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

Wu Lin†, Mohammad Emtiyaz Khan*, Mark Schmidt†
†University of British Columbia, *RIKEN Center for AI Project
wlin2018@cs.ubc.ca, emtiyaz.khan@riken.jp, schmidt@cs.ubc.ca

Abstract

Natural-gradient methods enable fast and simple algorithms for variational inference, but currently, their use is restricted to exponential-family approximations due to computational difficulties. In this paper, we extend the application of natural-gradients beyond exponential-family approximations. By proposing new types of expectation parameterizations, we derive simple updates for mixture of exponential-family approximations. Empirical results demonstrate a faster convergence of our method compared to black-box variational inference. Our work expands the scope of natural gradients for approximate Bayesian inference and makes them more widely applicable than before.

1. Introduction

Variational Inference provides a cheap and quick approximation to the posterior distribution and is now widely used in many areas of machine learning (Kingma and Welling, 2013; Furmston and Barber, 2010; Wainwright and Jordan, 2008; Hensman et al., 2013; Nguyen et al., 2017). In recent years, many natural-gradient methods have been proposed for variational inference (VI). However, most of them are restricted to exponential-family approximations (Honkela et al., 2011; Hensman et al., 2012; Hoffman et al., 2013; Khan and Lin, 2017; Salimbeni et al., 2018; Khan et al., 2018; Zhang et al., 2018). For such approximations, the natural gradient admits a simple form and can be computed without an explicit computation of the Fisher information matrix. Unfortunately, this simplicity does not extend to other types of approximations, and derivation of such simple updates for a generic approximation remains an open problem.

In this paper, we present a new approach to obtain simple natural-gradient updates for several types of approximations outside the class of exponential-family distributions. Our main focus is the structured approximations obtained by using the mixture of exponential-family distributions. For such distributions, we propose a new type of expectation-parameterization that enables the derivation of simple natural-gradient updates. We apply our method to a variety of models. Our experiments establish faster convergence than black-box VI methods.

2. Natural-Gradient Variational Inference (NGVI)

Given a probabilistic model $p(D, z)$ to model the data $D$ using a latent vector $z$, a common choice to approximate the posterior is to employ an exponential-family approximation: $p(z|D) \approx q(z|\lambda_z) := h_z(z) \exp \left[ \langle \phi_z(z), \lambda_z \rangle - A_z(\lambda_z) \right]$, where $q$ denotes the approximating distribution, $\lambda_z$ is the natural parameter, $\phi_z(z)$ are the sufficient statistics, $A_z(\lambda_z)$ is the log-partition function, and $h_z$ is the base measure. The approximation can be estimated
by maximizing the variational lower bound $\mathcal{L}(\lambda_z)$. This optimization can be carried out by either using gradient-descent (denoted by BBVI) or using natural-gradient method (denoted by NGVI) as shown below:

$$
\text{BBVI: } \lambda_z \leftarrow \lambda_z + \alpha \nabla_{\lambda_z} \mathcal{L}(\lambda_z), \quad \text{NGVI: } \lambda_z \leftarrow \lambda_z + \beta \left[ F_z(\lambda_z) \right]^{-1} \nabla_{\lambda_z} \mathcal{L}(\lambda_z),
$$

where $\nabla$ denotes the gradient, $F_z(\lambda_z)$ is the Fisher information matrix (FIM) with respect to the natural parameters, and $\alpha, \beta$ are scalar learning rates. The FIM is assumed to be invertible. In general, BBVI is computationally cheaper to perform and scales well to large data by using stochastic approximations of the gradients (Ranganath et al., 2014). However, NGVI could be simpler to compute than BBVI for some cases, e.g., in the stochastic variational inference (SVI) algorithm (Hoffman et al., 2013).

In general, a simple update can be obtained by using the expectation parameter of a minimal exponential family (Wainwright and Jordan, 2008) defined as $m_z(\lambda_z) := \mathbb{E}_q[\phi_z(z)] = \nabla_{\lambda_z} A(\lambda_z)$. The FIM is invertible in the minimal exponential family. As shown in Khan and Nielsen (2018), the NGVI update can then be expressed as follows:

$$
\text{NGVI for Exp-Family: } \lambda_z \leftarrow \lambda_z + \beta \nabla_{m_z} \mathcal{L}(\lambda_z),
$$

where the gradient is computed with respect to $m_z$. When the gradient with respect to $m_z$ is easier to compute than the gradient with respect to $\lambda_z$, we get a simpler algorithm that can also potentially converge faster. This approach has been used in some recent works to derive simple natural-gradient updates (Hensman et al., 2012; Khan and Lin, 2017; Khan et al., 2018; Salimbeni et al., 2018).

Unfortunately, when the approximations are not exponential-family distributions, the NGVI update may not be written in the form (2). This is because there may not even be a notion of the expectation parameter since the distribution may not be describable by sufficient statistics. This work addresses this issue for some types of non-exponential-family approximations and derives simple natural-gradient update similar to (2).

3. NGVI for Mixtures of Exponential-Family Distributions

We consider the mixture distributions $q(z) := \int q(z|w)q(w)dw$ where $q(w)$ is a mixing distribution that takes an exponential-family form as shown in (3) and $q(z|w)$ is the mixture component that takes the conditional exponential-family distribution as shown in (4).

$$
q(z|w) := h_w(w) \exp \left[ \langle \phi_w(w), \lambda_w \rangle - A_w(\lambda_w) \right], \quad q(z|w, \lambda_z) := h_z(z, w) \exp \left[ \langle \phi_z(z, w), \lambda_z \rangle - A_z(\lambda_z, w) \right].
$$

Note that, in $q(z|w, \lambda_z)$, the sufficient statistics $\phi_z(z, w)$ and log-partition $A_z(\lambda_z, w)$ depend on $w$. However, conditioned on $w$, the distribution is an exponential family parametrized by natural parameter $\lambda_z$. Therefore, this is a type of conditional exponential-family distribution (Xing et al., 2002; Liang et al., 2009; Lindsey, 1996; Feigin, 1981). We will give examples of this family soon, but first we define the FIM of $q(z, w)$. Denoting the set of natural parameters by $\lambda := \{\lambda_z, \lambda_w\}$, we can define the following FIM: $F_{wz}(\lambda) := -\mathbb{E}_{q(z,w)} \left[ \nabla_{\lambda}^2 \log q(z, w) \right]$. The FIM is defined by using the joint distribution of $z$ and $w$, and it is different from the one defined over the marginal distribution $q(z)$. We also assume the minimality holds in $q(w|\lambda_w)$ and $q(z|w, \lambda_z)$ so that the FIM is invertible. The minimality is discussed in Appendix A.
We consider the natural gradient in the Riemannian metric defined by $F_{wz}(\lambda)$. To simplify the natural-gradient update, we define the following expectation-parameters:

$$
\mathbf{m}_z(\lambda_z, \lambda_w) := \mathbb{E}_{q(z|w)q(w)}[\phi_z(z, w)], \quad \mathbf{m}_w(\lambda_w) := \mathbb{E}_{q(w)}[\phi_w(w)].
$$

Denote the set of expectation parameters by $\mathbf{m} := \{\mathbf{m}_z, \mathbf{m}_w\}$. The following theorem states our main result regarding the computation of natural gradients.

**Theorem 1** The natural-gradient update $\lambda \leftarrow \lambda + \beta \left[ F_{wz}(\lambda) \right]^{-1} \nabla_{\lambda} \mathcal{L}(\lambda)$ can be performed by using the following two updates which use the gradient with respect to $\mathbf{m}_z$ and $\mathbf{m}_w$:

$$
\lambda_z \leftarrow \lambda_z + \beta \nabla_{\mathbf{m}_z} \mathcal{L}(\lambda), \quad \lambda_w \leftarrow \lambda_w + \beta \nabla_{\mathbf{m}_w} \mathcal{L}(\lambda).
$$

A proof-sketch is given in Appendix A. The following examples show the simplicity of the proposed update (6).

**Example 1 (Finite Mixture of Gaussians)** The first example is regarding a finite mixture of Gaussians: $q(z) = \sum_{c=1}^{K} q(z|w=c)q(w=c)$ where $q(z|w=c) := \mathcal{N}(z|\mu_c, \Sigma_c)$ and $q(w=c) = \pi_c$ with $\sum_c \pi_c = 1$. The natural-parameter $\lambda_z := \{\Sigma^{-1}_c \mu_c, -\frac{1}{2} \Sigma^{-1}_c\}^{K}_{c=1}$ and the sufficient statistics is $\phi_z(z, w) := \{I_c(w)z, I_c(w)zz^T\}^{K}_{c=1}$ where $I_c(w)$ is an indicator taking a value 1 when $w = c$ and 0 otherwise. The mixing distribution $q(w)$ is a categorical distribution with $\lambda_w = \{\log \frac{\pi_c}{\pi_K}\}^{K-1}_{c=1}$ and $\phi_w(w) := \{I_c(w)\}^{K-1}_{c=1}$. The parameters $\mathbf{m}_z$ for the $c$-th component are simply equal to the mean parameter of the Gaussian $\mathcal{N}(z|\mu_c, \Sigma_c)$ multiplied by $\pi_c$, and $\mathbf{m}_w := \{\pi_c\}^{K-1}_{c=1}$. The gradient can be computed using the reparameterization trick. Figure 1 demonstrates the fast convergence of our algorithm over BBVI. This result can be extended to a finite mixture of exponential-family distributions discussed in Appendix B.

**Example 2 (Multivariate t-distribution)** The second example is the multivariate Student’s t-distribution which can be expressed as a scale mixture of Gaussian distribution:

$$
q(z) = \int q(z|w)q(w)dw = \int \mathcal{N}(z|\mu, w\Sigma) \mathrm{InvGam}(w|a, a)dw.
$$

The natural parameter and the sufficient statistics of $q(z|w)$ are $\lambda_z := \{\Sigma^{-1}, \mu, -\frac{1}{2} \Sigma^{-1}\}$ and $\phi_z(z, w) = \{w^{-1}z, w^{-1}zz^T\}$. These quantities for $q(w)$ are $\lambda_w := a$ and $\phi_w(w) := -1/w - \log w$. The expectation parameters defined in (5) are $\mathbf{m}_z := \{\mu, \mu\mu^T + \Sigma\}$ and $\mathbf{m}_w := -1 - \log a + \psi(a)$, where $\psi$ is the digamma function. Since $\mathbf{m}_z$ are essentially the expectation parameters of a Gaussian, the gradient with respect to them can be computed by adapting methods described in Khan et al. (2018). The resulting update can be expressed as a perturbed Newton update. The update with respect to $\mathbf{m}_w$ can be easily computed using the chain rule. The full update and experimental results are included in Appendix C and G, respectively. This result can be generalized to the general class of scale mixture of Gaussian distributions. Note that the joint distribution $\mathcal{N}(z|\mu, w\Sigma)\mathrm{InvGam}(w|a, a)$ indeed is a curved exponential family (Lehmann and Casella, 2006) as shown in Appendix C. Existing methods such as Khan and Lin (2017); Khan et al. (2018) cannot be directly used.

