do
                case 1 // Control Variates
                    $a = \frac{Cov(h,f)}{Var(h)}$
                end case
                case 2 // Natural Gradient
                    $F \leftarrow \frac{1}{|X[p]|}\sum_{s=1}^{|X[p]|} h[s](h[s])^T$
                    $z[n] \leftarrow F^{-1}\left(\frac{1}{|X[p]|}\sum_{s=1}^{|X[p]|} f[s]\right)$
                end case
            end switch
        end for
    end for
    // Update via Adam
    $m = \beta_1 \cdot m + (1-\beta_1) \cdot z$
    $v = \beta_2 \cdot v + (1-\beta_2) \cdot \|z\|^2$
    $\hat{m} = \frac{m}{1-\beta_1^t}$ ; $\hat{v} = \frac{v}{1-\beta_2^t}$
    $\lambda = \lambda + \eta \cdot \frac{\hat{m}}{\sqrt{\hat{v}}+\epsilon}$
until $||d\lambda|| < c$ // Convergence Criteria
```

IV. Related work

Since the publication of the BBVI algorithm, there have been various proposals for improving different aspects of the approach. The focus of these follow-up publications varies. Ruiz et al. [11] suggest the idea to adapt the formally fix structure of $q(x|\lambda)$ dynamically during the inference procedure. Titsias et alt. [12] propose an inference procedure for approximation models with a quasi mean-field structure based on local expectations and an efficient sampling scheme within Monte-Carlo integration. However, similarly to original BBVI algorithm, these proposals try to improve the algorithm under the assumption that parameter manifold of a variational problem is an Euclidean space and therefore the standard gradient of the ELBO represents the best search direction within a SGA framework. But such a manifold characterization is not guaranteed within an arbitrary variational problem. This aspect was first addressed by Honkela et al. [4], long before the publication of the BBVI algorithm. As a solution, the authors propose a SGA-based inference algorithm that uses the natural gradient [1] to consider the metric characteristics of the parameter manifold. They suggest the FIM as an appropriate RMT equivalent. The authors present their idea in a very general form. But unlike our algorithm, they do not offer an integrated algorithmic concept.

V. Experiments

In order to evaluate the performance of NG-BBVI with regard to its computational efficiency, we examine three applications. At the core of each application lies a probabilistic regression problem. We assume that the probabilistic properties of these problems are similar and accurately modeled by the generative Bayesian network depicted in Fig. 1 (top). Furthermore, we suppose, due to the natural characteristics of perception noise and latent-parameter distributions, that within each application the prior $p(x)$ and the likelihood $p(y|x)$ are most precisely modeled as Gaussian distributions. In that context, our proposed approach will be used to approximate the complex-structured posterior $p(x|y)$ by a simple, traceable model $q(x|\lambda)$. After that, the identified approximations will be used to perform a MAP analysis and thereby estimate the optimal latent parameters x of the regression models. Within each application, we will analyze the performance of our NG-BBVI algorithm based on its computational-convergence efficiency (CCE). Here, the CCE describes the ratio of the approximation quality of $q(x|\lambda)$ relative to the corresponding computational cost. We use the value of the ELBO as a representative quality indicator. For practicability reasons, we measure the computational cost indirectly by the corresponding calculation time. In order to rank the CCE of our method, we will compare it with the original BBVI algorithm. Within this comparison, we will process both approaches with different stochastic learning rates to separate the influence of the search direction and the stochastic learning rate on the CCE. We implemented the experiments completely in MATLAB. Hence, we realized the Importance Sampling within the Monte-Carlo integrations of both approaches with the help of MATLAB's standard sampling method the Ziggurat algorithm proposed by Marsaglia and Tsang [9]. To increase the significance of our results, we repeated all experiments several times and averaged the single-run results.

