1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–26

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, schmidtm@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 $p(\mathcal{D}, \mathbf{z})$ to model the data $\mathcal{D}$ using a latent vector $\mathbf{z}$, a common choice to approximate the posterior is to employ an exponential-family approximation: $p(\mathbf{z}|\mathcal{D}) \approx q(\mathbf{z}|\boldsymbol{\lambda}_z) := h_z(\mathbf{z}) \exp\left[\langle \boldsymbol{\phi}_z(\mathbf{z}), \boldsymbol{\lambda}_z \rangle - A_z(\boldsymbol{\lambda}_z)\right]$, where q denotes the approximating distribution, $\boldsymbol{\lambda}_z$ is the natural parameter, $\boldsymbol{\phi}_w(\mathbf{z})$ are the sufficient statistics, $A_z(\boldsymbol{\lambda}_z)$ is the log-partition function, and h_z is the base measure. The approximation can be estimated

by maximizing the variational lower bound $\mathcal{L}(\boldsymbol{\lambda}_z)$. 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:

$$\text{BBVI: } \boldsymbol{\lambda}_z \leftarrow \boldsymbol{\lambda}_z + \alpha \nabla_{\lambda_z} \mathcal{L}(\boldsymbol{\lambda}_z), \qquad \text{NGVI: } \boldsymbol{\lambda}_z \leftarrow \boldsymbol{\lambda}_z + \beta \ [\mathbf{F}_z(\boldsymbol{\lambda}_z)]^{-1} \nabla_{\lambda_z} \mathcal{L}(\boldsymbol{\lambda}_z), \tag{1}$$

where ∇ denotes the gradient, $\mathbf{F}_z(\boldsymbol{\lambda}_z)$ is the Fisher information matrix (FIM) with respect to the natural parameters, and α, β 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 $\mathbf{m}_z(\boldsymbol{\lambda}_z) := \mathbb{E}_q\left[\boldsymbol{\phi}_z(\mathbf{z})\right] = \nabla_{\lambda_z} A(\boldsymbol{\lambda}_z)$. 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:

$$\text{NGVI for Exp-Family: } \boldsymbol{\lambda}_z \leftarrow \boldsymbol{\lambda}_z + \beta \ \nabla_{m_z} \mathcal{L}(\boldsymbol{\lambda}_z), \tag{2}$$

where the gradient is computed with respect to $\mathbf{m}_z$. When the gradient with respect to $\mathbf{m}_z$ is easier to compute than the gradient with respect to $\boldsymbol{\lambda}_z$, 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 $q(\mathbf{z}) := \int q(\mathbf{z}|\mathbf{w}) q(\mathbf{w}) d\mathbf{w}$ where $q(\mathbf{w})$ is a mixing distribution that takes an exponential-family form as shown in (3) and $q(\mathbf{z}|\mathbf{w})$ is the mixture component that takes the *conditional* exponential-family distribution as shown in (4).

$$q(\mathbf{w}|\boldsymbol{\lambda}_w) := h_w(\mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_w(\mathbf{w}), \boldsymbol{\lambda}_w \rangle - A_w(\boldsymbol{\lambda}_w)\right], \tag{3}$$

$$q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) := h_z(\mathbf{z}, \mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle - A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right]. \tag{4}$$

Note that, in $q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$, the sufficient statistics $\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})$ and log-partition $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ depend on $\mathbf{w}$. However, conditioned on w, the distribution is an exponential family parametrized by natural parameter $\boldsymbol{\lambda}_z$. 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 $q(\mathbf{z}, \mathbf{w})$. Denoting the set of natural parameters by $\boldsymbol{\lambda} := \{\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w\}$, we can define the following FIM: $\mathbf{F}_{wz}(\boldsymbol{\lambda}) := -\mathbb{E}_{q(z,w)}\left[\nabla^2_{\lambda} \log q(\mathbf{z}, \mathbf{w})\right]$. The FIM is defined by using the joint distribution of $\mathbf{z}$ and $\mathbf{w}$, and it is different from the one defined over the marginal distribution $q(\mathbf{z})$. We also assume the minimality holds in $q(\mathbf{w}|\boldsymbol{\lambda}_w)$ and $q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$ so that the FIM is invertible. The minimality is discussed in Appendix A.