4. Extension to Multi-Linear Exponential Family

The mixture distribution discussed above is a hierarchical distribution that contains one block of parameters at each level. We can extend our approach to a distribution with multiple blocks of parameters at a single level. Suppose we divide the vector $\lambda := \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ where $\lambda_j$ is the $j$-th block of parameters. We consider a family of distribution, which takes the
following form: $q(z|\lambda) = h_z(z) \exp[f(z, \lambda) - A_z(\lambda)]$ with $f(\cdot)$ being a multi-linear function of $\lambda_j$, i.e., for all $j$, there exist functions $\phi_j$ and $r_j$ such that $f$ is linear with respect to $\lambda_j$, i.e., $f(z, \lambda) = (\lambda_j, \phi_j(z, \lambda_{-j})) + r_j(z, \lambda_{-j})$, where $\lambda_{-j}$ is the parameter vector containing all $\lambda$ except $\lambda_j$. We call this family the multi-linear exponential family. Obviously, an exponential family distribution parametrized by its natural parameter is a member of the multi-linear exponential family.

First, we define the natural parameter for the $j$-th block as $\lambda_j$. The expectation parameter can be defined in a similar fashion as the mixture case. The expectation parameter for the $j$-th block is defined as $m_j(\lambda) := E_{q(z)}[\phi_j(z, \lambda_{-j})]$. By defining an FIM for the $j$-th block as $F_j(\lambda) := -E_{q(z)}[\nabla^2_{\lambda_j} \log q(z|\lambda)]$, we can perform a block natural-gradient update: $\lambda_j \leftarrow \lambda_j + \beta_j [F_j(\lambda)]^{-1} \nabla_{\lambda_j} \mathcal{L}(\lambda)$. Similarly, we assume the minimality holds for all blocks so that $F_j(\lambda)$ is invertible. The following theorem establishes an update by using the expectation parameter $m_j$.

**Theorem 2** The above block natural-gradient update can be performed by using the following sequential update $\lambda_j \leftarrow \lambda_j + \beta \nabla_{m_j} \mathcal{L}(\lambda)$.

The proof of this theorem is similar to Theorem 1. More theoretical results can be found in Appendix E and F. The following example shows the simplicity of the proposed update (6).

**Example 3 (Matrix-Variate Gaussian (MVG))** As shown in Appendix D, the MVG distribution $MN(Z|W, U, V)$ can be written in the multi-linear form. The natural-gradient update is also derived in Appendix D and below we summarize the update. We first expressing the lower bound as $\mathcal{L}(\lambda) = E_q[-h(Z)]$. Under the Gauss-Newton approximation (Graves, 2011), the block natural-gradient update is,

$$W \leftarrow W - \beta_1 UGV, \quad U^{-1} \leftarrow U^{-1} + \beta_2 GV^\top, \quad V^{-1} \leftarrow V^{-1} + \beta_2 G^\top UG,$$  

(7)

where we sample $Z$ from the MVG distribution and evaluate the gradient $G := \nabla_Z h(Z)$. These updates extend the Newton-like update obtained in Khan et al. (2018) to MVG approximations. The gradient $G$ is pre-conditioned, which is very similar to other preconditioned algorithms, such as K-FAC (Martens and 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).
References


Appendix A. Proof of Theorem 1

The NGVI update of (6) updates the parameters $\lambda_z$ and $\lambda_w$ separately. This is possible because the FIM is in fact block-diagonal. The following lemma states this result.

**Lemma 3** The FIM $F_{wz}(\lambda)$ is block-diagonal with two blocks:

$$F_{wz}(\lambda) = \begin{bmatrix} F_z(\lambda) & 0 \\ 0 & F_w(\lambda_w) \end{bmatrix},$$

(8)

with the blocks defined as follows:

$$F_z(\lambda) := -\mathbb{E}_{q(z,w)} [\nabla^2_{\lambda z} \log q(z,w,\lambda)] , \quad F_w(\lambda_w) := -\mathbb{E}_{q(w)} [\nabla^2_{\lambda w} \log q(w,\lambda_w)]$$

(9)

**Proof** By definition of FIM, we have

$$F_{wz}(\lambda) = -\mathbb{E}_{q(z,w)\lambda} \left[ \nabla^2_{\lambda z} \log q(z,w,\lambda) \right]$$

(10)

$$= -\mathbb{E}_{q(z,w)\lambda} \left[ \begin{array}{c} \nabla^2_{\lambda z} \log q(z,w,\lambda) \\ \nabla_{\lambda w} \nabla_{\lambda z} \log q(z,w,\lambda) \\ \nabla^2_{\lambda w} \log q(z,w,\lambda) \end{array} \right]$$

(11)

$$= -\mathbb{E}_{q(z,w)\lambda} \left[ \begin{array}{c} \nabla^2_{\lambda z} \log q(z,w,\lambda_z) + \log q(z,w,\lambda_w) \\ \nabla_{\lambda w} \nabla_{\lambda z} \log q(z,w,\lambda) \\ \nabla^2_{\lambda w} \log q(z,w,\lambda) \end{array} \right]$$

(12)

$$= -\mathbb{E}_{q(z,w)\lambda} \left[ \begin{array}{c} \nabla^2_{\lambda z} \log q(z,w,\lambda_z) \\ 0 \\ \nabla^2_{\lambda w} \log q(w,\lambda_w) \end{array} \right]$$

(13)

$$= -\left[ \mathbb{E}_{q(z,w)\lambda} \left[ \nabla^2_{\lambda z} \log q(z,w,\lambda_z) \right] \begin{array}{c} 0 \\ 0 \end{array} \right] + \mathbb{E}_{q(w)\lambda_w} \left[ \nabla^2_{\lambda w} \log q(w,\lambda_w) \right]$$

(14)

$$= \begin{bmatrix} F_z(\lambda) & 0 \\ 0 & F_w(\lambda_w) \end{bmatrix},$$

(15)

where we can move from step 12 to step 13 since $\lambda_z$ and $\lambda_w$ are not tied.

By Lemma 3, the updates can therefore be carried out separately. The updates for $\lambda_w$ can be expressed in terms of its mean parameters by using the results of Khan and Nielsen (2018) because $q(w)$ is an exponential family distribution. Therefore, we only need to derive the update for $\lambda_z$. The following theorem states that the corresponding FIM can be written as the derivative of $m_z$ with respect to $\lambda_z$.

**Lemma 4** The FIM matrix $F_z$ is equal to the derivative of the expectation parameter $m_z(\lambda)$:

$$F_z(\lambda) := \nabla_{\lambda z} m_z(\lambda)$$

(16)

**Proof** Recall that

$$\mathbb{E}_{q(z,w)\lambda} \left[ \nabla_{\lambda z} \log q(z,w,\lambda) \right] = \mathbb{E}_{q(z,w)\lambda} \left[ \nabla_{\lambda z} \log q(z,w,\lambda_z) \right]$$

(17)

$$= \mathbb{E}_{q(z,w)\lambda} \left[ \nabla_{\lambda z} q(z,w,\lambda_z) / q(z,w,\lambda_z) \right]$$

(18)

$$= \mathbb{E}_{q(w)\lambda_w} \left[ \nabla_{\lambda z} q(z,w,\lambda_z) \right]$$

(19)

$$= \nabla_{\lambda z} \mathbb{E}_{q(z,w)\lambda} [1]$$

(20)

$$= 0$$

(21)
Since \( q(z|w, \lambda_z) = h_z(z, w) \exp [(\phi_z(z, w), \lambda_z) - A_z(\lambda_z, w)] \), we have

\[
0 = E_{q(z,w)} [\nabla_{\lambda_z} \log q(z, w|\lambda)] \\
= E_{q(z,w)} [\nabla_{\lambda_z} \log q(z|w, \lambda_z)] \\
= E_{q(z,w)} [\nabla_{\lambda_z} (\log h_z(z, w) + \langle \phi_z(z, w), \lambda_z \rangle - A_z(\lambda_z, w))] \\
= E_{q(z,w)} [\nabla_{\lambda_z} (\langle \phi_z(z, w), \lambda_z \rangle - A_z(\lambda_z, w))] \\
= E_{q(z,w)} [\phi_z(z, w)] - E_{q(z,w)} [\nabla_{\lambda_z} A_z(\lambda_z, w)] \\
= m_z - E_{q(z,w|\lambda)} [\nabla_{\lambda_z} A_z(\lambda_z, w)] \\
= m_z - E_{q(w|\lambda_z)} [\nabla_{\lambda_z} A_z(\lambda_z, w)] \\
(22)
\]

Therefore, we have

\[
F_z(\lambda) = -E_{q(z,w)} [\nabla_{\lambda_z}^2 \log q(z|w, \lambda_z)] \\
= -E_{q(z,w)} [\nabla_{\lambda_z}^2 (\log h_z(z, w) + \langle \phi_z(z, w), \lambda_z \rangle - A_z(\lambda_z, w))] \\
= E_{q(z,w)} \nabla_{\lambda_z}^2 A_z(\lambda_z, w) \\
= E_{q(w|\lambda_z)} \nabla_{\lambda_z}^2 A_z(\lambda_z, w) \\
= \nabla_{\lambda_z} \nabla_{\lambda_z} m_z \\
(29)
\]

Using this result and applying chain rule, we can write the following:

\[
F_z(\lambda) [\nabla_{m_z} L(\lambda)] = \nabla_{\lambda_z} m_z(\lambda) [\nabla_{m_z} L(\lambda)] = \nabla_{\lambda_z} L(\lambda) \\
(35)
\]

If the FIM is invertible, we can express the natural-gradient with respect to \( \lambda_z \) as the gradient with respect to \( m_z \). The following lemma establishes the conditions under which the FIM is invertible. This condition is just a sufficient condition. There are cases where the FIM is invertible but the condition does not hold.

**Lemma 5** Let’s denote the probability space of \( q(w) \) by \( \Omega \). If there exists an area \( S \subseteq \Omega \) such that \( S \) is a non-zero measure and \( q(z|w) \) is a minimal exponential-family distribution for all \( w \in S \), then \( F_z(\lambda) \) is invertible.