A. *Classify synthetic feature data*

First, we want to train a neural network for the classification of synthetic 2D feature vectors. The vectors are divided into 4 classes and thereby are distributed in the feature space according to Fig. 1 (middle). For the supervised training we use 70% of the total of 600 feature vectors. The remaining 30% define the test set. As a result of the low dimension of feature vectors and their distribution within the feature space, we rely on a small neural network with just one hidden layer of 10 neurons. As a consequence, the neural network has a 2-10-4 layer design and therefore offers 74 trainable latent parameters. Besides that, the network uses hyperbolic tangent shaped activation functions. At the beginning of the training, the latent parameters are initialized randomly with the help of a normal distribution $N(0, 1)$. In order to estimate the parameters during the training, we rely on the proposed NG-BBVI algorithm. Within NG-BBVI, we use a Gaussian-distributed model $q(x|\lambda)$ with a diagonal covariance. We approximate the expectations of the ELBO and the ELBO gradient by $N = 20$ samples per iteration of SGA. The control variates a_{opt} in the ELBO gradient are estimated with half of the samples. We apply the same sampling setup within the original BBVI algorithm. The large amount of training data by only 4 classes with a small inner class variance causes a significant informative redundancy in the data. This allows us to use of a mini-batch scheme to accelerate the inference. Hence, we only use $K = 20$ training samples per SGA iteration within both BBVI algorithms. The convergence dynamics of all variants are presented in Fig. 2 (top). The Fig. also includes the true positive rates (TPR) which describe the proportion of correct classified test data relative to all test data.

B. *Classify MNIST feature data*

As a second example, we train a neural network that classifies images of handwritten numbers. As database, we use the MNIST data set [8]. The images of the MNIST data set correspond to normalized grayscale images with a resolution of 28x28 pixels. Each image contains one handwritten number in the range of 0 to 9. In total, the data set consists of 60000 training examples and 10000 test examples. Within both subsets, all 10 classes are distributed almost uniformly. We do not preprocess the images. To adapt the structure of the data to the architecture of the neural network, we transform each image into a vector of dimension 1x784. In order to keep the network complexity low, we rely on a simple design with just an input layer and an output layer. As a result, the neural network has 7850 latent parameters to be trained. Besides that, the network uses hyperbolic tangent shaped activation functions. At the beginning of the training, the latent parameters are initialized randomly with the help of a normal distribution $N(0, 1)$. In order to estimate the parameters during the training, we rely on the featured algorithms. Within each approach, we use a Gaussian-distributed model $q(x|\lambda)$ with a diagonal covariance. We approximate the expectations of the ELBO and ELBO gradient by $N = 20$ samples per iteration of SGA. The control varieties a_{opt} are estimated with half of the samples. Similar to the synthetic-data example above, the large amount of training samples by only 10 classes causes an informative redundancy within the data. This again allows the application of a mini-batch scheme to accelerate the inference. Hence, we only use $K = 200$ training samples per SGA iteration in both inference algorithms. The convergence dynamics, as well as the test data set TPRs, of both variants

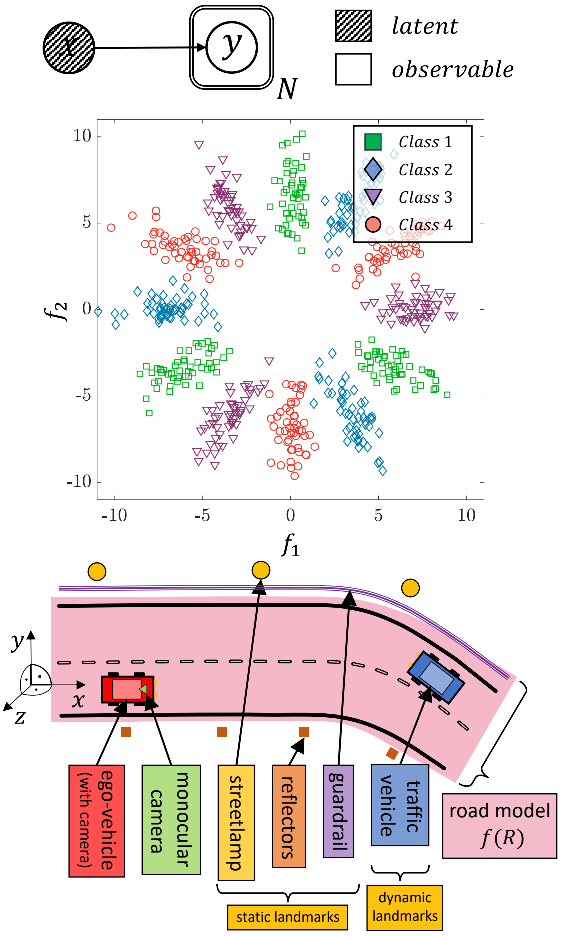


Fig. 1: Probabilistic Model (top), Feature space of *Exp. 1* (middle), Environment model of *Exp. 3* (bottom)

are presented in Fig. 2 (middle).