We consider the natural gradient in the Riemannian metric defined by $\mathbf{F}_{wz}(\boldsymbol{\lambda})$. To simplify the natural-gradient update, we define the following expectation-parameters:

$$\mathbf{m}_z(\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w) := \mathbb{E}_{q(z|w)q(w)}\left[\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})\right], \qquad \mathbf{m}_w(\boldsymbol{\lambda}_w) := \mathbb{E}_{q(w)}\left[\boldsymbol{\phi}_w(\mathbf{w})\right]. \tag{5}$$

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

$$\boldsymbol{\lambda}_z \leftarrow \boldsymbol{\lambda}_z + \beta\ \nabla_{m_z} \mathcal{L}(\boldsymbol{\lambda}), \qquad \boldsymbol{\lambda}_w \leftarrow \boldsymbol{\lambda}_w + \beta\ \nabla_{m_w} \mathcal{L}(\boldsymbol{\lambda}). \tag{6}$$

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:* $q(\mathbf{z}) = \sum_{c=1}^{K} q(\mathbf{z}|w=c)q(w=c)$ *where* $q(\mathbf{z}|w=c) := \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ *and* $q(w=c) = \pi_c$ *with* $\sum_c \pi_c = 1$*. The natural-parameter* $\boldsymbol{\lambda}_z := \{\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c, -\frac{1}{2}\boldsymbol{\Sigma}_c^{-1}\}_{c=1}^{K}$ *and the sufficient statistics is* $\boldsymbol{\phi}_z(\mathbf{z}, w) := \{\mathbb{I}_c(w)\mathbf{z}, \mathbb{I}_c(w)\mathbf{z}\mathbf{z}^T\}_{c=1}^{K}$ *where* $\mathbb{I}_c(w)$ *is an indicator taking a value 1 when* $w = c$ *and 0 otherwise. The mixing distribution* $q(w)$ *is a categorical distribution with* $\boldsymbol{\lambda}_w = \{\log \frac{\pi_c}{\pi_K}\}_{c=1}^{K-1}$ *and* $\phi_w(w) := \{\mathbb{I}_c(w)\}_{c=1}^{K-1}$*. The parameters* $\mathbf{m}_z$ *for the c-th component are simply equal to the mean parameter of the Gaussian* $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ *multiplied by* π_c*, and* $\mathbf{m}_w := \{\pi_c\}_{c=1}^{K-1}$*. 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:* $q(\mathbf{z}) = \int q(\mathbf{z}|w)q(w)dw = \int \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\ \mathrm{InvGam}(w|a,a)dw$*. The natural parameter and the sufficient statistics of* $q(\mathbf{z}|w)$ *are* $\boldsymbol{\lambda}_z = \{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\}$ *and* $\phi_z(\mathbf{z}, w) = \{w^{-1}\mathbf{z}, w^{-1}\mathbf{z}\mathbf{z}^T\}$*. These quantities for* $q(w)$ *are* $\boldsymbol{\lambda}_w := a$ *and* $\phi_w(w) := -1/w - \log w$*. The expectation parameters defined in (5) are* $\mathbf{m}_z := \{\boldsymbol{\mu}, \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}\}$ *and* $m_w := -1 - \log a + \psi(a)$*, where* ψ *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 using methods described in Khan et al. (2018). The resulting update can be expressed as a perturbed Newton update. The update with respect to* 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* $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\mathrm{InvGam}(w|a,a)$ *indeed is a curved exponential family 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 $\boldsymbol{\lambda} := \{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_N\}$ where $\boldsymbol{\lambda}_j$ is the j-th block of parameters. We consider a family of distribution, which takes the following form: $q(\mathbf{z}|\boldsymbol{\lambda}) = h_z(\mathbf{z}) \exp\left[f\left(\mathbf{z}, \boldsymbol{\lambda}\right) - A_z(\boldsymbol{\lambda})\right]$. with $f(\cdot)$ being a multi-linear function