**Proof** Conditioned on \( w \), we know that \( q(z|w) \) is an exponential-family distribution as shown below.

\[
q(z|w) = h_z(z, w) \exp [(\phi_z(z, w), \lambda_z) - A_z(\lambda_z, w)] \\
(36)
\]

Note that

\[
E_{q(z|w)} [\nabla_{\lambda_z}^2 \log q(z|w, \lambda_z)] \\
= E_{q(z|w)} [\nabla_{\lambda_z} \phi_z(z, w) - \nabla_{\lambda_z} A_z(\lambda_z, w)] \\
= E_{q(z|w)} [-\nabla_{\lambda_z} A_z(\lambda_z, w)] \\
= -\nabla_{\lambda_z}^2 A_z(\lambda_z, w) \\
(37)
\]

\[
E_{q(z|w)} [\phi_z(z, w)] - E_{q(z|w)} [\nabla_{\lambda_z} A_z(\lambda_z, w)] \\
= E_{q(z|w)} [-\nabla_{\lambda_z} A_z(\lambda_z, w)] \\
= -\nabla_{\lambda_z}^2 A_z(\lambda_z, w) \\
(38)
\]

\[
E_{q(z|w)} [\phi_z(z, w)] - E_{q(z|w)} [\nabla_{\lambda_z} A_z(\lambda_z, w)] \\
= E_{q(z|w)} [-\nabla_{\lambda_z} A_z(\lambda_z, w)] \\
= -\nabla_{\lambda_z}^2 A_z(\lambda_z, w) \\
(39)
\]

\[
E_{q(z|w)} [\phi_z(z, w)] - E_{q(z|w)} [\nabla_{\lambda_z} A_z(\lambda_z, w)] \\
= E_{q(z|w)} [-\nabla_{\lambda_z} A_z(\lambda_z, w)] \\
= -\nabla_{\lambda_z}^2 A_z(\lambda_z, w) \\
(40)
\]
Let’s denote $e_z = \frac{\nabla_{\lambda_z} q(z|w, \lambda_z)}{q(z|w, \lambda_z)}$. Conditioned on $w$, we have
\[
\nabla_{\lambda_z}^2 A_z(\lambda_z, w)
= -\mathbb{E}_{q(z|w)} \left[ \nabla_{\lambda_z}^2 \log q(z|w, \lambda_z) \right]
\]
\[
= -\mathbb{E}_{q(z|w)} \left[ \nabla_{\lambda_z} \left( \frac{\nabla_{\lambda_z} q(z|w, \lambda_z)}{q(z|w, \lambda_z)} \right) \right]
\]
\[
= -\mathbb{E}_{q(z|w)} \left[ \frac{\nabla_{\lambda_z} q(z|w, \lambda_z)}{q(z|w, \lambda_z)} - e_z e_z^T \right]
\]
\[
= \mathbb{E}_{q(z|w)} \left[ -\frac{\nabla_{\lambda_z}^2 q(z|w, \lambda_z)}{q(z|w, \lambda_z)} \right] + \mathbb{E}_{q(z|w)} \left[ e_z e_z^T \right]
\]
\[
= \mathbb{E}_{q(z|w)} \left[ e_z e_z^T \right] \geq 0,
\]
where we use the chain rule to move from step 43 to step 44.
Since $q(z|w)$ is a minimal exponential-family distribution for all $w \in S$, by Proposition 3.1 of Wainwright and Jordan (2008), we know that $A_z(\lambda_z, w)$ is strictly convex with respect to $\lambda_z$ given $w \in S$ is known, which implies that $\nabla_{\lambda_z}^2 A_z(\lambda_z, w) \succ 0$.
Using Eq. 32, we have
\[
F_z(\lambda) = -\mathbb{E}_{q(z,w)} \left[ \nabla_{\lambda_z}^2 \log q(z|w, \lambda_z) \right]
\]
\[
= \mathbb{E}_{q(z,w)} \left[ \nabla_{\lambda_z} A_z(\lambda_z, w) \right]
\]
Given any nonzero vector $a$, we have
\[
a^T F_z(\lambda) a
\]
\[
= a^T \left[ \int_{w \in S} q(w|\lambda_w) \nabla_{\lambda_z} A_z(\lambda_z, w) dw + \int_{w \in (\Omega - S)} q(w|\lambda_w) \nabla_{\lambda_z} A_z(\lambda_z, w) dw \right] a
\]
\[
= \int_{w \in S} q(w|\lambda_w) \left[ a^T \left( \nabla_{\lambda_z}^2 A_z(\lambda_z, w) \right) a \right] \nabla_{\lambda_z} A_z(\lambda_z, w) dw + \int_{w \in (\Omega - S)} q(w|\lambda_w) \left[ a^T \left( \nabla_{\lambda_z}^2 A_z(\lambda_z, w) \right) a \right] \geq 0
\]
\[
\geq \int_{w \in S} q(w|\lambda_w) \left[ a^T \left( \nabla_{\lambda_z}^2 A_z(\lambda_z, w) \right) a \right] \nabla_{\lambda_z} A_z(\lambda_z, w) dw
\]
\[
> 0.
\]
We use Eq 47 to move from step 52 to step 53. Furthermore, we use the strict convexity of $A_z(\lambda_z, w)$ with respect to $\lambda_z$ when $w \in S$ and the nonzero measure $S$ to move from step 53 to step 54.
Therefore, we know that $F_z(\lambda) \succ 0$ and $F_z(\lambda)$ is invertible.
Appendix B. Updates for Finite Mixture of Gaussian Distributions

Let’s consider \( q(w) = \text{Cate}(w|\pi) \) and \( q(z|w) = \text{ExpFmy}(z|\lambda_w) \). We assume \( q(z|w) \) is (implicitly) re-parameterizable.

**Lemma 6** \( q(z|w) \) is a conditional exponential family distribution.

**Proof**

\[
q(z|w) = \sum_{c=1}^{K} \mathbb{I}_c(w) \text{ExpFmy}(z|\lambda_{zc}) \tag{55}
\]

\[
= \sum_{c=1}^{K} \mathbb{I}_c(w) h_z(z) \exp \left[ \lambda_{zc}^T \phi_z(z) - A_z(\lambda_{zc}) \right] \tag{56}
\]

\[
= h_z(z) \exp \left\{ \sum_{c=1}^{K} \left[ \mathbb{I}_c(w) \phi_z(z) , \lambda_{zc} \right] - \sum_{c'=1}^{K} \mathbb{I}_{c'}(w) A_z(\lambda_{zc'}) \right\} \tag{57}
\]

Firstly, we give the update for \( q(z|w) \). The natural parameter and sufficient statistics \( \{\lambda_{zc}\}_{c=1}^{K} \) and \( \{\mathbb{I}_c(w) \phi_z(z)\}_{c=1}^{K} \).

The update is

\[
\lambda_{zc} \leftarrow \lambda_{zc} + \beta_z \nabla m_zc \mathcal{L}(\lambda), \tag{58}
\]

where \( m_zc = \mathbb{E}_{q(w,z)} [\mathbb{I}_c(w) \phi_z(z)] \).

We consider the following model.

\[
p(D, z) = \prod_{n=1}^{N} p(D_n|z)p(z) \tag{59}
\]

Let denote \( h_n(z) := -(\log p(D_n|z) + \log p(z)/N) \). The lower bound can be expressed as

\[
\mathcal{L}(\lambda) = \mathbb{E}_{q(w,z)} \left[ -\sum_{n=1}^{N} h_n(z) - \log q(z, w) \right] \tag{60}
\]

\[
= \mathbb{E}_{q(w,z)} \left[ -\sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(z|\lambda_{zc'}) \right]. \tag{61}
\]

In general, \( \nabla m_zc \mathcal{L}(\lambda) \) can be computed

\[
\nabla m_zc \mathcal{L}(\lambda) = (\nabla_{\lambda_{zc}} m_zc)^{-1} \nabla_{\lambda_{zc}} \mathcal{L}(\lambda) \tag{62}
\]

\[
= (\nabla_{\lambda_{zc}} \mathbb{E}_{q(w,z)} [\mathbb{I}_c(w) \phi_z(z)])^{-1} \nabla_{\lambda_{zc}} \mathcal{L}(\lambda), \tag{63}
\]

where

\[
\nabla_{\lambda_{zc}} \mathbb{E}_{q(w,z)} [\mathbb{I}_c(w) \phi_z(z)] = \nabla_{\lambda_{zc}} \int \pi_c q(z|w=c) \phi_z(z) dz \tag{64}
\]

\[
= \int \pi_c q(z|w=c) \nabla_z [\phi_z(z)] \nabla_{\lambda_{zc}} z dz \tag{65}
\]
\( \nabla_{\lambda_c} \mathcal{L}(\lambda) \) can be computed as below.

\[
\nabla_{\lambda_c} \mathcal{L}(\lambda) = \nabla_{\lambda_c} \mathbb{E}_{q(z)} \left[ -\sum_{n=1}^{N} h_n(z) \right] - \nabla_{\lambda_c} \mathbb{E}_{q(z)} \left[ \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(z|\lambda_{z,c'}) \right] \tag{66}
\]

\[
= \int \pi_c \nabla_{\lambda_c} q(z|w = c) \left[ -\sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(z|\lambda_{z,c'}) \right] dz + \int_{0}^{q(z)} \nabla_{\lambda_c} \log q(z) dz \tag{67}
\]

\[
= \int \pi_c q(z|w = c) \left[ \nabla_z \left( -\sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(z|\lambda_{z,c'}) \right) \right] \left[ \nabla_{\lambda_c} z \right] dz \tag{68}
\]

Now, we give the update for \( q(w|\lambda_w) = \text{Cate}(w|\pi) \), which is an exponential family distribution as shown below.

\[
\text{Cate}(w|\pi) = \exp \left\{ \left( \sum_{c=1}^{K-1} \pi_c \log \frac{\pi_c}{\pi_K} \right) + \log \pi_K \right\}, \tag{69}
\]

where \( \pi_K = 1 - \sum_{c=1}^{K-1} \pi_c \)

Its natural parameter and expectation parameter are

\[
\lambda_w = \{ \log \frac{\pi_c}{\pi_K} \}_{c=1}^{K-1} \tag{70}
\]

\[
m_w = \{ \mathbb{E}_{q(z)} \left[ \mathbb{I}_c(w) \right] \}_{c=1}^{K-1} = \{ \pi_c \}_{c=1}^{K-1} \tag{71}
\]

Recall that \( \mathcal{L}(\lambda) \) can be expressed as

\[
\mathcal{L}(\lambda) = \mathbb{E}_{q(w,z)} \left[ -\sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(z|\lambda_{z,c'}) \right]. \tag{72}
\]

\( \nabla_{\pi_c} \mathcal{L}(\lambda) \) can be computed as below.

\[
\nabla_{\pi_c} \mathcal{L}(\lambda) = \nabla_{\pi_c} \mathbb{E}_{q(z)} \left[ -\sum_{n=1}^{N} h_n(z) \right] - \nabla_{\pi_c} \mathbb{E}_{q(z)} \left[ \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(z|\lambda_{z,c'}) \right] \tag{73}
\]

\[
= \int q(z|w = c) \left[ -\sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(z|\lambda_{z,c'}) \right] dz + \int_{0}^{q(z)} \nabla_{\pi_c} \log q(z) dz \tag{74}
\]

\[
= \int q(z|w = c) \left[ -\sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(z|\lambda_{z,c'}) \right] dz \tag{75}
\]

Observe that in Eq 65, 68, 75, we have to use at least \( K \) MC samples to compute the gradients for a \( K \)-mixture distribution. To address this, we propose to use importance
sampling to reduce the number of MC samples. The gradients are computed as below.