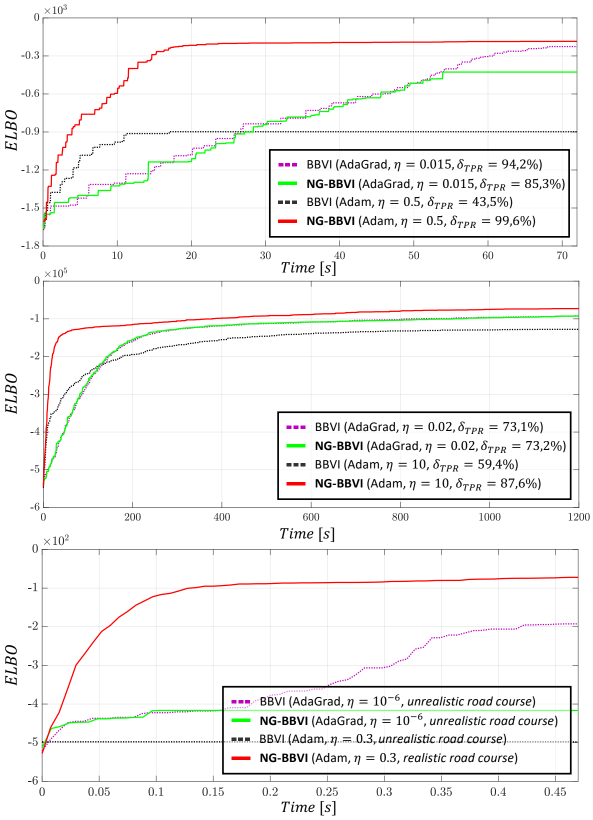
C. ADAS: Estimating the course of a road

The third highlighted regression problem originates from the environment of automated driving and therefore has a direct practical application [13]. In detail, the regression problem formalizes the estimation of the $3D$ road course in front of an automated vehicle on the basis of course-correlated static and dynamic landmarks. The estimate therefore relies on position and scene flow information from these landmarks measured by a monocular camera system. In detail, this system is installed behind the windscreen of the vehicle and looks forward in the direction of travel. Fig. 1 (bottom) illustrates the setup. Before the estimation of the road course, the acquired camera measurements are preprocessed by means of image processing, machine learning and structure-from-motion. After that, the camera data is classified, structured and transformed into the $3D$ space. The road course is modeled by a parametric nonlinear model $Y_{pre} = f(R)$. The dimension of R varies in practice between 6 to 14 and thereby characterizes the application as a regression problem with a low-dimensional parameter space. Based on that road model, the objective of the regression problem is the estimation of the configuration R_{opt} which optimally predicts the preprocessed measurements Y_{mea}. As a result of various influences, the measurement data Y_{mea} is considered as a noisy signal. Therefore, we model the regression problem as a probabilistic process with the graphical structure according to Fig. 1 (top). Again, we will solve the regression problem with both previously discussed inference algorithms. Since the road scene in front of the vehicle is continuously changing, the estimation of the road course has to be executed permanently. Within the environment of driver assistance systems this is a very challenging task due to the limited computational capabilities of these systems. Therefore, it is necessary to implement the estimation as efficient as possible or correspondingly apply a solver with a good CCE. Within both algorithms we use a Gaussian-distributed model $q(x|\lambda)$ with a diagonal covariance. To simplify the complexity of the experiment, we initialize λ randomly. In both approaches, we approximate the expectations of the ELBO as well as the ELBO gradient by $N = 10$ samples per iteration of SGA. The control variates a_{opt} are estimated with half of the samples. Since the amount of acquired camera measurements per estimation procedure is generally very limited, there is no significant informative redundancy within that data. Hence, we do not apply a mini-batch scheme within the inference. The convergence dynamics of all discussed approaches are depicted in Fig. 2 (bottom).

