Fast and Simple Natural-Gradient Variational Inference with Mixture of Exponential-family Approximations

Wu Lin [1] **Mohammad Emtiyaz Khan** [2] **Mark Schmidt** [1]

Abstract

Natural-gradient methods enable fast and simple algorithms for variational inference, but due to computational difficulties, their use is mostly limited to *minimal* exponential-family (EF) approximations. In this paper, we extend their application to estimate *structured* approximations such as mixtures of EF distributions. Such approximations can fit complex, multimodal posterior distributions and are generally more accurate than unimodal EF approximations. By using a *minimal conditional-EF* representation of such approximations, we derive simple natural-gradient updates. Our empirical results demonstrate a faster convergence of our natural-gradient method compared to black-box gradient-based methods. Our work expands the scope of natural gradients for Bayesian inference and makes them more widely applicable than before.

1. Introduction

Variational Inference (VI) provides a cheap and quick approximation to the posterior distribution, and is now widely used in many areas of machine learning (Kingma & Welling, 2013; Furmston & Barber, 2010; Wainwright & Jordan, 2008; Hensman et al., 2013; Nguyen et al., 2017). In recent years, many natural-gradient methods have been proposed for VI (Sato, 2001; Honkela et al., 2007; 2011; Hensman et al., 2012; Hoffman et al., 2013; Khan & Lin, 2017). These works have shown that, for specific types of models and approximations, natural-gradient methods can result in simple updates which converge faster than gradient-based methods. For example, stochastic variational inference (SVI) (Hoffman et al., 2013) is a popular natural-gradient method for conjugate exponential-family models. Unfortunately, the simplicity of natural-gradient updates is currently limited to VI with *minimal* exponential-family (EF) approximations. For such approximations, we can efficiently compute natural-gradients in the natural-parameter space without explicitly computing the Fisher information matrix (FIM) (Khan & Nielsen, 2018). Unfortunately, this property does not extend to many other approximations such as mixtures of EF distributions. Such *structured* approximations are more appropriate for complex and multi-modal posterior distributions, giving a more accurate fit than minimal EF distributions. However, computation of natural-gradients is challenging for them.

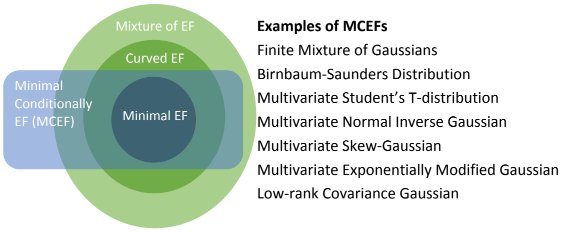
Mixture of EF
Curved EF
Minimal Conditionally EF (MCEF)
Minimal EF
Examples of MCEFs
Finite Mixture of Gaussians
Birnbaum-Saunders Distribution
Multivariate Student's T-distribution
Multivariate Normal Inverse Gaussian
Multivariate Skew-Gaussian
Multivariate Exponentially Modified Gaussian
Low-rank Covariance Gaussian

Figure 1. We derive simple natural-gradient updates for approximations with a *minimal-conditional EF* (MCEF) representation. Such approximations include all minimal (normalized) EF distributions, and some curved and mixture of EF distributions.

In this paper, we propose a simple new natural-gradient method for VI with structured approximations. We define a class of distributions which take a *minimal conditional-EF* (MCEF) form. This includes many members of mixture and curved EF distributions (see Fig. 1). Using the MCEF representation and an expectation parameterization associated with it, we derive simple natural-gradient updates. We show examples on a variety of models where simple natural-gradient updates can be used to estimate flexible and accurate structured approximations. Our empirical results show faster convergence of our method compared to gradient descent for VI. Our work extends the simplicity of natural-gradient methods making them more widely applicable than before, while maintaining their fast convergence.

1.1. Related Works

Existing work on natural-gradient VI to obtain EF approximations all assume a minimal representation to ensure invertiblility of the FIM. Such methods show fast convergence,

[1] University of British Columbia, Vancouver, Canada. [2] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. Correspondence to: Wu Lin <wlin2018@cs.ubc.ca>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

result in simple updates, and lead to a straightforward implementation for many types of models. They have been used for conditionally-conjugate EF models (Sato, 2001; Hoffman et al., 2013) and non-conjugate models (Khan & Lin, 2017), including deep neural networks (Khan et al., 2018; Zhang et al., 2018; Mishkin et al., 2018) and Gaussian processes (Khan et al., 2016; Salimbeni et al., 2018). Our work presents fast and simple updates for approximations that are beyond the reach of these works.

For structured approximations, existing works employ stochastic-gradient methods (Salimans & Knowles, 2013; Hoffman & Blei, 2015; Ranganath et al., 2016; Titsias & Ruiz, 2018; Yin & Zhou, 2018). Such methods are widely applicable, but not as widely used as their mean-field counterparts. This is because they are computationally expensive and slow to converge. Our work attempts to improve these aspects for a flexible class of approximations.

2. Natural-Gradient Variational Inference

We begin with a description of natural-gradient descent for variational inference in probabilistic models. Given a probabilistic model $p(\mathcal{D}, \mathbf{z})$ to model data $\mathcal{D}$ using latent vector $\mathbf{z}$, the goal of Bayesian inference is to compute the posterior distribution: $p(\mathbf{z}|\mathcal{D}) = p(\mathcal{D}, \mathbf{z})/p(\mathcal{D})$. This requires computation of the *marginal likelihood* $p(\mathcal{D}) = \int p(\mathcal{D}, \mathbf{z}) d\mathbf{z}$, which is a high dimensional integral and difficult to compute. VI simplifies this problem by approximating the posterior distribution $p(\mathbf{z}|\mathcal{D})$ by another distribution whose normalizing constant is easier to compute. For example, a common choice to approximate the posterior is to use a *regular*[1] exponential-family (EF) approximation,

$$q(\mathbf{z}|\boldsymbol{\lambda}_z) := h_z(\mathbf{z}) \exp \left[\langle \boldsymbol{\phi}_z(\mathbf{z}), \boldsymbol{\lambda}_z \rangle - A_z(\boldsymbol{\lambda}_z) \right], \tag{1}$$

where q denotes the approximating distribution, $\boldsymbol{\phi}_z(\mathbf{z})$ are the sufficient statistics, $h_z(\mathbf{z})$ is the base measure, and $\boldsymbol{\lambda}_z \in \Omega$ is the natural parameter with Ω being the set of valid natural-parameters (the set of $\boldsymbol{\lambda}_z$ where the log-partition function $A_z(\boldsymbol{\lambda}_z)$ is finite) and $\langle \cdot, \cdot \rangle$ denotes an inner product.[2] Such parametrized approximations can be estimated by maximizing the variational lower bound:

$$\mathcal{L}(\boldsymbol{\lambda}_z) := \mathbb{E}_q \left[\log p(\mathcal{D}, \mathbf{z}) - \log q(\mathbf{z}|\boldsymbol{\lambda}_z) \right], \tag{2}$$

which can be solved by gradient descent, as shown below:

$$\text{GD}: \quad \boldsymbol{\lambda}_z \leftarrow \boldsymbol{\lambda}_z + \alpha \nabla_{\lambda_z} \mathcal{L}(\boldsymbol{\lambda}_z), \tag{3}$$

where ∇ denotes the gradient and $\alpha > 0$ is a scalar learning rate. The GD algorithm is simple and convenient to implement by using modern automatic-differentiation methods and the reparameterization trick (Ranganath et al., 2014; Titsias & Lázaro-Gredilla, 2014). Unfortunately, such *first-order* methods can show suboptimal rates of convergence and be slow in practice.