\[
\nabla_{\lambda_{xc}} E_{q(w,z)} \left[ \mathbb{I}(w = c) \phi_z(z) \right] = \int q(z) \frac{\pi_c q(z|w = c)}{q(z)} \left[ \nabla_z \phi_z(z) \right] \nabla_{\lambda_{xc}} z \, dz
\]

(76)

\[
= \mathbb{E}_{q(z)} \left[ q(w = c|z) \nabla_z \phi_z(z) \nabla_{\lambda_{xc}} z \right]
\]

(77)

\[
\nabla_{\lambda_{xc}} \mathcal{L}(\lambda) = \int q(z) \frac{\pi_c q(z|w = c)}{q(z)} \nabla_z \left[ - \sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \exp \mathcal{F}(z|\lambda_{zc'}) \right] \nabla_{\lambda_{xc}} z \, dz
\]

(78)

\[
= \mathbb{E}_{q(z)} \left[ q(w = c|z) \nabla_z \left[ - \sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \exp \mathcal{F}(z|\lambda_{zc'}) \right] \nabla_{\lambda_{xc}} z \right]
\]

(79)

\[
\nabla_{\pi_c} \mathcal{L}(\lambda) = \int q(z) \frac{q(z|w = c)}{q(z)} \left[ - \sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \exp \mathcal{F}(z|\lambda_{zc'}) \right] \, dz
\]

(80)

\[
= \mathbb{E}_{q(z)} \left[ q(z|w = c) \left[ - \sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \exp \mathcal{F}(z|\lambda_{zc'}) \right] \right]
\]

(81)

In the mixture of Gaussian case, we can directly compute the gradient \( \nabla_{m_{z,c}} \mathcal{L}(\lambda) \). Recall that the natural parameters of \( q(z|w) \) are \( \lambda_z = \lbrace \Sigma_c^{-1} \mu_c, -\frac{1}{2} \Sigma_c^{-1} \rbrace_{c=1}^{K} \). The expectation parameters of \( q(z|w) \) are \( m_z = \lbrace \pi_c \mu_c, \pi_c \left( \mu_c T + \Sigma_c \right) \rbrace_{c=1}^{K} \).

By the chain rule, we know that

\[
\nabla_{m_{z,c}} \mathcal{L}(\lambda) = \frac{1}{\pi_c} (\nabla_{\mu_c} \mathcal{L}(\lambda) - 2 \nabla_{\Sigma_c} \mathcal{L}(\lambda) \mu_c)
\]

(82)

\[
\nabla_{m_{\Sigma,c}} \mathcal{L}(\lambda) = \frac{1}{\pi_c} (\nabla_{\Sigma_c} \mathcal{L}(\lambda))
\]

(83)

By the Bonnet’s and Price’s theorem, we have

\[
\nabla_{\mu_c} \mathcal{L}(\lambda) = \mathbb{E}_{q(z)} \left[ q(w = c|z) \nabla_z \left( - \sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \mathcal{N}(z|\mu_{c'}, \Sigma_{c'}) \right) \right]
\]

(84)

\[
\nabla_{\Sigma_c} \mathcal{L}(\lambda) = \frac{1}{2} \mathbb{E}_{q(z)} \left[ q(w = c|z) \nabla_z^2 \left( - \sum_{n=1}^{N} h_n(z) - \log \sum_{c'=1}^{K} \pi_{c'} \mathcal{N}(z|\mu_{c'}, \Sigma_{c'}) \right) \right].
\]

(85)

### Appendix C. Update for Multivariate t-Distribution

In this section, we derive the update for the multivariate t-distribution. The update can be extended to the class of normal variance-mean mixture including the scale mixture of Gaussians.

**Lemma 7** \( \mathcal{N}(z|\mu, w \Sigma) \text{InvGam}(w|a, a), \) where \( z \in \mathbb{R}^d \), is a curved exponential family distribution.
We know that variational distribution. Note that when \( \text{et al. (2018).} \)

For this model, by using the Gauss-Newton approximation of the Hessian, the NGVI update 

\[
\lambda = -\frac{1}{2} \mu^T \Sigma^{-1} \mu, \lambda_2 = \Sigma^{-1} \mu, \lambda_3 = -\frac{1}{2} \mu^T \Sigma^{-1} \mu, \lambda_4 = -a.
\]

The distribution on \( z \in \mathcal{R}^d \) and \( w \in \mathcal{R}_{++} \) can be written as a weight-perturbed Newton method similar to Khan et al. (2018).

Proof Let \( \Lambda_1 = -\frac{1}{2} \Sigma^{-1}, \lambda_2 = \Sigma^{-1} \mu, \lambda_3 = -\frac{1}{2} \mu^T \Sigma^{-1} \mu, \lambda_4 = -a. \) The distribution on \( z \in \mathcal{R}^d \) and \( w \in \mathcal{R}_{++} \) can be expressed as follows.

\[
\mathcal{N}(z|\mu, w\Sigma)\text{InvGam}(w|a, a) = \text{det}(2\pi w \Sigma)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(z - \mu)^T (w \Sigma)^{-1} (z - \mu)\} \frac{a^a}{\Gamma(a)} w^{-a-1} \exp\{-\frac{a}{w}\}
\]

\[
= (2\pi w)^{-d/2} w^{-1} \exp\{-\frac{1}{2}(z - \mu)^T (w \Sigma)^{-1} (z - \mu) - \frac{1}{2} \log \det(\Sigma) \}
- \frac{a}{w} - a \log w - (\log \Gamma(a) - a \log a) \}
\]

\[
= (2\pi w)^{-d/2} w^{-1} \exp\{\langle -\frac{1}{2} \Sigma^{-1}, wz \rangle + \langle \Sigma^{-1} \mu, wz \rangle + \langle -\frac{1}{2} \mu^T \Sigma^{-1} \mu, w^{-1}\rangle
+ \langle -a, w^{-1} + \log w \rangle - [\log \Gamma(a) - a \log a + \frac{1}{2} \log \det(\Sigma)] \}
\]

\[
= (2\pi w)^{-d/2} w^{-1} \exp\{\langle \Lambda_1, wz \rangle + \langle \lambda_2, wz \rangle + \langle \lambda_3, w^{-1}\rangle
+ \langle \lambda_4, w^{-1} + \log w \rangle - [\log \Gamma(-\lambda_4) + \lambda_4 \log(-\lambda_4) - \frac{1}{2} \log \det(-2\Lambda_1)] \},
\]

It is easy to see that the joint distribution is an exponential family distribution. Note that \( \lambda_3 \) is fully determined by \( \Lambda_1 \) and \( \lambda_2 \) as shown below.

\[
\lambda_3 = -\frac{1}{2} \mu^T \Sigma^{-1} \mu
= -\frac{1}{2} (\Sigma \Sigma^{-1} \mu)^T (\Sigma^{-1} \mu)
= -\frac{1}{2} \left(( -2\Lambda_1)^{-1} \lambda_2 \right)^T \lambda_2
\]

We know that \( \mathcal{N}(z|\mu, w\Sigma)\text{InvGam}(w|a, a) \) is a curved exponential family distribution since \( \lambda_3 \) is fully determined by \( \Lambda_1 \) and \( \lambda_2 \).

We give the update for the following model where the prior over \( p(z, w) \) is a joint distribution of a Multivariate t-Distribution. Our method can be easily applied to models even when this is not the case. We use \( q(z, w) = \mathcal{N}(z|\mu, w\Sigma)\text{InvGam}(w|a, a) \) as the variational distribution. Note that when \( a \leq 1 \), the variance of the marginal distribution \( q(z) \) does not exist. We assume that \( a > 1 \) such that the variance exists.

Given \( z \in \mathcal{R}^d \) and \( w \in \mathcal{R}_{++} \), we consider the following model :

\[
p(D, z, w) = \prod_{n=1}^{N} p(D_n|z) \mathcal{N}(z|0, w\mathbf{I}) \text{InvGam}(w|a_0, a_0)
\]

For this model, by using the Gauss-Newton approximation of the Hessian, the NGVI update given in Theorem 1 can be written as a weight-perturbed Newton method similar to Khan et al. (2018).

1. Generate a sample \( (z_*, w_*) \sim q(z, w|\lambda) \).

2. Sample a random data example \( p(D_n|z) \) and compute its gradient at \( z_* \).

\[
g \leftarrow \nabla_z \log p(D_n|z)
\]
3. Compute a learning-rate modifier:
\[
u \leftarrow \frac{a}{a + d/2 - 1} \left[ 1 + \frac{1}{2} (z_s - \mu)^\top \Sigma^{-1} (z_s - \mu) / a \right]
\] (94)

4. Update \(a\) for the Inverse-Gamma distribution:
\[
a \leftarrow (1 - \beta_2) a + \beta_2 \left[ a_0 - \frac{N w^2}{2 (1 - w_s)} g^\top \Sigma g \right],
\] (95)

where we choose \(\beta_2\) such that \(a > 1\).

5. Take a Newton-like step to update \(\mu\) and \(\Sigma\).
\[
S \leftarrow (1 - \beta_1) S + \beta_1 (ugg^\top)
\] (96)
\[
\mu \leftarrow \mu - \beta_1 (S + I/N)^{-1} (g + \mu/N)
\] (97)

6. Repeat until convergence.

The update of \(\mu\) and \(S\) is very similar to the Gauss-Newton update of Khan et al. (2018), but here the learning rate \(\beta_1\) is multiplied by \(u\). The algorithm therefore can be implemented easily by using further approximations discussed in Khan et al. (2018).

To derive this algorithm, we prove a version of Bonnet’s and Price’s theorem for the case when the expectation is taken with respect to a scale-mixture of Gaussian instead of a Gaussian.