D. Discussion

We evaluate the computational convergence efficiency (CCE) of the discussed approaches based on the depicted convergence dynamics in Fig. 2. As mentioned above, the ELBO will be used as a representative indicator for the accuracy of the approximative solution of the MAP analysis within the regression problems. Here, the principle applies: The higher the value of the ELBO, the more accurate is the corresponding MAP solution of the problem. As illustrated in the figures, within each experiment we process both inference algorithms with different stochastic learning rates in various configurations. This procedure allows us to separate the influence of the search direction and the stochastic learning rate on the inference. As learning rates, we consider AdaGrad [3] and Adam [6] (see table in Fig. 2). In order to keep the results focused, we only vary those hyperparameters within each learning rate which affect the convergence dynamic the most. For both AdaGrad and Adam, this is the scaling parameter η. The remaining parameters for both learning rates are defined according to the usual recommendations. Thus, we set $N = 10$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. To identify the best η for each inference algorithm, we performed a grid-search based hyperparameter optimization in advance. In order to compare the algorithms, we apply the identified optimal learning rate of one algorithm correspondingly on the other algorithm.

The results of Exp. 1 (see Fig. 2 (top)) show that NG-BBVI in combination with Adam not only offers the best CCE, but also achieves the highest ELBO value of all compared methods.



AdaGrad	Adam
$Z = \sum_{i=t-N}^{t-1} p_i^T p_i$ $p_{eff}^t = \eta \ (Z^{-1/2} \circ p_t)$	$m_t = \beta_1 m_{t-1} + (1-\beta_1) p_t$ $v_t = \beta_2 v_{t-1} + (1-\beta_2) \lVert p_t \rVert^2$ $p_{eff}^t = \eta \frac{m_t}{1-\beta_1^t} \left(\sqrt{\frac{v_t}{1-\beta_2^t}} + \epsilon \right)^{-1}$

Fig. 2: Convergence dynamics of *Exp. 1* (top), *Exp. 2* (middle) and *Exp. 3* (bottom), Table of applied learning rates (bottom)

BBVI behaves noticeably worse with the same learning rate. This is a clear indicator that the parameter space of the problem does not correspond to an Euclidean space, because otherwise both algorithms would show a similar performance. BBVI only performs better when using AdaGrad. Here, the AdaGrad scheme, which considers variations between successive gradients and thereby adapts implicitly to local metric variations, generates more suitable search directions in SGA. On the other hand, the same AdaGrad scheme works suboptimal in conjunction with NG-BBVI. A combination of both seems to neutralize the adaptiveness of the natural gradient within NG-BBVI. Finally, these ELBO convergence dynamics translate into classifiers with corresponding performances. This is reflected by the TPR (see δ_{TPR} in legend in Fig. 2), which are achieved on the test data set. The results of Exp. 2 (see Fig. 2 (middle)) reflect a pattern which is already known from Exp. 1. NG-BBVI in combination with Adam achieves the best CCE. BBVI combined with that learning rate performs noticeably worse. This indicates, again, that the parameter space of this problem does not correspond to an Euclidean space. BBVI achieves the best CCE in combination with AdaGrad. Similar to Exp. 1, NG-BBVI, combined with that learning rate, shows a comparable convergence dynamic. Once again, AdaGrad seems to compensate the influence of the FIM in SGA. As in Exp. 1, the achieved convergence levels correlate directly with the quality of the corresponding classifiers. Here, it is worth to note that the best NG-BBVI based classifier offers a TPR which is close to the absolute best TPR rate of 88% for such a designed classifier (2-layer) [8]. However, this reference classifier is trained within a deterministic modeling and with the help of backpropagation. In Exp. 3 (see Fig. 2 (bottom)) the patterns of Exp. 1 and Exp. 2 repeat. But this time the differences between the algorithms are even more pronounced. This implies that the parameter space of road-course estimation problem differs even more from an Euclidean space than the parameter spaces of the classification problems. In the best configuration BBVI not only converges much slower than the best configuration of NG-BBVI, but it