An alternative approach is to use natural-gradient descent which exploits the information geometry of q to speed-up convergence. Assuming that the Fisher information matrix (FIM) of $q(\mathbf{z}|\boldsymbol{\lambda}_z)$, denoted by $\mathbf{F}_z(\boldsymbol{\lambda}_z)$, is positive-definite for all $\boldsymbol{\lambda}_z \in \Omega$, the natural-gradient descent (NGD) for VI in the natural-parameter space is given as follows:

$$\text{NGD}: \quad \boldsymbol{\lambda}_z \leftarrow \boldsymbol{\lambda}_z + \beta \; \left[\mathbf{F}_z(\boldsymbol{\lambda}_z)\right]^{-1} \nabla_{\lambda_z} \mathcal{L}(\boldsymbol{\lambda}_z), \tag{4}$$

The preconditioning of the gradients by the FIM leads to a proper scaling of the gradient in each dimension, and takes into account dependencies between variables. This often leads to faster convergence, particularly when the FIM is well-conditioned.

A naive implementation of (4) would require computation and inversion of the FIM. However, for specific types of models and approximations, NGVI could be simpler to compute than GD. This is true for mean-field VI in conjugate-exponential family models (Hoffman et al., 2013) as well as Bayesian neural networks (Khan et al., 2018; Zhang et al., 2018). This computational efficiency is a result of a simple NGD update to estimate EF approximations, as shown by Khan & Nielsen (2018). For EFs, we can use the expectation-parameter, defined as the function $\mathbf{m}_z(\boldsymbol{\lambda}_z) := \mathbb{E}_q \left[\boldsymbol{\phi}_z(\mathbf{z})\right]$ from $\Omega \to \mathcal{M}$, to compute natural-gradients. This is possible because of the following relation,

$$\nabla_{\lambda_z} \mathcal{L} = \left[\nabla_{\lambda_z} \mathbf{m}_z^T\right] \nabla_{m_z} \mathcal{L} = \left[\mathbf{F}_z(\boldsymbol{\lambda}_z)\right] \nabla_{m_z} \mathcal{L} \tag{5}$$

where the first equality is obtained by applying the chain rule and second equality is obtained by noting that $\nabla_{\lambda_z} \mathbf{m}_z^T = \nabla_{\lambda_z}^2 A(\boldsymbol{\lambda}_z) = \mathbf{F}_z(\boldsymbol{\lambda}_z)$. When the FIM is invertible, we get a simple update for NGD,

$$\boldsymbol{\lambda}_z \leftarrow \boldsymbol{\lambda}_z + \beta \; \nabla_{m_z} \mathcal{L}(\boldsymbol{\lambda}_z), \tag{6}$$

where the gradient is computed with respect to $\mathbf{m}_z$. When the above gradient is easier to compute than the gradient with respect to $\boldsymbol{\lambda}_z$, NGD admits a simpler form than GD. This is the case for many existing works on NGD for VI.

To rewrite NGD as in (6), we need the FIM to be invertible. For EFs, a sufficient condition for invertibility is to use a *minimal* representation which is defined[3] below using the definition given in Wainwright & Jordan (2008).

Definition 1 (Minimal EF) *A regular EF representation is said to be minimal when there does not exist a nonzero vectors $\boldsymbol{\lambda}$ such that $\langle \boldsymbol{\phi}_z(\mathbf{z}), \boldsymbol{\lambda} \rangle$ is equal to a constant.*

[1] An EF is called regular when Ω is an open set.

[2] We assume $A_z(\boldsymbol{\lambda}_z)$ can be efficiently computed.

[3] A more complete definition is given in Definition 1.3 of Johansen (1979).

This essentially means that they are no linear dependencies in the parameterization of the distribution. When such a nonzero vector exists, we can add/subtract it from $\boldsymbol{\lambda}$ without changing the distribution. Minimality ensures that this never happens and the parametrization $\boldsymbol{\lambda}_z$ is unique and identifiable up to multiplication with a nonsingular affine transformation.

Under a minimal representation, the log-partition function $A_z(\boldsymbol{\lambda}_z)$ is *strictly* convex, implying that the FIM is positive-definite.[4] For other types of representation, like *curved* EFs, the FIM may not be positive-definite. Minimality ensures that the FIM is positive-definite and that NGD is well defined [5]. The NGD update can be then carried out using the expectation parameter which are a one-to-one function, as stated below (Wainwright & Jordan, 2008).

Theorem 1 *The representation* (1) *is minimal if and only if the mapping* $\mathbf{m}_z(\cdot) : \Omega \to \mathcal{M}$ *is one-to-one.*

Unfortunately, minimal EF approximations are not always appropriate. Such approximations, especially unimodal ones, usually yield poor approximations of multimodal posterior distributions. Structured approximations, such as mixtures of EF distributions, are more suitable for approximating multimodal posteriors. Unfortunately, for such approximations, there is no straightforward way to define a minimal EF representation which can be exploited to derive a simple NGD update. For example, a mixture of EF distributions expressed as

$$q(\mathbf{z}) := \int q(\mathbf{z}|\mathbf{w})q(\mathbf{w})d\mathbf{w}, \tag{7}$$

with $q(\mathbf{z}|\mathbf{w})$ as the component and $q(\mathbf{w})$ as the mixing distribution, may not even have an EF form even when both the terms above are in the EF. A famous example is the finite mixture of Gaussians. The conditions under which the FIM of (7) is invertible are also difficult to characterize in general. Due to these reasons, it is difficult to simplify the NGD update for such structured distributions.

In this paper, we propose a new way to derive simple natural-gradients for structured approximations (7) using a minimal conditional representation which we define next.

3. Minimal Conditional-EF Representation

In this section we define a minimal conditional-EF (MCEF) representation of the joint distribution $q(\mathbf{z}, \mathbf{w})$ for structured approximations that take the form (7). Using this, we derive conditions under which the FIM of the joint is invertible, and show that it leads to a simple NGD similar to (6).

We begin with a definition of the conditional EF distribution.

Definition 2 (Conditional EF) *We call the joint distribution* $q(\mathbf{z}, \mathbf{w})$ *defined in* (7) *a conditional EF when its components take the following form:*[6]

$$\begin{aligned} q(\mathbf{z}|\mathbf{w}) &:= h_z(\mathbf{z}, \mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle - A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right], \\ q(\mathbf{w}) &:= h_w(\mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_w(\mathbf{w}), \boldsymbol{\lambda}_w \rangle - A_w(\boldsymbol{\lambda}_w)\right], \end{aligned} \tag{8}$$

with $\boldsymbol{\lambda}_w, \boldsymbol{\phi}_w(\mathbf{w})$ *and* $A_w(\boldsymbol{\lambda}_w)$ *being the natural parameters, sufficient statistics, and log-partition function of* $q(\mathbf{w})$*, and* $\boldsymbol{\lambda}_z, \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})$ *and* $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ *are the same for* $q(\mathbf{z}|\mathbf{w})$*. We denote the set of natural parameters for* $q(\mathbf{z}|\mathbf{w})$ *and* $q(\mathbf{w})$ *by* Ω_z *and* Ω_w *respectively, and assume them to be open.*[7]

Note that the sufficient statistics $\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})$ and log-partition $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ both depend on $\mathbf{w}$, but, conditioned on $\mathbf{w}$, the distribution is parametrized by the natural parameter $\boldsymbol{\lambda}_z$. For a well-defined conditional EF distribution, $q(\mathbf{w})$ is a regular EF distribution and conditioned on $\mathbf{w}$, $q(\mathbf{z}|\mathbf{w})$ is also a regular EF distribution. This is a type of conditional exponential-family distribution (Xing et al., 2002; Liang et al., 2009; Lindsey, 1996; Feigin, 1981) with a special conditional structure.