First of all, we show how to update \(q(z|w) = N(z|\mu, w \Sigma)\). The natural parameter and the sufficient statistics of \(q(z|w)\) are \(\lambda_z = \{\Sigma^{-1} \mu, -1/2 \Sigma^{-1}\} \text{ and } \phi_z(z, w) = \{w^{-1} z, w^{-1} zz^T\}\). The expectation parameters are
\[
m_{z1} = \mathbb{E}_{q(z,w)} [w^{-1} z] = \mu
\] (98)
\[
m_{z2} = \mathbb{E}_{q(z,w)} [w^{-1} zz^T] = \mu \mu^T + \Sigma
\] (99)

By the chain rule, we know that
\[
\nabla_{m_{z1}} L(\lambda) = \nabla_\mu L(\lambda) - 2 \nabla_\Sigma L(\lambda) \mu
\] (100)
\[
\nabla_{m_{z2}} L(\lambda) = \nabla_\Sigma L(\lambda)
\] (101)

Let’s denote \(h_n(z) := -\log p(D_n|z)\). In the model, the lower bound \(L(\lambda)\) can be expressed as
\[
L(\lambda) = \mathbb{E}_{q(z,w)} \left[ -\sum_{n=1}^N h_n(z) + \log N(z|0, wI) + \log \frac{\text{InvGam}(w|a_0, a_0)}{\text{InvGam}(w|a, a)} \right],
\] (102)

The gradients of the lower bound are
\[
\nabla_\mu L(\lambda) = \nabla_\mu \mathbb{E}_{q(z,w)} \left[ -\sum_{n=1}^N h_n(z) \right] - \mu
\] (103)
\[
\nabla_\Sigma L(\lambda) = \nabla_\Sigma \mathbb{E}_{q(z,w)} \left[ -\sum_{n=1}^N h_n(z) \right] - \frac{1}{2} I + \frac{1}{2} \Sigma^{-1}
\] (104)
The natural gradient update is

\[ \lambda_{z_1} \leftarrow \lambda_{z_1} + \beta_1 \nabla_{m_{z_1}} L(\lambda) \]  
\[ \lambda_{z_2} \leftarrow \lambda_{z_2} + \beta_1 \nabla_{m_{z_2}} L(\lambda) \]  

(105)

(106)

Since \( \lambda_{z_1} = \Sigma^{-1} \mu, \lambda_{z_2} = -\frac{1}{2} \Sigma^{-1} \), using Eq.100,101,103,104, the update can be re-expressed as

\[ \Sigma^{-1} \leftarrow (1 - \beta_1) \Sigma^{-1} + 2\beta_1 \nabla_{\Sigma} E_{q(z,w)} \left[ \sum_{n=1}^{N} h_n(z) \right] + \beta_1 I \]  

(107)

\[ \mu \leftarrow \mu - \beta_1 \Sigma \left( \nabla_{\mu} E_{q(z,w)} \left[ \sum_{n=1}^{N} h_n(z) \right] + \mu \right) \]  

(108)

By defining \( S = \left[ \Sigma^{-1} - I \right] / N \), we have

\[ S \leftarrow (1 - \beta_1) S + 2\beta_1 \nabla_{\Sigma} E_{q(z,w)} \left[ \sum_{n=1}^{N} h_n(z)/N \right] \]  

(109)

\[ \mu \leftarrow \mu - \beta_1 \left( S + I/N \right)^{-1} \left( \nabla_{\mu} E_{q(z,w)} \left[ \sum_{n=1}^{N} h_n(z)/N \right] + \mu / N \right) \]  

(110)

Now, we show the update for \( q(w) = \text{InvGam}(w|a,a) \). The natural parameter and the sufficient statistics of \( q(w) \) are \( \lambda_w = a \) and \( \phi_w(w) = -w^{-1} - \log w \).

The gradient can be expressed as

\[ \nabla_{m_w} L(\lambda) = \nabla_{m_w} E_{q(z,w)} \left[ - \sum_{n=1}^{N} h_n(z) \right] + a_0 - a \]  

(111)

The update can be expressed in terms of \( a \) as

\[ a \leftarrow (1 - \beta_2) a + \beta_2 \left( a_0 - \sum_{n=1}^{N} \nabla_{m_w} E_{q(z,w)} \left[ h_n(z) \right] \right) \]  

(112)

While the gradient w.r.t. the expectation parameter does not admit a close-form expression, we can compute the gradient using the re-parametrization trick. Recall that the gradient \( \nabla_{m_w} E_{q(z,w)} \left[ h_n(z) \right] \) can be computed as

\[ \nabla_{m_w} E_{q(z,w)} \left[ h(z) \right] = (\nabla_{\lambda_w} m_w)^{-1} \nabla_{\lambda_w} E_{q(z,w)} \left[ h_n(z) \right] \]  

(113)

\[ = (\nabla_{\lambda_w} E_{q(w)} \left[ \phi_w(w) \right])^{-1} \nabla_{\lambda_w} E_{q(z,w)} \left[ h_n(z) \right] \]  

(114)

\[ = (\nabla_{a} E_{q(w)} \left[ \phi_w(w) \right])^{-1} \nabla_{a} E_{q(z,w)} \left[ h_n(z) \right] \]  

(115)

Note that \( \nabla_{a} E_{q(w)} \left[ \phi_w(w) \right] = \nabla a m_w \) has a closed-form expression.
Even when $\nabla_a m_w$ does not admit a closed-form expression, we can also compute it when $q(w)$ is re-parameterizable shown below. Since $q(w)$ is (implicitly) re-parameterizable, the gradient can be computed as

$$\nabla_a \mathbb{E}_{q(w)}\left[\phi_w(w)\right] = - \int \text{InvGam}(w|a, a) \left(\nabla_w \left[w^{-1} + \log w\right]\right) (\nabla_a w) dw$$  \hfill (116)

$$\nabla_a \mathbb{E}_{q(z,w)} \left[h_n(z)\right] = \int \int \left(\nabla_w q(z|w)\right) \left(\nabla_a w\right) h_n(z) dw dz$$

$$= \int \int \text{InvGam}(w|a, a) \left(\nabla_w \mathcal{N}(z|\mu, w \Sigma)\right) (\nabla_a w) h_n(z) dw dz$$  \hfill (118)

Now, we discuss about gradient approximation. Recall that the joint distribution is $\text{InvGam}(w|a, a) \mathcal{N}(z|\mu, w \Sigma)$. The marginal distribution is a multivariate t-distribution as shown below

$$q(z) = \det (\pi \Sigma)^{-1/2} \frac{\Gamma(a + d/2) \left(2a + (z - \mu)^T \Sigma^{-1} (z - \mu)\right)^{-a-d/2}}{\Gamma(a)^2 (2a)}.$$  \hfill (119)

When $a > 1$, we denote the following function as $u(z)$

$$u(z) := \int wq(z|w)q(w)dw = \frac{a}{(a + d/2 - 1)} \left(1 + (2a)^{-1} (z - \mu)^T \Sigma^{-1} (z - \mu)\right) q(z).$$  \hfill (120)

**Lemma 8** Let $h(z)$ be a twice continuously differentiable function and $q(w) = \text{InvGam}(w|a, a)$, where $a > 1$. If $|h(z)|, |z_j h(z)|$, and $|\nabla_z \phi(h(z))|$ are integrable for any index $j$ and $k$, the following gradient identities hold.

$$\nabla_{\mu} \mathbb{E}_{q(z,w) \mathcal{N}(z|\mu, w \Sigma)} \left[h(z)\right] = \int q(z) \nabla_z h(z) dz$$

$$\nabla_{\Sigma} \mathbb{E}_{q(z,w) \mathcal{N}(z|\mu, w \Sigma)} \left[h(z)\right] = \frac{1}{2} \int u(z) \nabla^2_{zz} h(z) dz.$$  \hfill (122)

The lemma can be generalized to the case when $q(w)$ is a generalized inverse Gaussian distribution.

**Proof**

$$\nabla_{z_j} \mathbb{E}_{q(z,w) \mathcal{N}(z|\mu, w \Sigma)} \left[h(z)\right] = \int \int q(w) \left[\nabla_{z_j} \mathcal{N}(z|\mu, w \Sigma)\right] h(z) dw dz$$

$$= - \int q(w) \left\{ \int \left[\nabla_{z_j} \mathcal{N}(z|\mu, w \Sigma)\right] h(z) dz \right\} dw$$

$$= - \int q(w) \left\{ \int \mathcal{N}(z|\mu, w \Sigma) h(z) dz \bigg|_{z_j = +\infty}^{z_j = -\infty} - \int \mathcal{N}(z|\mu, w \Sigma) \nabla_{z_j} h(z) dz \right\} dw$$

$$= - \int q(z) \left[ h(z) \bigg|_{z_j = +\infty}^{z_j = -\infty} + \int q(w) \int \mathcal{N}(z|\mu, w \Sigma) \nabla_{z_j} h(z) dz dw \right]$$

$$= \int q(z) \nabla_{z_j} h(z) dz.$$  \hfill (127)
We use integration by part twice to move from Eq. 124 to Eq. 125. Since \(|h(\mathbf{z})|\) is integrable, we know that the first term in Eq. 126 is 0.

Similarly, we have the following result by using integration by part twice.

\[
\nabla_{\Sigma} \mathbb{E}_{q(z|\mathbf{z},\mathbf{\mu},\Sigma)} [h(\mathbf{z})] = \int q(w) [\nabla_{\Sigma} \mathcal{N}(\mathbf{z}|\mathbf{\mu}, w \Sigma)] h(\mathbf{z}) dw d\mathbf{z}
\]

\[
= \frac{1}{2} \int \int q(w) [w \nabla_{z_j} \nabla_{z_k} \mathcal{N}(\mathbf{z}|\mathbf{\mu}, w \Sigma)] h(\mathbf{z}) dw d\mathbf{z}
\]

\[
= \frac{1}{2} \int q(w) w \int [\nabla_{z_j} \mathcal{N}(\mathbf{z}|\mathbf{\mu}, w \Sigma)] h(\mathbf{z}) d\mathbf{z} dw |_{z_j=+\infty}^{z_j=-\infty} - \frac{1}{2} \int \int q(w) \nabla_{z_k} \mathcal{N}(\mathbf{z}|\mathbf{\mu}, w \Sigma) \nabla_{z_j} h(\mathbf{z}) d\mathbf{z} dw
\]

\[
= -\frac{1}{2} \int q(w) \int \nabla_{z_k} \mathcal{N}(\mathbf{z}|\mathbf{\mu}, w \Sigma) \nabla_{z_j} h(\mathbf{z}) d\mathbf{z} dw
\]

\[
= -\frac{1}{2} \int q(w) w \left\{ \int \mathcal{N}(\mathbf{z}|\mathbf{\mu}, w \Sigma) \nabla_{z_j} h(\mathbf{z}) d\mathbf{z}_{-k} |_{z_k=+\infty}^{z_k=-\infty} - \int \mathcal{N}(\mathbf{z}|\mathbf{\mu}, w \Sigma) \nabla_{z_k} \nabla_{z_j} h(\mathbf{z}) d\mathbf{z} \right\} dw
\]

\[
= -\frac{1}{2} \int \left[ u(\mathbf{z}_k, \mathbf{z}_{-k}) \nabla_{z_j} h(\mathbf{z}) d\mathbf{z}_{-k} |_{z_k=+\infty}^{z_k=-\infty} \right] + \frac{1}{2} \int q(w) \int \mathcal{N}(\mathbf{z}|\mathbf{\mu}, w \Sigma) \nabla_{z_k} \nabla_{z_j} h(\mathbf{z}) d\mathbf{z} dw
\]

\[
= \frac{1}{2} \int u(\mathbf{z}) \nabla_{z_k} \nabla_{z_j} h(\mathbf{z}) d\mathbf{z}
\]

We use integration by part twice to move from Eq. 128 to Eq. 130 and from Eq. 131 to Eq. 132. The first term in Eq. 130 is 0 since \(|z_k h(\mathbf{z})|\) is integrable. The first term in Eq. 133 is 0 since \(|z_k h(\mathbf{z})|\) and \(|\nabla_{z_j} h(\mathbf{z})|\) are integrable.