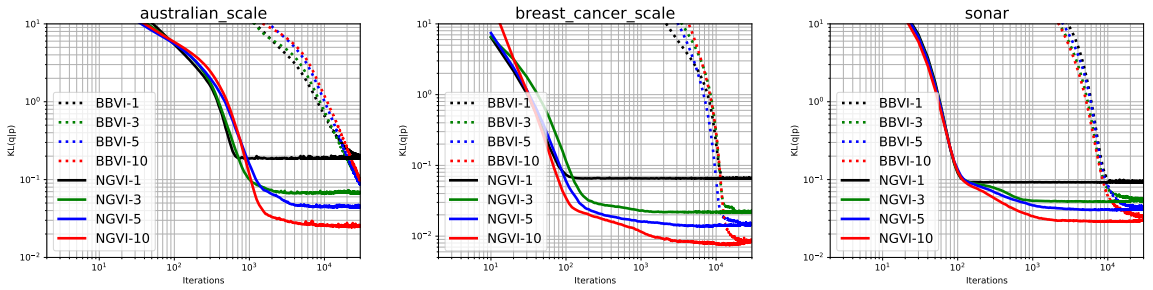

Figure 1: This figure demonstrates a fast convergence of NGVI over BBVI to approximate the posterior distribution of Bayesian logistic regression. We use a mixture of Gaussians with full covariance matrix as the approximating distribution. The number indicates the number of mixture components used. The plot shows the KL obtained using 10^6 MC samples, where p is the true posterior. For both algorithms, we used full-batches by using 20 MC samples to compute stochastic approximations. For BBVI, we use the Adam optimizer. We can see that our method is faster than BBVI using Adam. The main reason behind this is the simplicity of the natural-gradient update where we do not need to maintain a separate scaling vector for mean and covariance of the Gaussian, unlike Adam. This is similar to the Variational Adam method of Khan et al. (2018) and Zhang et al. (2018) which has a simpler update than Adam.

of $\boldsymbol{\lambda}_j$, i.e., for all j, there exist functions ϕ_j and r_j such that f is linear with respect to $\boldsymbol{\lambda}_j$, i.e., $f(\mathbf{z}, \boldsymbol{\lambda}) := \langle \boldsymbol{\lambda}_j, \phi_j(\mathbf{z}, \boldsymbol{\lambda}_{-j}) \rangle + r_j(\mathbf{z}, \boldsymbol{\lambda}_{-j})$, where $\boldsymbol{\lambda}_{-j}$ is the parameter vector containing all $\boldsymbol{\lambda}$ except $\boldsymbol{\lambda}_j$. 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 $\boldsymbol{\lambda}_j$. 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 $\mathbf{m}_j(\boldsymbol{\lambda}) := \mathbb{E}_{q(z)}[\phi_j(\mathbf{z}, \boldsymbol{\lambda}_{-j})]$. By defining an FIM for the j-th block as $\mathbf{F}_j(\boldsymbol{\lambda}) := -\mathbb{E}_{q(z)}\left[\nabla^2_{\lambda_j} \log q(\mathbf{z}|\boldsymbol{\lambda})\right]$, we can perform a *block* natural-gradient update: $\boldsymbol{\lambda}_j \leftarrow \boldsymbol{\lambda}_j + \beta_j \left[\mathbf{F}_j(\boldsymbol{\lambda})\right]^{-1} \nabla_{\lambda_j} \mathcal{L}(\boldsymbol{\lambda})$. Similarly, we assume the minimality holds for all blocks so that $\mathbf{F}_j(\boldsymbol{\lambda})$ is invertible. The following theorem establishes an update by using the expectation parameter $\mathbf{m}_j$.

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

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* $\mathcal{MN}(\mathbf{Z}|\mathbf{W}, \mathbf{U}, \mathbf{V})$ *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* $\mathcal{L}(\boldsymbol{\lambda}) = E_q[-h(\mathbf{Z})]$. *Under the Gauss-Newton approximation (Graves, 2011), the block natural-gradient update is,*

$$\mathbf{W} \leftarrow \mathbf{W} - \beta_1 \mathbf{U}\mathbf{G}\mathbf{V}, \qquad \mathbf{U}^{-1} \leftarrow \mathbf{U}^{-1} + \beta_2 \mathbf{G}\mathbf{V}\mathbf{G}^\top, \qquad \mathbf{V}^{-1} \leftarrow \mathbf{V}^{-1} + \beta_2 \mathbf{G}^\top \mathbf{U}\mathbf{G}, \quad (7)$$

where we sample $\mathbf{Z}$ *from the MVG distribution and evaluate the gradient* $\mathbf{G} := \nabla_Z h(\mathbf{Z})$. *These updates extend the Newton-like 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 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

Paul D Feigin. Conditional exponential families and a representation theorem for asympotic inference. *The Annals of Statistics*, pages 597–603, 1981.

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

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

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1842–1850, 2018.