We are interested in an NGD update that can exploit the FIM of the joint distribution. We denote the set of natural parameters by $\boldsymbol{\lambda} := \{\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w\}$ and the set of valid $\boldsymbol{\lambda}_z$ and $\boldsymbol{\lambda}_w$ by Ω_z and Ω_w respectively. We define the following FIM in the natural-parameter space $\Omega_z \times \Omega_w$ as follows:

$$\mathbf{F}_{wz}(\boldsymbol{\lambda}) := -\mathbb{E}_{q(z,w)}\left[\nabla^2_\lambda \log q(\mathbf{z}, \mathbf{w})\right]. \tag{9}$$

This is the FIM of the joint distribution $q(\mathbf{z}, \mathbf{w})$ which is different from the one for the marginal distribution $q(\mathbf{z})$. Similar to the minimal EF case, our goal now is to find representations where the above FIM is invertible and can be exploited to compute NGD using expectations of the sufficient statistics. The expectation parameters of a CEF can be defined as shown below:

$$\begin{aligned} \mathbf{m}_z(\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w) &:= \mathbb{E}_{q(z|w)q(w)}\left[\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})\right], & (10) \\ \mathbf{m}_w(\boldsymbol{\lambda}_w) &:= \mathbb{E}_{q(w)}\left[\boldsymbol{\phi}_w(\mathbf{w})\right]. & (11) \end{aligned}$$

We denote the ranges of $\mathbf{m}_z$ and $\mathbf{m}_w$ by $\mathcal{M}_z$ and $\mathcal{M}_w$ respectively, and the whole set of expectation parameters

[4] A formal proof can be found in Johansen (1979).

[5] In some cases, when FIM is not invertible, it is still possible to perform NGD by, for example, ignoring the zero eigenvalues or using damping, but the simplification shown in (6) may not be possible.

[6] We assume that $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ and $A_w(\boldsymbol{\lambda}_w)$ can be efficiently computed.

[7] It is possible for Ω_z to be non-open since it is an intersection of all of the valid natural-parameter of $q(\mathbf{z}|\mathbf{w})$ for each $\mathbf{w} \sim q(\mathbf{w})$. The set is open when the set of valid natural-parameters for $q(\mathbf{z}|\mathbf{w})$ conditioned on $\mathbf{w}$ does not depend on $\mathbf{w}$, or when the cardinality of the support of $\mathbf{w}$ is finite. For all the examples given in this paper, the set is an open set.

by $\mathbf{m}(\boldsymbol{\lambda}) := \{\mathbf{m}_z, \mathbf{m}_w\}$. Since $\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})$ and $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ both depend on $\mathbf{w}$, we may or may not be able to perform NGD using these expectation parameters. However, we next show NGD is always possible by restricting to a minimal representation of CEFs.

Below we define a minimal representation for CEF which ensures that the $\mathbf{F}_{wz}(\boldsymbol{\lambda})$ is positive definite, and NGD can be performed using $\mathbf{m}$.

Definition 3 (Minimal Conditional-EF (MCEF)) *A conditional EF defined in Definition 2 is said to have a minimal representation when* $\mathbf{m}_w(\cdot) : \Omega_w \to \mathcal{M}_w$ *and* $\mathbf{m}_z(\cdot, \boldsymbol{\lambda}_w) : \Omega_z \to \mathcal{M}_z$ *are both one-to-one,* $\forall \boldsymbol{\lambda}_w \in \Omega_w$.

In the next section, we will show that all the examples shown in Figure 1 (See “Examples of MCEFs”) have an MCEF representation. Similar to minimal EFs, an MCEF representation implies that $\mathbf{F}_{wz}(\boldsymbol{\lambda})$ is positive-definite and invertible, as stated in the following theorem (see a proof in Appendix A).

Theorem 2 *For an MCEF representation given in Definition 3, the FIM* $\mathbf{F}_{wz}(\boldsymbol{\lambda})$ *given in* (9) *is positive-definite and invertible for all* $\boldsymbol{\lambda} \in \Omega$.

Since the FIM is well-defined, it is reasonable to perform natural-gradient steps in the Riemannian manifold defined by $\mathbf{F}_{wz}(\boldsymbol{\lambda})$. The steps can be taken using an update that takes a simple form, similar to the one shown in (6). As shown in Lemma 5 in Appendix A, similar to (5) for minimal EF, we have the following relationship for MCEFs:

$$\nabla_\lambda \mathcal{L} = \left[\nabla_\lambda \mathbf{m}^T\right] \nabla_m \mathcal{L} = \left[\mathbf{F}_{wz}(\boldsymbol{\lambda})\right] \nabla_m \mathcal{L} \qquad (12)$$

Since the FIM is invertible, we can compute the natural-gradient by using gradients with respect to $\mathbf{m}$. The following theorem shows the simplicity of NGD.

Theorem 3 *For minimal conditional-EF approximations, the following updates are equivalent:*

$$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \beta \ \left[F_{wz}(\boldsymbol{\lambda})\right]^{-1} \nabla_\lambda \mathcal{L}(\boldsymbol{\lambda}) \qquad (13)$$

$$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \beta \ \nabla_m \mathcal{L}(\boldsymbol{\lambda}). \qquad (14)$$

The above theorem generalizes the previous result for minimal EF approximations to minimal conditional-EF approximations. It also implies that using the expectation parameterization $\mathbf{m}$ enables us to exploit the geometry of the joint $q(\mathbf{z}, \mathbf{w})$ to improve convergence.

4. Examples

In this section, we give examples of approximations with a minimal conditional-EF form. We discuss finite mixtures of EFs, scale mixtures of Gaussians, and multivariate Skew-Gaussians. We give updates that can be implemented efficiently by using existing NGD implementations, e.g., the variational online Newton (VON) method of Khan et al. (2018). We also propose many new versions. To derive these updates, we use the extended Bonnet’s and Price’s theorems (Lin et al., 2019) for Gaussian mixtures. For exponential-family mixtures, we use the implicit reparameterization trick (Salimans & Knowles, 2013; Figurnov et al., 2018). Lin et al. (2019) discuss a weaker version of the reparameterization trick for exponential-family mixtures.

4.1. Finite Mixture of Exponential Family Distribution

Finite mixtures of EFs are a powerful approximation where components in EF form are mixed using a discrete distributions such as a multinomial distribution, as shown below:

$$q(\mathbf{z}) = \sum_{c=1}^{K} \pi_c q(\mathbf{z}|\boldsymbol{\lambda}_c), \text{ such that } \sum_{c=1}^{K} \pi_c = 1, \qquad (15)$$

where $q(\mathbf{z}|\boldsymbol{\lambda}_c)$ are EF distributions with natural parameters $\boldsymbol{\lambda}_c$. The support of $\mathbf{z}$ of all the components is assumed to be the same, and π_K is fixed to $1 - \sum_{c=1}^{K-1} \pi_c$.

This distribution cannot be written in an exponential form in general, and therefore none of the existing NGD methods can be used to derive a simple expression for NGVI. Directly applying NGD using the FIM of $q(\mathbf{z})$ would be too expensive since the number of parameters are in $O(D^2K)$, and the FIM could be extremely large when computed naively. Fortunately, the joint distribution does take an MCEF form when all $q(\mathbf{z}|\boldsymbol{\lambda}_c)$ are minimal EFs. A formal statement is given in Appendix B. Using this conditional EF form, we can derive a much simpler NGD update which reduces to simple parallel updates on mixture components.