Using Lemma 8, the update for \(q(\mathbf{z}|w)\) can be expressed as

\[
\mathbf{S} \leftarrow (1 - \beta_1) \mathbf{S} + \beta_1 \mathbb{E}_{q(z)} \left[ \frac{a}{(a + d/2 - 1)} \left( 1 + (2a)^{-1} (\mathbf{z} - \mathbf{\mu})^T \Sigma^{-1} (\mathbf{z} - \mathbf{\mu}) \right) \left( \sum_{n=1}^{N} \nabla^2 h_n(\mathbf{z})/N \right) \right]
\]

\[
\mathbf{\mu} \leftarrow \mathbf{\mu} - \beta_1 (\mathbf{S} + \mathbf{I}/N)^{-1} \left( \mathbb{E}_{q(z)} \left[ \sum_{n=1}^{N} \nabla h_n(\mathbf{z})/N \right] + \mathbf{\mu}/N \right)
\]

Using MC approximation, we have the following update

\[
\mathbf{S} \leftarrow (1 - \beta_1) \mathbf{S} + \beta_1 \frac{a}{(a + d/2 - 1)} \left( 1 + (2a)^{-1} (\mathbf{z}_s - \mathbf{\mu})^T \Sigma^{-1} (\mathbf{z}_s - \mathbf{\mu}) \right) \left( \sum_{n=1}^{N} \nabla^2 h_n(\mathbf{z})/N \right)
\]

\[
\mathbf{\mu} \leftarrow \mathbf{\mu} - \beta_1 (\mathbf{S} + \mathbf{I}/N)^{-1} \left( \sum_{n=1}^{N} \nabla h_n(\mathbf{z})/N + \mathbf{\mu}/N \right)
\]
Similarly, the update for \( q(w) \) can be expressed as

\[
a \leftarrow (1 - \beta_2)a + \beta_2 \left( a_0 - \frac{N}{\nabla_a \mathbb{E}_{q(w)}[\phi_w(w)]} \sum_{n=1}^{N} \nabla_a \mathbb{E}_{q(z,w)}[h_n(z)] / N \right) \tag{139}
\]

where

\[
\nabla_a \mathbb{E}_{q(w)}[\phi_w(w)] = - \int \text{InvGam}(w|a, a) (\nabla_w \left[w^{-1} + \log w\right]) (\nabla_a w) dw \tag{140}
\]

\[
\nabla_a \mathbb{E}_{q(z,w)}[h_n(z)] = \int \int \text{InvGam}(w|a, a) (\nabla_w \mathcal{N}(\mathbf{z}|\mu, w \mathbf{\Sigma})) (\nabla_a w) h_n(\mathbf{z}) dwdzd\mathbf{z} \tag{141}
\]

Using one MC sample, we have

\[
\nabla_a \mathbb{E}_{q(w)}[\phi_w(w)] \approx (w_x^{-2} - w_x^{-1}) \nabla_a w_x \tag{142}
\]

\[
\nabla_a \mathbb{E}_{q(z,w)}[h_n(z)] \approx \int (\nabla_{w_x} \mathcal{N}(\mathbf{z}|\mu, w \mathbf{\Sigma})) (\nabla_a w_x) h_n(\mathbf{z}) dz \tag{143}
\]

\[
= \int \text{Tr} \left( \mathbf{\Sigma} \nabla_z \mathcal{N}(\mathbf{z}|\mu, \hat{\mathbf{\Sigma}}) \right) (\nabla_a w_x) h_n(\mathbf{z}) dz \tag{144}
\]

\[
= \nabla_a w_x \text{Tr} \left( \mathbf{\Sigma} \nabla_z \mathbb{E}_{q(z\mid \mu, \mathbf{\Sigma})}[h_n(\mathbf{z})] \right) \tag{145}
\]

\[
= \frac{1}{2} \nabla_a w_x \text{Tr} \left( \mathbf{\Sigma} \nabla_z^2 h_n(\mathbf{z}) \right) \tag{146}
\]

\[
\approx \frac{1}{2} \nabla_a w_x \text{Tr} \left( \mathbf{\Sigma} \nabla_z^2 h_n(\mathbf{z}) \right), \tag{147}
\]

where \( \hat{\mathbf{\Sigma}} = w_x \mathbf{\Sigma} \) and we use the Price’s theorem \( \nabla_{\mathbf{\Sigma}} \mathbb{E}_{q(z\mid \mu, \mathbf{\Sigma})}[h_n(\mathbf{z})] = \frac{1}{2} \mathbb{E}_{q(z\mid \mu, \mathbf{\Sigma})}[\nabla_{\mathbf{\Sigma}}^2 h_n(\mathbf{z})] \).

**Appendix D. Updates for Matrix-Variate Gaussian Distribution**

We first show that MVG is a multi-linear exponential-family distribution.

**Lemma 9** Matrix Gaussian distribution is a member of the multi-linear exponential family

**Proof** Let \( \mathbf{\Lambda}_1 = \mathbf{W}, \mathbf{\Lambda}_2 = \mathbf{U}^{-1}, \) and \( \mathbf{\Lambda}_3 = \mathbf{V}^{-1}. \) The distribution on \( \mathbf{Z} \in \mathbb{R}^{d \times p} \) can be expressed as follows.

\[
\mathcal{MN}(\mathbf{Z}|\mathbf{W}, \mathbf{U}, \mathbf{V}) = (2\pi)^{-dp/2} \exp \left\{ -\frac{1}{2} \text{Tr} \left( \mathbf{V}^{-1}(\mathbf{Z} - \mathbf{W})^T \mathbf{U}^{-1}(\mathbf{Z} - \mathbf{W}) \right) - (d/2 \log \det(\mathbf{V}) + p/2 \log \det(\mathbf{U})) \right\}
\]

\[
= (2\pi)^{-dp/2} \exp \left\{ \text{Tr} \left( \mathbf{\Lambda}_3 (-\frac{1}{2} \mathbf{Z} + \mathbf{\Lambda}_1)^T \mathbf{\Lambda}_2 \mathbf{Z} \right) \right. 
- \frac{1}{2} \left[ \text{Tr} \left( \mathbf{\Lambda}_3 \mathbf{\Lambda}_1^T \mathbf{\Lambda}_2 \mathbf{\Lambda}_1 \right) + d \log \det(\mathbf{\Lambda}_3) + p \log \det(\mathbf{\Lambda}_2) \right] \right\}. \tag{149}
\]

The function \( \text{Tr} \left( \mathbf{\Lambda}_3 (-\frac{1}{2} \mathbf{Z} + \mathbf{\Lambda}_1)^T \mathbf{\Lambda}_2 \mathbf{Z} \right) \) is linear with respect each \( \mathbf{\Lambda}_j \) given others. \( \blacksquare \)
We now derive the NGVI update using our new expectation parameterization. We can obtain function $\phi_1$, $\phi_2$, and $\phi_3$ from the multi-linear function

$$f(Z, \Lambda) := \text{Tr} \left( \Lambda_3 \left( -\frac{1}{2} Z + \Lambda_1 \right)^T \Lambda_2 Z \right).$$  \hspace{1cm} (150)

For example, we can obtain function $\phi_1$ from $f(Z, \Lambda)$ as shown below:

$$f(Z, \Lambda) = \langle \Lambda_1, \frac{1}{2} \text{Tr} \left( \Lambda_3 Z^T \Lambda_2 Z \right) \rangle$$  \hspace{1cm} (151)

Similarly, we can obtain functions $\phi_2$ and $\phi_3$. The corresponding expectation parameters of the Matrix Gaussian distribution can then be derived as below:

$$M_1 = E_{MN(Z|W,U,V)}[\Lambda_2 Z \Lambda_3] = \Lambda_2 \Lambda_1 \Lambda_3$$  \hspace{1cm} (152)

$$M_2 = E_{MN(Z|W,U,V)} \left[ -\frac{1}{2} Z \Lambda_3 Z^T + Z \Lambda_3 \Lambda_1^T \right] = \frac{1}{2} (\Lambda_1 \Lambda_3 \Lambda_1^T - p \Lambda_2^{-1})$$  \hspace{1cm} (153)

$$M_3 = E_{MN(Z|W,U,V)} \left[ -\frac{1}{2} Z^T \Lambda_2 Z + \Lambda_1^T \Lambda_2 Z \right] = \frac{1}{2} (\Lambda_1^T \Lambda_2 - d \Lambda_3^{-1})$$  \hspace{1cm} (154)

We can then compute the gradient with respect to the expectation parameters using chain-rule:

$$\nabla_{M_1} E_{q(z|\Lambda)}[h(Z)] = (\Lambda_2)^{-1} \nabla \delta E_{MN(Z|W,U,V)}[h(Z)] (\Lambda_3)^{-1}$$  \hspace{1cm} (155)

$$\nabla_{M_2} E_{q(z|\Lambda)}[h(Z)] = -\frac{2}{p} \nabla \delta E_{MN(Z|W,U,V)}[h(Z)]$$  \hspace{1cm} (156)

$$\nabla_{M_3} E_{q(z|\Lambda)}[h(Z)] = -\frac{2}{d} \nabla \delta E_{MN(Z|W,U,V)}[h(Z)]$$  \hspace{1cm} (157)