also achieves a significantly lower convergence level. Translating the estimated latent parameters of this solution into a corresponding road-course estimation (see the qualitative ratings within the legend) results in a road model which adapts poorly to the measured landmarks. The NG-BBVI based solution, on the other hand, adapts realistically and therefore is useful in practice. In addition, this experiment shows even more that NG-BBVI and AdaGrad are not a good match to each other. The influence of the FIM (see Eq. 6) on the gradient, here, seems completely compensated by AdaGrad in SGA. In summary, it can be stated that in all experiments NG-BBVI, in conjunction with Adam, offers significant advantages over the combination of BBVI and AdaGrad. We assume that this is caused by the explicit consideration of the non-Euclidian character of the parameter spaces of these problems in NG-BBVI. Within none of the experiments, the additional computational cost to calculate the FIM, has a detrimental effect on the CCE performance. Therefore, from a practical point of view, the usage of NG-BBVI, combined with Adam, has to be preferred in all our experiments.

VI. Conclusion & Future Work

In this contribution we presented a modification of the BBVI algorithm which incorporates the idea of the natural gradient in its inference framework and uses Adam as an optimal companion. This effort was driven by the desire to realize a procedure that approximates the posterior $p(x|y)$ of non-differentiable probabilistic regression problems by a mean-field structured model $q(x|\lambda)$ under minimal numerical costs. In an empirical study, we were able to prove that our approach has advantages over the original BBVI algorithm [10] within the considered examples. Thereby, despite the additional effort to calculate the Fisher information matrix, our proposed algorithm was able to solve the inference problems more efficiently than the original BBVI algorithm, thanks to more suitable search directions within SGA.

In our algorithm, we see potential for further improvements. Hence, in future work we would like to investigate in detail, whether the applied SGA scheme and thereby the efficiency of the inference can profit from a combination of search directions generated by the natural gradient (see Eq. 6) as well as a Quasi-Newton method (see Eq. 7). These expectations derive from the fact that a Quasi-Newton method perhaps delivers more suitable search directions close to the extremal point than the natural gradient does.

References

[1] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning*, 12:2121–2159, 2011.

[4] A. Honkela, R. T, M. Kuusela, M. Torni, and J. Karhunen. Approximate riemannian conjugate gradient learning for fixed-form variational bayes. *SIAM Journal on Optimization*, 11:3235–3268, 2010.

[5] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An algorithm for least-squares estimation of nonlinear parameters. *Machine Learning*, 37:183–233, 1999.

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. arXiv:1412.6980.

[7] D. P. Kingma and P. Welling. Auto-encoding variational bayes, 2014. Proceedings of the 2nd International Conference on Learning Representations (ICLR).

[8] Y. LeCun, C. Cortes, and C. Burges. The mnist database of handwritten digits. 1998. http://yann.lecun.com/exdb/mnist/.

[9] G. Marsaglia and W. W. Tsang. A fast, easily implemented method for sampling from decreasing or symmetric unimodal density functions. *SIAM Journal on Scientific and Statistical Computing*, 5, 1984.

[10] R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference, 2014. 17th International Conference on Artificial Intelligence and Statistics (AISTATS), Reykjavik, Iceland.

[11] F. Ruiz, M. Titsias, and D. Blei. Overdispersed black-box variational inference, 2016. The Conference on Uncertainty in Artificial Intelligence, New York, USA.

[12] M. Titsias and M. Lazaro-Gredilla. Local expectation gradients for black box variational inference, 2015. 28th International Conference on Neural Information Processing Systems (NIPS), Montreal, Canada.

[13] F. Trusheim, A. Condurache, and A. Mertins. Visual landmark based 3D road course estimation with black box variational inference, 2017. 17th International Conference on Computer Analysis of Images and Patterns (CAIP), Ystad, Sweden.