We will now demonstrate the simplicity of our update for a finite mixture of Gaussians (MOG) where components $q(\mathbf{z}|\boldsymbol{\lambda}_c) := \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ are Gaussian. The sufficient statistics, natural parameters, and expectation parameters of the corresponding CEF are given in Table 1.

We consider the following model where the likelihood $p(\mathcal{D}_n|\mathbf{z})$ is defined using neural-network weights $\mathbf{z}$ with a prior $p(\mathbf{z})$: $p(\mathcal{D}, \mathbf{z}) = \prod_{n=1}^{N} p(\mathcal{D}_n|\mathbf{z})p(\mathbf{z})$. We approximate the posterior with a MOG by optimizing the following variational lower bound: $\mathcal{L}(\boldsymbol{\lambda}) := \mathbb{E}_{q(z)}[-h(\mathbf{z})]$ where $h(\mathbf{z}) := \log\left[q(\mathbf{z})/p(\mathbf{z})\right] - \sum_n \log p(\mathcal{D}_n|\mathbf{z})$.

We now summarize the NGD update derived in Appendix B. We first generate samples $(\mathbf{z}, w)$ from $q(\mathbf{z}, w)$. The mean $\boldsymbol{\mu}_c$ and covariance $\boldsymbol{\Sigma}_c$ are similarly updated by the variational online Newton (VON) algorithm (Khan et al., 2018):

$$\begin{aligned} \boldsymbol{\Sigma}_c^{-1} &\leftarrow \boldsymbol{\Sigma}_c^{-1} + \beta\delta_c \left[\nabla_z^2 h(\mathbf{z})\right], \\ \boldsymbol{\mu}_c &\leftarrow \boldsymbol{\mu}_c - \beta\delta_c \boldsymbol{\Sigma}_c \left[\nabla_z h(\mathbf{z})\right], \\ \log\left(\pi_c/\pi_K\right) &\leftarrow \log\left(\pi_c/\pi_K\right) - \beta(\delta_c - \delta_K) h(\mathbf{z}) \end{aligned} \qquad (16)$$

Table 1. We give expressions for various components of $q(\mathbf{z}, \mathbf{w})$ as specified in (8) for a minimal conditional-EF representation. We give three examples. More examples can be found at the Appendix. The first two rows give expressions for the sufficient statistics, and the subsequent rows show natural and expectation parameters. For the first column, $\mathbb{I}_c(w)$ denotes the indicator function which is 1 when $w = c$ and 0 otherwise. For the second column, ψ is the digamma function. For the third column, $c := \sqrt{2/\pi}$. Note that both the natural and expectation parameters may lie in a constrained set. Due to space limitations, we have not explicitly given the description of these sets.

	Mixture of Gaussians	T-Distribution	Skew-Gaussian
$\phi_w(w)$	$\{\mathbb{I}_c(w)\}_{c=1}^{K-1}$	$-1/w - \log w$	$w,\ w^2$
$\boldsymbol{\phi}_z(\mathbf{z}, w)$	$\left\{\mathbb{I}_c(w)\mathbf{z},\ \mathbb{I}_c(w)\mathbf{z}\mathbf{z}^T\right\}_{c=1}^{K}$	$\left\{\mathbf{z}/w,\ \mathbf{z}\mathbf{z}^T/w\right\}$	$\left\{\mathbf{z},\ \|w\|\mathbf{z},\ \mathbf{z}\mathbf{z}^T\right\}$
$\boldsymbol{\lambda}_w$	$\{\log(\pi_c/\pi_K)\}_{c=1}^{K-1}$	a	constants $\left\{0,\ -\frac{1}{2}\right\}$
$\mathbf{m}_w(\boldsymbol{\lambda}_w)$	$\{\pi_c\}_{c=1}^{K-1}$	$-1 - \log a + \psi(a)$	constants $\{0,\ 1\}$
$\boldsymbol{\lambda}_z$	$\left\{\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c,\ -\frac{1}{2}\boldsymbol{\Sigma}_c^{-1}\right\}_{c=1}^{K}$	$\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\right\}$	$\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha},\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\right\}$
$\mathbf{m}_z(\boldsymbol{\lambda})$	$\left\{\pi_c\boldsymbol{\mu}_c, \pi_c\left(\boldsymbol{\mu}_c\boldsymbol{\mu}_c^T + \boldsymbol{\Sigma}_c\right)\right\}_{c=1}^{K}$	$\left\{\boldsymbol{\mu}, \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}\right\}$	$\{\boldsymbol{\mu} + c\boldsymbol{\alpha},\ \boldsymbol{\alpha} + c\boldsymbol{\mu},$ $\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\alpha}\boldsymbol{\alpha}^T + \boldsymbol{\Sigma} + c\left(\boldsymbol{\mu}\boldsymbol{\alpha}^T + \boldsymbol{\alpha}\boldsymbol{\mu}^T\right)\}$

where $\delta_c := \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)/\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. This update can be implemented efficiently using approximations discussed in Khan et al. (2018); Zhang et al. (2018); Mishkin et al. (2018). For example, implementation of the Adam optimizer can be utilized. This is simpler and more efficient than the update which requires computation of the FIM.

A similar example to MOG is the fatigue life distribution (Birnbaum & Saunders, 1969) discussed in Appendix C.

4.2. Scale Mixture of Gaussians

A multivariate scale-mixture of Gaussian (SMG) distribution (Andrews & Mallows, 1974; Eltoft et al., 2006) takes the following form where the covariance matrix is "scaled" by a vector $\mathbf{w}$ sampled from $q(\mathbf{w})$ such that $w_i > 0$:

$$q(\mathbf{z}) = \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \mathbf{L}\mathbf{W}\mathbf{L}^T)\prod_{i=1}^{d} q(w_i)d\mathbf{w},$$

where $\mathbf{W} := \mathrm{diag}(\mathbf{w})$ is a diagonal matrix containing $\mathbf{w}$ as the diagonal and $\mathbf{L}$ is a matrix with determinant 1 (e.g., a Cholesky factor) that determines the covariance matrix. SMG includes many well-known distributions like Student's t, Laplace, logistic, doubly-exponential, normal-gamma, normal-inverse Gaussian, normal-Jeffreys, and their non-parametric extensions (Caron & Doucet, 2008). For example, the multivariate Student's t-distribution is obtained by using a scalar w from an inverse-gamma distribution:

$$\mathcal{T}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, a) := \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\ \mathcal{IG}(w|a, a)dw, \quad (17)$$

where $\boldsymbol{\Sigma} := \mathbf{L}\mathbf{L}^T$ and both the shape and scale parameters of the inverse-gamma distribution are equal to a. We assume $a > 1$, since for $a \leq 1$ the variance of $q(\mathbf{z})$ does not exist.

SMG is another class of approximations where existing methods cannot be applied to obtain a simple natural-gradient update. For example, for Student's t the joint-distribution $q(\mathbf{z}, w)$ takes an EF form, but, as we show in Lemma 8 in Appendix D, the number of free parameters is equal to 3, while the number of natural parameters is equal to 4. Such representations are not minimal EFs but *curved* where the FIM is not positive-definite because the number of free parameters is less than the number of natural parameters. Therefore, the update (6) does not apply. In contrast, our update can be applied since the joint distribution is a minimal CEF. This can be verified from Table 1 where both $m_w(\lambda_w)$ and $\mathbf{m}_z(\boldsymbol{\lambda})$ are one-to-one functions. A formal proof is given in Lemma 10 in Appendix D.