We will now express the gradients in terms of the gradient of the function $h(Z)$. This leads to a simple update because gradient of $h(Z)$ can be obtained using automatic gradients (or backpropagation when using a neural network). Let $z = \text{vec}(Z)$ and $Z = \text{Mat}(z)$. The distribution can be re-expressed as a multivariate Gaussian distribution $\mathcal{N}(z|\mu, \Sigma)$, where $\mu = \text{vec}(W)$, $\Sigma = V \otimes U$, and $\otimes$ denotes the Kronecker product. Furthermore, the lower bound can be re-expressed as $E_{\mathcal{N}(z|\mu, \Sigma)}[\hat{h}(z)]$, where $\hat{h}(z) = h(Z)$. We make use of the Bonnet’s and Price’s theorems (Opper and Archambeau, 2009):

$$\nabla_{\mu} E_{\mathcal{N}(z|\mu, \Sigma)}[\hat{h}(z)] = E_{\mathcal{N}(z|\mu, \Sigma)}[\nabla_{\mu} \hat{h}(z)]$$  \hspace{1cm} (158)

$$\nabla_{\Sigma} E_{\mathcal{N}(z|\mu, \Sigma)}[\nabla_{\mu} \hat{h}(z)] = \frac{1}{2} E_{\mathcal{N}(z|\mu, \Sigma)}[\nabla_{\mu}^2 \hat{h}(z)]$$  \hspace{1cm} (159)
These identities can be used to express the gradient with respect to the expectation parameters in terms of the gradient with respect to $Z$:

$$
\nabla_W \mathbb{E}_{MN(z|w,u,v)} [h(Z)] = \text{Mat} \left( \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)} \left[ \nabla_z \hat{h}(z) \right] \right)
$$

$$
= \mathbb{E}_{MN(z|w,u,v)} [\nabla_Z h(Z)]
$$

$$
\nabla_U \mathbb{E}_{MN(z|w,u,v)} [h(Z)] = (\nabla_U \Sigma) \nabla_Z \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)} \left[ \nabla_z \hat{h}(z) \right]
$$

$$
= \frac{1}{2} (\nabla_U \Sigma) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)} \left[ \nabla_z^2 \hat{h}(z) \right]
$$

$$
\approx \frac{1}{2} (\nabla_U \Sigma) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)} \left[ \nabla_z \hat{h}(z) \nabla_z \hat{h}(z)^T \right]
$$

$$
= \frac{1}{2} \mathbb{E}_{MN(z|w,u,v)} \left[ \nabla_Z h(Z) V \nabla_Z h(Z)^T \right]
$$

$$
\nabla_V \mathbb{E}_{MN(z|w,u,v)} [h(Z)] = (\nabla_V \Sigma) \nabla_Z \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)} \left[ \nabla_z \hat{h}(z) \right]
$$

$$
= \frac{1}{2} (\nabla_V \Sigma) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)} \left[ \nabla_z^2 \hat{h}(z) \right]
$$

$$
\approx \frac{1}{2} (\nabla_V \Sigma) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)} \left[ \nabla_z \hat{h}(z) \nabla_z \hat{h}(z)^T \right]
$$

$$
= \frac{1}{2} \mathbb{E}_{MN(z|w,u,v)} \left[ \nabla_Z h(Z)^T U \nabla_Z h(Z) \right].
$$

To avoid computation of the Hessian, we have used the Gauss-Newton approximation in Eq 164 and Eq. 168 (Khan et al., 2018).

We choose the step-size as $\beta = \{\beta_1, p\beta_2, d\beta_2\}$. The update with the Gauss-Newton approximation can be expressed as

$$
\Lambda_1 \leftarrow \Lambda_1 - \beta_1 (\Lambda_2)^{-1} \mathbb{E}_{MN(z|w,u,v)} \left[ \nabla_Z h(Z) \right] (\Lambda_3)^{-1}
$$

$$
\Lambda_2 \leftarrow \Lambda_2 + \beta_2 \mathbb{E}_{MN(z|w,u,v)} \left[ \nabla_Z h(Z) V \nabla_Z h(Z)^T \right]
$$

$$
\Lambda_3 \leftarrow \Lambda_3 + \beta_2 \mathbb{E}_{MN(z|w,u,v)} \left[ \nabla_Z h(Z)^T U \nabla_Z h(Z) \right]
$$

We can re-express these in terms of $\{W, U^{-1}, V^{-1}\}$ to get the final updates:

$$
W \leftarrow W - \beta_1 U \mathbb{E}_{MN(z|w,u,v)} \left[ \nabla_Z h(Z) \right] V
$$

$$
(U)^{-1} \leftarrow (U)^{-1} + \beta_2 \mathbb{E}_{MN(z|w,u,v)} \left[ \nabla_Z h(Z) V \nabla_Z h(Z)^T \right]
$$

$$
(V)^{-1} \leftarrow (V)^{-1} + \beta_2 \mathbb{E}_{MN(z|w,u,v)} \left[ \nabla_Z h(Z)^T U \nabla_Z h(Z) \right]
$$

**Appendix E. Connection to Block Mirror Descent**

As discussed in Section 4, the following theorem shows that the block natural-gradient update is indeed a block mirror descent.

**Theorem 10** *The block natural-gradient update is a block mirror descent.*

In fact, we can perform block natural-gradient update for the class of mixture of exponential family. The following lemmas give the proof-sketch of the theorem. Recall that $j$ is the block index of a variational distribution studied in this paper. We first define the function $\Psi_j(\Lambda_j) = \mathbb{E}_{\mathcal{N}(\Lambda_j, A_j)} [A_j(\cdot)]$, where $A_j(\cdot)$ is the log-partition function of the distribution at
block $j$. As shown in the following examples, $\Psi_j(\lambda_j)$ is convex w.r.t. $\lambda_j$ given $\lambda_{-j}$ is known. If $\Psi_j(\lambda_j)$ is strictly convex w.r.t. $\lambda_j$, we say the minimality holds. We then define the convex conjugate of function $\Psi_j$ as below.

$$\Psi_j^*(x) := \sup_{\eta} v_j(x, \eta),$$

(176)

where $v_j(\eta) = \langle x, \eta \rangle - \Psi_j(\eta)$.

**Example 4 (Mixture of exponential family distribution)** The variational distribution is a mixture of exponential family distribution: $q(z, w | \lambda_z, \lambda_w)$. When $j = w$, the log-partition function is $A_j(\lambda_j) = A_w(\lambda_w)$. The expectation parameter is $m_j = E_q[\phi_w(w)]$. Since $q(w)$ is an exponential family distribution when $j = w$, we know that $A_w(\lambda_w)$ is convex w.r.t. $\lambda_w$. Furthermore, $\Psi_j(\lambda_j) = A_w(\lambda_w)$ is convex w.r.t. $\lambda_j$ given $\lambda_{-j}$ is known. The minimal condition implies that $q(w)$ is a minimal exponential family distribution. When $j = z$, the log-partition function is $A_j(\lambda_j) = A_z(\lambda_z, w)$. The expectation parameter is $m_j = E_q[\phi_z(z, w)]$. According Eq. 47, we know that $A_z(\lambda_z, w)$ is convex w.r.t. $\lambda_j$ for any valid $w$, which implies that $\Psi_j(\lambda_j) = E_{q(w)}[A_z(\lambda_z, w)]$ is convex w.r.t. $\lambda_j$ given $\lambda_{-j}$ is known. If Lemma 5 holds, we know that the minimal condition holds.

**Example 5 (Multi-linear exponential family distribution)** The variational distribution is a multi-linear exponential family distribution: $q(z|\lambda_1, \ldots, \lambda_N)$. When $j \in \{1, \ldots, N\}$, the log-partition function is $A_j(\lambda_j) = A_z(\lambda_j, \lambda_{-j})$. The expectation parameter is $m_j = E_q[\phi_j(z, \lambda_{-j})]$. Since $\lambda_{-j}$ is known, we know that $q(z|\lambda_j, \lambda_{-j}) = q_j(z|\lambda_j)$ becomes a one-parameter exponential family distribution. We know that $\Psi_j(\lambda_j) = A_z(\lambda_j, \lambda_{-j})$ is convex w.r.t. $\lambda_j$ given that $\lambda_{-j}$ is known. The minimal condition implies that for all $j \in \{1, \ldots, N\}$, $q_j(z|\lambda_j)$ is a minimal exponential family distribution.

**Lemma 11** If the minimality holds, we have

$$\Psi_j^*(m_j) = \sup_{\eta} v_j(m_j, \eta)$$

(177)

$$= \langle m_j, \lambda_j \rangle - \Psi_j(\lambda_j),$$

(178)

where $\lambda_j$ and $m_j$ are the natural parameter and the expectation parameter of the variational distribution at block $j$.

**Proof** We give the proof for the case when the variational distribution is a mixture of exponential family distribution. It is easy to show the lemma holds in the multi-linear exponential family case. When $j = w$, since $q(w)$ is an exponential family distribution, we know that

$$m_w = \nabla_{\lambda_w} A_w(\lambda_w),$$

(179)

which implies that $m_w - \nabla_{\lambda_w} A_w(\lambda_w) = 0$.

Note that the gradient of $v_j(m_w, \eta)$ is

$$\nabla_\eta v_j(m_w, \eta) = \nabla_\eta \langle m_w, \eta \rangle - \Psi_j(\eta) = m_w - \nabla_\eta A_w(\eta).$$

(180)
When $\eta = \lambda_w$, the gradient is $0$. Since $\Psi_j(\lambda_w)$ is convex w.r.t. $\lambda_w$, we know that $\lambda_w$ is an optimal solution. Since the minimality holds, we know that $\Psi_j(\lambda_w)$ is strictly convex, which implies that the solution is unique. Therefore,\
$$\frac{\Psi_w^*(m_w)}{\psi^*(m_j)} = \langle m_w, \lambda_w \rangle - A_w(\lambda_w).$$

When $j = z$, by Eq. 21, we know that\
$$m_z = \nabla_{\lambda_z} \mathbb{E}_{q(w)} [A_z(\lambda_z, w)],$$
which implies that $m_z - \nabla_{\lambda_z} \mathbb{E}_{q(w)} [A_z(\lambda_z, w)] = 0$. Similarly, since the minimality holds, when $\lambda_w$ is known, we know $\frac{\Psi_w^*(m_z)}{\psi^*(m_j)} = \langle m_z, \lambda_z \rangle - \mathbb{E}_q [A_z(\lambda_z, w)].$

Now, we show that the connection between block mirror descent update and the natural-gradient update.

**Lemma 12** If the minimality holds, given that $\lambda_{-j} = \lambda_{-j}^k$, the following identity is true.\
$$\mathbb{B}_{\psi_j}(m_j \| m_j^k) = \mathbb{D}_{KL}(q(\cdot | \lambda_j, \lambda_{-j}^k) \| q(\cdot | \lambda_j^k, \lambda_{-j}^k))$$
$$= \Psi_j(\lambda_j^k) - \Psi_j(\lambda_j) - \left< \nabla_{\lambda_j} \Psi_j(\lambda_j), (\lambda_j^k - \lambda_j) \right>,$$
where $\mathbb{B}(\cdot)$ denotes the Bregman divergence.