We now demonstrate the simplicity of our update to obtain a Student's t-approximation on a Bayesian neural network. We assume the following model with a likelihood $p(\mathcal{D}_n|\mathbf{z})$ specified using a neural network with weights $\mathbf{z}$ and a Student's t-prior: $p(\mathcal{D}, \mathbf{z}) = \prod_{n=1}^{N} p(\mathcal{D}_n|\mathbf{z})\mathcal{T}(\mathbf{z}|\mathbf{0}, \mathbf{I}, a_0)$. The Student's t-prior is better than a Gaussian one if we expect the weights to follow a heavy-tailed distribution. We approximate the posterior by the approximation (17) using parameters $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}$, and a. We use the variational lower bound defined in the joint distribution $p(\mathcal{D}, \mathbf{z}, w)$ space:

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(z,w|\lambda)}\Big[\sum_{n=1}^{N}\overbrace{\log p(\mathcal{D}_n|\mathbf{z})}^{:=-f_n(\boldsymbol{z})} + \log\frac{p(\mathbf{z}, w)}{q(\mathbf{z}, w)}\Big] \quad (18)$$

Below, we summarize the NGD updates derived in Appendix D. We first sample $(\mathbf{z}, w)$ from q and randomly sampled an example n. The update is then a small modification of the VON update (Khan et al., 2018):

$$\boldsymbol{\Sigma}^{-1} \leftarrow (1-\beta)\boldsymbol{\Sigma}^{-1} + \beta\left[u\nabla_z^2 f_n(\mathbf{z}) + \mathbf{I}/N\right], \quad (19)$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta\boldsymbol{\Sigma}\left[\nabla_z f_n(\mathbf{z}) + \boldsymbol{\mu}/N\right], \quad (20)$$

$$a \leftarrow (1-\beta)a + \beta\left[a_0 - \delta\mathrm{Tr}\left(\nabla_z^2 f_n(\mathbf{z})\boldsymbol{\Sigma}\right)\right], \quad (21)$$

where $\beta > 0$, $\delta \leftarrow Nw^2/(2(1-w))$, and u is a pre-multiplier defined below (d is length of $\mathbf{z}$): $u = (a - 1 + d/2)^{-1}\left[a + (\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})/2\right]$. Similarly to the previous section, this update can be implemented efficiently using the method of Khan et al. (2018).

Another example is given at Appendix H. Extensions using an Adam-like optimizer for this kind of mixtures are given in Appendix J.

4.3. Gaussian Mean Mixture

We consider the following Gaussian mixture.

$$q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + \sum_{i=1}^{k} u(w_i)\boldsymbol{\alpha}_i, \boldsymbol{\Sigma}) \prod_{i=1}^{k} q(w_i) d\mathbf{w},$$

where $\boldsymbol{\alpha}$ is a d-by-k matrix and $\mathbf{z} \in \mathcal{R}^d$. An example is the rank-1 covariance Gaussian with $k = 1$, $u(w) = w$, and $\mathbf{D}$ as a diagonal covariance matrix: $q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\mathbf{D}) = \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + w\boldsymbol{\alpha}, \mathbf{D})\mathcal{N}(w|0,1)dw = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}\boldsymbol{\alpha}^T + \mathbf{D})$.

The multivariate skew Gaussian (Azzalini & Valle, 1996; Genton, 2004), defined below, is another example and allows for non-zero skewness (asymmetric approximations):

$$q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + |w|\boldsymbol{\alpha}, \boldsymbol{\Sigma})\mathcal{N}(w|0,1)dw.$$

This distribution is not a minimal EF and the FIM of the marginal $q(\mathbf{z})$ can be singular (Azzalini, 2013). However, the joint distribution is a minimal conditional-EF distribution as we show in Lemma 12 and 13 in Appendix E. The sufficient statistics, natural parameters, expectation parameters of the conditional EF form are given in Table 1.

Similar to other examples, we get a simple and efficient NGD update. We summarize the updates for a model with a neural-network likelihood $p(\mathcal{D}_n|\mathbf{z})$ using weights $\mathbf{z}$ and a Gaussian prior $\mathcal{N}(\mathbf{z}|0,\mathbf{I})$. Denoting $f_n(\mathbf{z}) = -\log p(\mathcal{D}_n|\mathbf{z})$, the lower bound is $\mathcal{L}(\boldsymbol{\lambda}) := \mathbb{E}_{q(z)}[-\sum_n f_n(\mathbf{z}) + \log p(\mathbf{z}) - \log q(\mathbf{z})]$. The updates derived in Appendix E is a variant of the VON update:

$$\boldsymbol{\Sigma}^{-1} \leftarrow (1-\beta)\boldsymbol{\Sigma}^{-1} + \beta(\mathbf{I} + N\mathbf{g}_S^n) \qquad (22)$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta\boldsymbol{\Sigma}\left[\bar{c}(\mathbf{g}_\mu^n - c\mathbf{g}_\alpha^n) + \boldsymbol{\mu}\right] \qquad (23)$$

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \beta\boldsymbol{\Sigma}\left[\bar{c}(\mathbf{g}_\alpha^n - c\mathbf{g}_\mu^n) + \boldsymbol{\alpha}\right] \qquad (24)$$

where $\bar{c} := N/(1-2/\pi)$ and $\mathbf{g}_S^n, \mathbf{g}_\mu^n, \mathbf{g}_\alpha^n$ are gradients obtained by gradient and Hessian of $f_n(\mathbf{z})$ at a sample of $q(\mathbf{z})$. The gradients are defined in (53)-(55) in Appendix E.

Another example of the mixture is the exponentially modified Gaussian distribution (Grushka, 1972; Carr & Madan, 2009) given in Appendix F. Extensions using an Adam optimizer for the class of mixtures are given in Appendix J.

5. Generalization to Multilinear EF

We now extend the approach to an approximation with multilinear EFs which contain blocks of natural parameters. We start by specifying a distribution over $\mathbf{z}$ by a function $f(\cdot)$:

$$q(\mathbf{z}|\boldsymbol{\lambda}) = h_z(\mathbf{z}) \exp\left[f(\mathbf{z},\boldsymbol{\lambda}) - A_z(\boldsymbol{\lambda})\right]. \qquad (25)$$

Then we divide the vector $\boldsymbol{\lambda} := \{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_K\}$ into K blocks with $\boldsymbol{\lambda}_j \in \Omega_j$ being the j-th block of parameters. In EF, the function f is just linear in $\boldsymbol{\lambda}_z$ at (1). We can generalize the notion of linearity to multiple blocks of parameters by considering f to be a *multilinear* function.

Definition 4 (Minimal Multilinear-EF) *We call f a multilinear function when, for each block j, there exist functions $\boldsymbol{\phi}_j$ and r_j such that f is linear with respect to $\boldsymbol{\lambda}_j$, i.e.,*

$$f(\mathbf{z},\boldsymbol{\lambda}) = \langle \boldsymbol{\lambda}_j, \boldsymbol{\phi}_j(\mathbf{z},\boldsymbol{\lambda}_{-j})\rangle + r_j(\mathbf{z},\boldsymbol{\lambda}_{-j}), \qquad (26)$$

where $\boldsymbol{\lambda}_{-j}$ is the parameter vector containing all $\boldsymbol{\lambda}$ except $\boldsymbol{\lambda}_j$. A distribution $q(\mathbf{z}|\boldsymbol{\lambda})$ defined as in (25)*, but with a multilinear f, is called a multilinear EF. Additionally, when Ω_j is open and the following expectation parameters $\mathbf{m}_j(\boldsymbol{\lambda}) := \mathbb{E}_{q(z)}\left[\boldsymbol{\phi}_j(\mathbf{z},\boldsymbol{\lambda}_{-j})\right]$ are one-to-one, we call the distribution a minimal multilinear EF distribution.*

Clearly, minimal EFs are minimal multilinear EFs. The following theorem give a result about a *block* NGD update performed on individual blocks of parameters $\boldsymbol{\lambda}_j$.

Theorem 4 *For approximations with multilinearly-minimal EF representation, the following updates are equivalent:*

$$\boldsymbol{\lambda}_j \leftarrow \boldsymbol{\lambda}_j + \beta \left[\mathbf{F}_j(\boldsymbol{\lambda})\right]^{-1} \nabla_{\lambda_j}\mathcal{L}(\boldsymbol{\lambda}) \qquad (27)$$

$$\boldsymbol{\lambda}_j \leftarrow \boldsymbol{\lambda}_j + \beta \nabla_{m_j}\mathcal{L}(\boldsymbol{\lambda}) \qquad (28)$$

The proof of this theorem is similar to Theorem 3. We now give an example and demonstrate the simplicity of the NGD update. Let's consider the Matrix-Variate Gaussian (MVG) distribution defined as follows:

$$\mathcal{MN}(\mathbf{Z}|\mathbf{W},\mathbf{U},\mathbf{V}) := \mathcal{N}(\text{vec}(\mathbf{Z})|\text{vec}(\mathbf{W}), \mathbf{U}\otimes\mathbf{V}).$$

This distribution has been used for Bayesian neural networks (Louizos & Welling, 2016; Sun et al., 2017). An approximate NGD update is also derived by Zhang et al. (2018) where the FIM is approximated by a block-diagonal matrix and K-FAC approximation. Our update have a similar block-diagonal approximation, but the update for the each block is an exact NGD unlike Zhang et al. (2018) where the steps are approximated by K-FAC.

In Appendix I, we show that the MVG distribution $\mathcal{MN}(\mathbf{Z}|\mathbf{W},\mathbf{U},\mathbf{V})$ can be written in the minimal multilinear form. The NGD update, derived in Appendix I, is summarized below to optimize the lower bound as $\mathcal{L}(\boldsymbol{\lambda}) = E_q[-h(\mathbf{Z})]$ where $h(\mathbf{Z}) := -\log p(\mathcal{D},\mathbf{Z}) + \log q(\mathbf{Z})$. To simplify our implementation, we use the Gauss-Newton approximation (Graves, 2011) although it is not necessary to do so. The resulting block NGD update is shown below,

$$\mathbf{W} \leftarrow \mathbf{W} - \beta_1\mathbf{U}\mathbf{G}\mathbf{V}, \qquad (29)$$

$$\mathbf{U}^{-1} \leftarrow \mathbf{U}^{-1} + \beta_2\mathbf{G}\mathbf{V}\mathbf{G}^\top, \qquad (30)$$

$$\mathbf{V}^{-1} \leftarrow \mathbf{V}^{-1} + \beta_2\mathbf{G}^\top\mathbf{U}\mathbf{G}, \qquad (31)$$

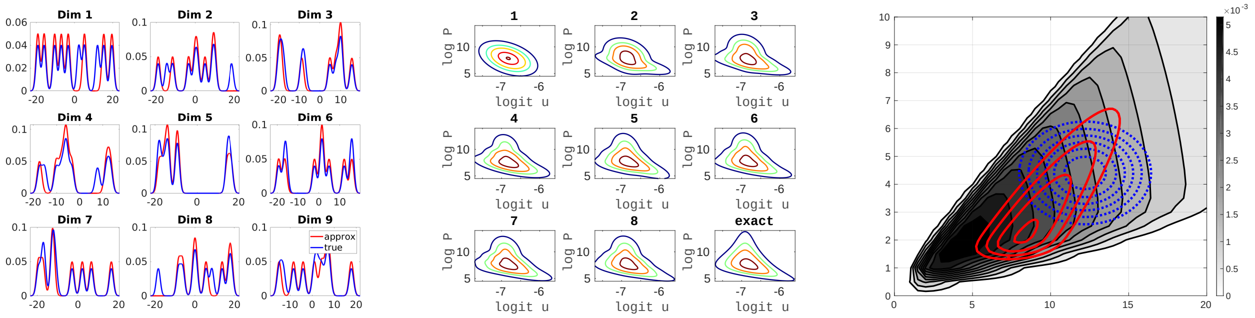

Figure 2. Quantitative results on three toy examples showing the flexibility obtained by using structured approximations considered in this paper. The leftmost figure shows the MOG approximation (with $K = 20$) to fit an MOG model with 10 components in a 20 dimensional problem. The first 9 dimensions are shown in the figure where we see that MOG approximation fits the marginals well. The middle figure shows MOG approximation fit to a beta-binomial model for a 2-D problem. The number indicates the number of mixture components used. By increasing the number of components in our approximation, we get better results. The last figure shows a Skew Gaussian (in red) and a Gaussian (in blue) fit on a 2-D logistic regression posterior. We see that skew-Gaussian captures the skewness in the distribution in the right direction, and gives better approximation than a single Gaussian.

where we sample $\mathbf{Z}$ from the MVG distribution and evaluate the gradient $\mathbf{G} := \nabla_Z h(\mathbf{Z})$. These updates extend the VON update obtained in Khan et al. (2018) to MVG approximations. The gradient $\mathbf{G}$ is pre-conditioned, which is very similar to other preconditioned algorithms, such as K-FAC (Martens & Grosse, 2015; Zhang et al., 2018) and Shampoo (Gupta et al., 2018). The update can be extended to Tensor-Variate Gaussian (Ohlson et al., 2013).

6. Experimental Results

The code is available at:
`https://github.com/yorkerlin/VB-MixEF`.

6.1. Qualitative Results on Synthetic Examples

First, we show qualitative results on three toy examples and visualize the results obtained by structured approximations.

The first toy example is the Gaussian mixture example from Wang et al. (2018). In this example, the true distribution is a finite mixture of Gaussians (MOG) $p(\mathbf{z}) = \sum_{i=1}^{C} \frac{1}{C}\mathcal{N}(\mathbf{z}|\mathbf{u}_i, \mathbf{I}), \mathbf{z} \in \mathcal{R}^d$, where each element of $\mathbf{u}_i$ is uniformly drawn from the interval $[-s, s]$. We approximate the posterior distribution by an MOG approximation described in Section 4.1. We consider a case with $K = 20$, $C = 10$, $s = 20$, and $d = 20$. We initialize $\pi_c = \frac{1}{K}$ and $\mathbf{\Sigma}_c = 100\mathbf{I}$. Each element of $\boldsymbol{\mu}_c$ is randomly initialized by Gaussian noise with mean 0 and variance 100. We use 10 Monte Carlo samples to compute the gradients. The leftmost plot in Figure 2 shows the first 9 marginal distributions of the true distribution and its approximations, where we clearly see that MOG closely matches the marginals. All 20 marginal distributions are in Figure 6 in Appendix B.5.

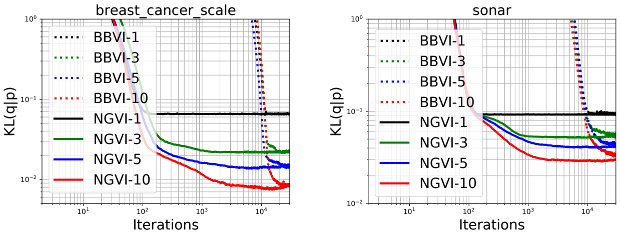

Figure 3. Bayesian logistic regression approximated by MoG: This figure demonstrates a fast convergence of NGVI over BBVI. We use a mixture of Gaussians with full covariance matrix as the approximating distribution. The number next to the method name indicates the number of mixture components used. The plot shows the KL obtained using 10^6 MC samples, where p is the true posterior distribution. For both algorithms, we used full-batches by using 20 MC samples to compute stochastic approximations.