**Proof**

We give the proof for the case when the variational distribution is a mixture of exponential family distribution. It is easy to show the lemma holds in the multi-linear exponential family case. According to the duality of the Bregman divergence and Lemma 11, we have\
$$\mathbb{B}_{\psi_j}(m_j \| m_j^k) = \mathbb{B}_{\psi_j}(\lambda_j^k \| \lambda_j).$$

When $j = w$, $\Psi_w(\lambda_w) = A_w(\lambda_w)$. Therefore, we have the following results.

$$\mathbb{D}_{KL}(q(z, w|\lambda_w, \lambda_{-w}^k) \| q(z, w|\lambda_w^k, \lambda_{-w}^k))$$
$$= \mathbb{D}_{KL}(q(w|\lambda_w) \| q(w|\lambda_w^k))$$
$$= \mathbb{E}_{q(w|\lambda_w)} [\phi_w(w), (\lambda_w - \lambda_w^k)] - A_w(\lambda_w) + A_w(\lambda_w^k)$$
$$= \left< m_w, (\lambda_w - \lambda_w^k) \right> - A_w(\lambda_w) + A_w(\lambda_w^k)$$
$$= \Psi_w(\lambda_w^k) - \Psi_w(\lambda_w) - \left< \nabla_{\lambda_w} \Psi_w(\lambda_w), (\lambda_w^k - \lambda_w) \right>$$
$$= \mathbb{B}_{\psi}(\lambda_w^k \| \lambda_w),$$
where $m_w = \nabla_{\lambda_w} A_w(\lambda_w) = \nabla_{\lambda_w} \Psi_w(\lambda_w)$.
When \( j = z \), \( \Psi_z(\lambda_z) = \mathbb{E}_{\eta(z|\lambda_z)} [A_z(\lambda_z, w)] \). Therefore, we have the following results.

\[
\mathbb{D}_{KL}(q(z, w|\lambda_z, \lambda^k_w) \| q(z, w|\lambda^k_z, \lambda^k_w)) = \mathbb{D}_{KL}(q(w|\lambda^k_z)q(z|w, \lambda_z) \| q(w|\lambda^k_z)q(z|w, \lambda^k_z)) = \left( \mathbb{E}_{\eta(z|\lambda_z)} [\phi_z(z, w)], \left( \lambda_z - \lambda^k_z \right) \right) - \mathbb{E}_{\eta(z|\lambda_z)} [A_z(\lambda_z, w)] + \mathbb{E}_{\eta(z|\lambda_z)} \left[ A_z(\lambda^k_z, w) \right]
\]

(191)

(192)

(193)

(194)

(195)

(196)

where \( m_z = \nabla_{\lambda_z} \Psi_z(\lambda_z) \) due to Eq. 181.

**Lemma 13** If the minimality holds, given \( \lambda_{-j} = \lambda^k_{-j} \), a valid step-size \( \beta^k_j \), and the following optimization problem

\[
\max_{m_j} \left\{ m_j, \nabla_{m_j} \mathcal{L}(\lambda) \right\} - \frac{1}{\beta^k_j} \mathbb{E}_{\Psi_j} (m_j||m^k_j)
\]

(197)

the solution of the optimization problem is

\[
\lambda^k_{j+1} = \lambda^k_j + \beta^k_j \nabla_{m_j} \mathcal{L}(\lambda)
\]

(198)

**Proof** According to Lemma 12, Eq. 197 can be also re-written as

\[
\left\{ m_j, \nabla_{m_j} \mathcal{L}(\lambda) \right\} - \frac{1}{\beta^k_j} \mathbb{E}_{\Psi_j} (m_j||m^k_j)
\]

(199)

(200)

(201)

Similarly, by taking the gradient of Eq. 201 w.r.t. \( m_j \) to be 0, we have

\[
0 = \nabla_{m_j} \left\{ m_j, \nabla_{m_j} \mathcal{L}(\lambda) \right\} - \frac{1}{\beta^k_j} \left\{ m_j, \left( \lambda_j - \lambda^k_j \right) \right\} - \Psi_j(\lambda_j) + \Psi_j(\lambda^k_j)
\]

(202)

\[
= \nabla_{m_j} \mathcal{L}(\lambda) - \frac{1}{\beta^k_j} \left[ \left( \lambda_j - \lambda^k_j \right) + \left( \nabla_{m_j} \lambda_j \right) m_j - \nabla_{m_j} \Psi_j(\lambda_j) \right]
\]

(203)

\[
= \nabla_{m_j} \mathcal{L}(\lambda) - \frac{1}{\beta^k_j} \left[ \left( \lambda_j - \lambda^k_j \right) + \left( \nabla_{m_j} \lambda_j \right) m_j - \left( \nabla_{m_j} \lambda_j \right) \nabla_{\lambda_j} \Psi_j(\lambda_j) \right]
\]

(204)
Finally, we obtain the following solution by solving Eq. 204
\[ \lambda_{j}^{k+1} = \lambda_{j}^{k} + \beta_{j}^{k} \nabla_{m_{j}} L(\lambda) \tag{205} \]

Appendix F. A Variant of Block Mirror Descent

Now, we discuss a variant of block mirror descent. Instead of sequentially performing update block by block, we can simultaneously update all blocks at each iteration. This variant is known as a Jacobi update. In the case of mixture of exponential family distribution, we have the following theorem.

**Theorem 14**  The Jacobi update is a natural-gradient update when the variational distribution is a member of the mixture of exponential family distribution.

**Proof**  When the variational distribution is a mixture of exponential family distribution, according to Lemma 13, the Jacobi update is
\[ \lambda_{w}^{k+1} = \lambda_{w}^{k} + \beta_{w} \nabla_{m_{w}} L(\lambda) \tag{206} \]
\[ \lambda_{z}^{k+1} = \lambda_{z}^{k} + \beta_{z} \nabla_{m_{z}} L(\lambda), \tag{207} \]
where \( \beta_{w} = \beta_{z} = \beta \). It is easy to see that the Jacobi update is the natural-gradient update due to Theorem 1.

Note that Lemma 3 shows that the Fisher information matrix is block-diagonal in the case of mixture of exponential family distribution. We can generalize Theorem 15 and applied it to the multi-linear exponential family which gives the following theorem.

**Theorem 15**  The Jacobi update is a scaled gradient descent update where the scaling matrix is a block-diagonal approximated Fisher information matrix, if the variational distribution is either a mixture of exponential family distribution or a multi-linear exponential family distribution.

**Proof**  When the variational distribution is a mixture of exponential family distribution, by Theorem 14, the Jacobi update is a natural-gradient update. Recall that natural-gradient update is a scaled gradient update, where the scaling matrix is the Fisher information matrix. By Lemma 3, we know that the Fisher information matrix is block-diagonal.

Now, we consider the case when the variational distribution is a multi-linear exponential distribution. According to according to Lemma 13, we know the Jacobi update is
\[ \lambda_{j}^{k+1} = \lambda_{j}^{k} + \beta_{j} \nabla_{m_{j}} L(\lambda) \quad j \in \{1, \ldots, N\}. \tag{208} \]
Since $\lambda^k_{-j}$ is known, by the chain rule, we have

$$\hat{\nabla}_{m_j} \tilde{L}(\lambda) = \left(\hat{\nabla}_{\lambda_j^k} m_j\right)^{-1} \hat{\nabla}_{\lambda_j^k} L(\lambda)$$

(209)

$$= \left(\hat{\nabla}_{\lambda_j^k} E_{q(z|m_j)} \left[ \phi_j(z, \lambda^k_{-j}) \right] \right)^{-1} \hat{\nabla}_{\lambda_j^k} L(\lambda)$$

(210)

Recall that $q(z|\lambda_j, \lambda^k_{-j})$ is an one-parameter exponential family distribution given that $\lambda_{-j}$ is known. We have

$$0 = E_{q(z|m_j)} \left[ \nabla_{\lambda_j} \log q(z|\lambda_j, \lambda^k_{-j}) \right]$$

(211)

$$= E_{q(z|m_j)} \left[ \nabla_{\lambda_j} \left[ \langle \phi_j(z, \lambda^k_{-j}), \lambda_j \rangle + r_j(\lambda^k_{-j}) - A_z(\lambda_j, \lambda^k_{-j}) \right] \right]$$

(212)

$$= E_{q(z|m_j)} \left[ \phi_j(z, \lambda^k_{-j}) - \nabla_{\lambda_j} A_z(\lambda_j, \lambda^k_{-j}) \right]$$

(213)

$$= E_{q(z|m_j)} \left[ \phi_j(z, \lambda^k_{-j}) - \nabla_{\lambda_j} A_z(\lambda_j, \lambda^k_{-j}) \right]$$

(214)

Let’s define

$$F_{\lambda_j} = E_{q(z|m_j)} \left[ \nabla^2_{\lambda_j} A_z(\lambda_j, \lambda^k_{-j}) \right] = \nabla^2_{\lambda_j} A_z(\lambda_j, \lambda^k_{-j}).$$

(215)

Due to Eq 214, we know that

$$\hat{\nabla}_{\lambda_j^k} E_{q(z|m_j)} \left[ \phi_j(z, \lambda^k_{-j}) \right] = F_{\lambda_j^k}.$$ 

(216)

Therefore, the Jacobi update is

$$\lambda_j^{k+1} = \lambda_j^k + \beta \left( F_{\lambda_j^k} \right)^{-1} \hat{\nabla}_{\lambda_j^k} L(\lambda) \quad j \in \{1, \ldots, N\}.$$ 

(217)

Recall that the Fisher information matrix of $q(z|\lambda)$ is

$$F_{\lambda} = -E_{q(z|m)} \left[ \nabla^2_{\lambda} \log q(z|\lambda) \right]$$

(218)

$$= -E_{q(z|m)} \left[ \begin{array}{ccc} \nabla^2_{\lambda_1} \log q(z|\lambda) & \ldots & \nabla^2_{\lambda_N} \log q(z|\lambda) \\ \vdots & \ldots & \vdots \\ \nabla^2_{\lambda_N} \log q(z|\lambda) & \ldots & \nabla^2_{\lambda_1} \log q(z|\lambda) \end{array} \right]$$

(219)

$$= -E_{q(z|m)} \left[ \begin{array}{ccc} \nabla^2_{\lambda_1} A_z(\lambda) & \ldots & \nabla^2_{\lambda_N} \log q(z|\lambda) \\ \vdots & \ldots & \vdots \\ \nabla^2_{\lambda_N} \log q(z|\lambda) & \ldots & \nabla^2_{\lambda_1} \log q(z|\lambda) \end{array} \right]$$

(220)

$$= \left[ \begin{array}{cc} F_{\lambda_1^k} & \ldots & \nabla^2_{\lambda_1^k} \log q(z|\lambda) \\ \ldots & \ldots & \ldots \\ \nabla^2_{\lambda_N} \log q(z|\lambda) & \ldots & F_{\lambda_N^k} \end{array} \right]$$

(221)

From Eq 221, it is easy to see that the Jacobi update is a scaled gradient descent update where the scaling matrix at iteration $k$ is the block approximation of the Fisher information.
Figure 2: This figure demonstrates a fast convergence of NGVI over BBVI to approximate the posterior distribution of Bayesian logistic regression using a scale mixture of Gaussians with full covariance matrix. The plot shows the ELBO of the whole training set obtained using $5 \times 10^4$ MC samples. For both algorithms, we used mini-batches by using 10 MC samples to compute stochastic approximations. $M$ denotes the size of a mini-batch. For BBVI, we use the Adam optimizer. We can see that our method is faster than BBVI using Adam.

Finally, it is easy to check that the scaling matrix $\hat{F}_{\lambda k}$ is positive semi-definitive. If the minimality holds, it is positive definitive.

\[ \hat{F}_{\lambda k} := \begin{bmatrix} F_{\lambda_1^k} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & F_{\lambda_N^k} \end{bmatrix} \quad (222) \]

Appendix G. Results for Multivariate t-Distribution

Figure 2 shows a fast convergence of NGVI over BBVI to approximate the posterior distribution of Bayesian logistic regression with a Student’s t prior expressed as a scale mixture of Gaussians discussed at Eq. 92. For this model, we use a scale mixture of Gaussians with full covariance matrix as the variational distribution.