In the second toy example we approximate the beta-binomial model for overdispersion considered in Salimans & Knowles (2013); Salimans et al. (2015) by using MOG ($N = 20$, $d = 2$). The model is to used to estimate the rates of death from stomach cancer for cities in Missouri. The exact posterior of the model is non-standard and extremely skewed. In the middle plot in Figure 2, we see that our MOG approximation approximates the true posterior better and better as we increase the number of mixture components.

In the last toy example, we visualize the skew-Gaussian approximations for the two-dimensional Bayesian Logistic regression example taken from Murphy (2013) ($N = 60$, $d = 2$). In the rightmost plot in Figure 2, we can see that the skewness of the true posterior is captured well by the skew-Gaussian distribution. The Gaussian approximation results in a worse approximation than the skew-Gaussian.

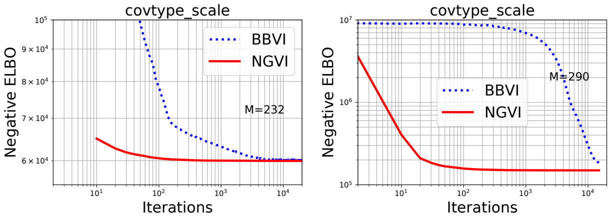

Figure 4. Bayesian logistic regression with Student's t (left) and skew-Gaussian approximations (right). We use 10 MC samples for training, and M denotes the mini-batch size.

6.2. Results on Real Data

Next, we show results on real-world datasets. We consider two models in our experiments. We start with Bayesian Logistic regression (BLR) and present results for MOG approximations on two small UCI datasets. The 'Breast-Cancer' dataset has $N = 683, d = 10$ with 341 chosen for training, and regularization parameter is set to 1.0. The 'Sonar' dataset has $N = 208, d = 6$ with 100 chosen for training, and regularization parameter is set to 0.204. We vary the number of mixture components to $K = 1, 3, 5, 10$. In Figure 3, we plot the KL divergence between the true posterior and the MOG approximation, and compare our method (referred to as 'NGVI') proposed in Section 4.1 to the black-box gradient method (referred to as 'BBVI') with the re-parametrization trick (Salimans & Knowles, 2013; Kingma & Welling, 2013; Figurnov et al., 2018). For both methods, we use a full batch for each update. We observe that NGVI always converges faster than BBVI. We also see that MoG is a better approximation than the single Gaussian, and the quality of the posterior approximation improves as the number of mixture increases.

Next, we show results on a larger dataset using two other kinds of variational approximation: Skew-Gaussian and Student's t. We use the UCI dataset "covtype-binary-scale" with $d = 54, N = 581,012$ with $464,809$ chosen for training and regularization parameter 0.002. We use the algorithm discussed in Section 4.2 and 4.3. For black-box methods, we use the Adam optimizer and refer to it as BBVI. In the skew-Gaussian case, we use a Gaussian prior, and, for the Student's t-distribution, we use a Student's t-prior as shown in (18). Figure 4 demonstrates the fast convergence of our method compared to BBVI.

Finally, we discuss results on Bayesian neural network (BNN) with a standard normal prior on weights. We use one hidden layer, 50 hidden units, and ReLU activation functions. We approximate the posterior by a skew-Gaussian distribution using NGVI. We also compare to two other methods where we used BBVI to fit a skew-Gaussian approximation as well as a Gaussian approximation. For scalability reasons, we use of a diagonal covariance for all methods. We use 10 Monte Carlo (MC) samples and mini-batch size of 32. For NGVI, we use the gradient magnitude-approximation as explained in Appendix J. Figure 5 shows the performance of all methods in terms of the test RMSE. We can see that our method converges faster than BBVI, although the performance of skew-Gaussian methods seem to be similar to a Gaussian.

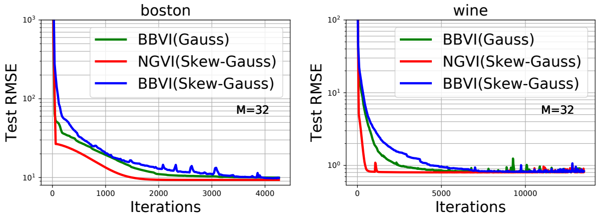

Figure 5. BNN using skew-Gaussian approximation: This figure shows a fast convergence of NGVI over BBVI to approximate the posterior of BNN. For all methods, the prior is a Gaussian. We use 10 MC samples for training. M in the figures denotes the size of a mini-batch. BBVI (Gauss) uses a Gaussian approximation while BBVI (Skew-Gauss) uses a skew-Gaussian approximation. Our NGVI method with skew-Gaussian approximation converges faster than the other two methods.

7. Discussion

In this paper, we present fast and simple NGD for VI with structured approximations. The approximations we have considered are currently beyond the reach of existing methods, and our approach extends these existing approaches to perform NGD updates with a simple update which can also be implemented efficiently in some cases. Our current proposal is limited to a certain class of approximations, and further work is needed to generalize our results to many other types of structured approximations. The minimality condition we proposed uses one-to-one mappings of the expectation parameterization. We believe that this condition can be relaxed which will enable simple NGD update for many types of approximations.

Our main focus has been on the derivation of simple updates. We have presented examples where the updates can also be implemented efficiently. There are however implementation bottlenecks in existing software frameworks to implement some of the reparameterization tricks used in our algorithms. It is important to find ways to enable efficient updates by modifying the existing software frameworks. Another issue is that the NGD update needs to make sure that the parameters stay inside Ω, and this issue deserves further exploration. Another important direction is to apply structured approximations to large problems, especially those involving deep networks. We hope to perform such extensive experiments in the future to establish the benefits obtained by our NGD update for Bayesian deep learning.

Acknowledgements

We would like to thank Voot Tangkaratt for useful discussions during this work. WL is supported by a UBC International Doctoral Fellowship.

References

Andrews, D. F. and Mallows, C. L. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 99–102, 1974.

Arellano-Valle, R. B., Contreras-Reyes, J. E., and Genton, M. G. Shannon Entropy and Mutual Information for Multivariate Skew-Elliptical Distributions. *Scandinavian Journal of Statistics*, 40(1):42–62, 2013.

Azzalini, A. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32 (2):159–188, 2005.

Azzalini, A. *The skew-normal and related families*, volume 3. Cambridge University Press, 2013.

Azzalini, A. and Valle, A. D. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

Balakrishnan, N. and Kundu, D. Birnbaum-Saunders distribution: A review of models, analysis, and applications. *Applied Stochastic Models in Business and Industry*, 35 (1):4–49, 2019.

Barndorff-Nielsen, O. E. Normal inverse Gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of statistics*, 24(1):1–13, 1997.

Batir, N. Some new inequalities for gamma and polygamma functions. *J. Inequal. Pure Appl. Math*, 6(4):1–9, 2005.

Birnbaum, Z. W. and Saunders, S. C. A new family of life distributions. *Journal of applied probability*, 6(2): 319–327, 1969.

Bonnet, G. Transformations des signaux aléatoires a travers les systemes non linéaires sans mémoire. In *Annales des Télécommunications*, volume 19, pp. 203–220. Springer, 1964.

Caron, F. and Doucet, A. Sparse Bayesian nonparametric regression. In *Proceedings of the 25th international conference on Machine learning*, pp. 88–95. ACM, 2008.

Carr, P. and Madan, D. Saddlepoint methods for option pricing. *The Journal of Computational Finance*, 13(1): 49, 2009.

Choi, J., Du, Y., and Song, Q. Inverse Gaussian quadrature and finite normal-mixture approximation of generalized hyperbolic distribution. *arXiv preprint arXiv:1810.01116*, 2018.

Contreras-Reyes, J. E. and Arellano-Valle, R. B. Kullback–Leibler divergence measure for multivariate skew-normal distributions. *Entropy*, 14(9):1606–1626, 2012.

Culham, R. Lecture Notes: Advance differential equations and special functions. `www.mhtlab.uwaterloo.ca/courses/me755/web_chap4.pdf`, 2004. Accessed: 2019/03/25.

Desmond, A. On the relationship between two fatigue-life models. *IEEE Transactions on Reliability*, 35(2):167–169, 1986.

Eltoft, T., Kim, T., and Lee, T.-W. Multivariate scale mixture of Gaussians modeling. In *International Conference on Independent Component Analysis and Signal Separation*, pp. 799–806. Springer, 2006.

Feigin, P. D. Conditional Exponential Families and a Representation Theorem for Asympotic Inference. *The Annals of Statistics*, pp. 597–603, 1981.

Figurnov, M., Mohamed, S., and Mnih, A. Implicit Reparameterization Gradients. 2018.

Furmston, T. and Barber, D. Variational methods for reinforcement learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 241–248, 2010.

Genton, M. G. *Skew-elliptical distributions and their applications: a journey beyond normality*. CRC Press, 2004.

Graves, A. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pp. 2348–2356, 2011.

Grushka, E. Characterization of exponentially modified Gaussian peaks in chromatography. *Analytical Chemistry*, 44(11):1733–1738, 1972.

Gupta, V., Koren, T., and Singer, Y. Shampoo: Preconditioned Stochastic Tensor Optimization. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1842–1850, 2018.

Hensman, J., Rattray, M., and Lawrence, N. D. Fast variational inference in the conjugate exponential family. In *Advances in neural information processing systems*, pp. 2888–2896, 2012.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

Hoffman, M. D. and Blei, D. M. Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, 2015.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Honkela, A., Tornio, M., Raiko, T., and Karhunen, J. Natural conjugate gradient in variational inference. In *International Conference on Neural Information Processing*, pp. 305–314. Springer, 2007.

Honkela, A., Raiko, T., Kuusela, M., Tornio, M., and Karhunen, J. Approximate Riemannian Conjugate Gradient Learning for Fixed-form Variational Bayes. *Journal of Machine Learning Research (JMLR))*, 11:3235–3268, 2011.

Johansen, S. Introduction to the Theory of Regular Exponential Famelies. 1979.

Jørgensen, B., Seshadri, V., and Whitmore, G. On the mixture of the inverse Gaussian distribution with its complementary reciprocal. *Scandinavian Journal of Statistics*, pp. 77–89, 1991.

Khan, M. and Lin, W. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *Artificial Intelligence and Statistics*, pp. 878–887, 2017.

Khan, M. E. and Nielsen, D. Fast yet Simple Natural-Gradient Descent for Variational Inference in Complex Models. *arXiv preprint arXiv:1807.04489*, 2018.

Khan, M. E., Babanezhad, R., Lin, W., Schmidt, M., and Sugiyama, M. Faster stochastic variational inference using Proximal-Gradient methods with general divergence functions. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 319–328. AUAI Press, 2016.

Khan, M. E., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2611–2620, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kotz, S. and Nadarajah, S. *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

Kumar, S. and Tsvetkov, Y. Von Mises-Fisher Loss for Training Sequence to Sequence Models with Continuous Outputs. *arXiv preprint arXiv:1812.04616*, 2018.

Liang, P., Jordan, M. I., and Klein, D. Learning from measurements in exponential families. In *Proceedings of the 26th annual international conference on machine learning*, pp. 641–648. ACM, 2009.

Lin, W., Khan, M. E., and Schmidt, M. Stein's Lemma for the Reparameterization Trick with Exponential-family Mixtures. `https://github.com/yorkerlin/VB-MixEF/blob/master/report.pdf`, 2019. Accessed: 2019/06.

Lindsey, J. K. *Parametric statistical inference*. Oxford University Press, 1996.

Louizos, C. and Welling, M. Structured and efficient variational deep learning with matrix gaussian posteriors. In *International Conference on Machine Learning*, pp. 1708–1716, 2016.

Martens, J. and Grosse, R. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pp. 2408–2417, 2015.

Mishkin, A., Kunstner, F., Nielsen, D., Schmidt, M., and Khan, M. E. SLANG: Fast Structured Covariance Approximations for Bayesian Deep Learning with Natural Gradient. In *Advances in Neural Information Processing Systems*, pp. 6246–6256, 2018.

Murphy, K. P. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013. ISBN 9780262018029 0262018020.

Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. Variational Continual Learning. *arXiv preprint arXiv:1710.10628*, 2017.

Oh, C., Adamczewski, K., and Park, M. Radial and Directional Posteriors for Bayesian Neural Networks. *arXiv preprint arXiv:1902.02603*, 2019.

Ohlson, M., Ahmad, M. R., and Von Rosen, D. The multilinear normal distribution: Introduction and some basic properties. *Journal of Multivariate Analysis*, 113:37–47, 2013.

Opper, M. and Archambeau, C. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

Price, R. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958.

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Artificial Intelligence and Statistics*, pp. 814–822, 2014.

Ranganath, R., Tran, D., and Blei, D. Hierarchical variational models. In *International Conference on Machine Learning*, pp. 324–333, 2016.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Ruiz-Antolín, D. and Segura, J. A new type of sharp bounds for ratios of modified Bessel functions. *Journal of Mathematical Analysis and Applications*, 443(2):1232–1246, 2016.

Salimans, T. and Knowles, D. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

Salimans, T., Kingma, D., and Welling, M. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pp. 1218–1226, 2015.

Salimbeni, H., Eleftheriadis, S., and Hensman, J. Natural Gradients in Practice: Non-Conjugate Variational Inference in Gaussian Process Models. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Sato, M.-A. Online model selection based on the variational Bayes. *Neural computation*, 13(7):1649–1681, 2001.

Staines, J. and Barber, D. Variational optimization. *arXiv preprint arXiv:1212.4507*, 2012.

Sun, S., Chen, C., and Carin, L. Learning Structured Weight Uncertainty in Bayesian Neural Networks. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, pp. 1283–1292, 2017.

Titsias, M. and Lázaro-Gredilla, M. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, pp. 1971–1979, 2014.

Titsias, M. K. and Ruiz, F. J. Unbiased Implicit Variational Inference. *arXiv preprint arXiv:1808.02078*, 2018.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1–2:1–305, 2008.

Wang, D., Liu, H., and Liu, Q. Variational Inference with Tail-adaptive f-Divergence. In *Advances in Neural Information Processing Systems*, pp. 5742–5752, 2018.

Xing, E. P., Jordan, M. I., and Russell, S. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pp. 583–591. Morgan Kaufmann Publishers Inc., 2002.

Yang, Z.-H. and Chu, Y.-M. On approximating the modified Bessel function of the second kind. *Journal of Inequalities and Applications*, 2017(1):41, 2017.

Yin, M. and Zhou, M. Semi-Implicit Variational Inference. *arXiv preprint arXiv:1805.11183*, 2018.

Zhang, G., Sun, S., Duvenaud, D., and Grosse, R. Noisy Natural Gradient as Variational Inference. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.