James Hensman, Magnus Rattray, and Neil D Lawrence. Fast variational inference in the conjugate exponential family. In *Advances in neural information processing systems*, pages 2888–2896, 2012.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

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

A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. 11:3235–3268, 2011.

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

Mohammad Emtiyaz Khan and Didrik Nielsen. Fast yet simple natural-gradient descent for variational inference in complex models. *arXiv preprint arXiv:1807.04489*, 2018.

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2611–2620, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Percy Liang, Michael I Jordan, and Dan Klein. Learning from measurements in exponential families. In *Proceedings of the 26th annual international conference on machine learning*, pages 641–648. ACM, 2009.

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

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

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017.

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

Manfred Opper and Cédric Archambeau. The variational gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

Hugh Salimbeni, Stefanos Eleftheriadis, and James Hensman. Natural gradients in practice: Non-conjugate variational inference in gaussian process models. 2018.

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

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

G. Zhang, S. Sun, D. Duvenaud, and R. Grosse. Noisy natural gradient as variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Appendix A. Proof of Theorem 1

The NGVI update of (6) updates the parameters $\boldsymbol{\lambda}_z$ and $\boldsymbol{\lambda}_w$ separately. This is possible because the FIM is in fact block-diagonal. The following lemma states this result.

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

$$\mathbf{F}_{wz}(\boldsymbol{\lambda}) = \begin{bmatrix} \mathbf{F}_z(\boldsymbol{\lambda}) & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_w(\boldsymbol{\lambda}_w) \end{bmatrix}, \tag{8}$$

with the blocks defined as follows:

$$\mathbf{F}_z(\boldsymbol{\lambda}) := -\mathbb{E}_{q(z,w)}\left[\nabla^2_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right], \qquad \mathbf{F}_w(\boldsymbol{\lambda}_w) := -\mathbb{E}_{q(w)}\left[\nabla^2_{\lambda_w} \log q(\mathbf{w}|\boldsymbol{\lambda}_w)\right] \tag{9}$$

Proof By definition of FIM, we have

$$\mathbf{F}_{wz}(\boldsymbol{\lambda}) = -\mathbb{E}_{q(z,w|\lambda)}\left[\nabla^2_{\lambda} \log q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda})\right] \tag{10}$$

$$= -\mathbb{E}_{q(z,w|\lambda)}\begin{bmatrix} \nabla^2_{\lambda_z} \log q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda}) & \nabla_{\lambda_w}\nabla_{\lambda_z} \log q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda}) \\ \nabla_{\lambda_z}\nabla_{\lambda_w} \log q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda}) & \nabla^2_{\lambda_w} \log q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda}) \end{bmatrix} \tag{11}$$

$$= -\mathbb{E}_{q(z,w|\lambda)}\begin{bmatrix} \nabla^2_{\lambda_z} \left(\log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) + \log q(\mathbf{w}|\boldsymbol{\lambda}_w)\right) & \nabla_{\lambda_w}\nabla_{\lambda_z} \left(\log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) + \log q(\mathbf{w}|\boldsymbol{\lambda}_w)\right) \\ \nabla_{\lambda_z}\nabla_{\lambda_w} \left(\log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) + \log q(\mathbf{w}|\boldsymbol{\lambda}_w)\right) & \nabla^2_{\lambda_w} \left(\log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) + \log q(\mathbf{w}|\boldsymbol{\lambda}_w)\right) \end{bmatrix} \tag{12}$$

$$= -\mathbb{E}_{q(z,w|\lambda)}\begin{bmatrix} \nabla^2_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) & \mathbf{0} \\ \mathbf{0} & \nabla^2_{\lambda_w} \log q(\mathbf{w}|\boldsymbol{\lambda}_w) \end{bmatrix} \tag{13}$$

$$= -\begin{bmatrix} \mathbb{E}_{q(z,w|\lambda)}\left[\nabla^2_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right] & \mathbf{0} \\ \mathbf{0} & \mathbb{E}_{q(w|\lambda_w)}\left[\nabla^2_{\lambda_w} \log q(\mathbf{w}|\boldsymbol{\lambda}_w)\right] \end{bmatrix} \tag{14}$$

$$= \begin{bmatrix} \mathbf{F}_z(\boldsymbol{\lambda}) & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_w(\boldsymbol{\lambda}_w), \end{bmatrix} \tag{15}$$

where we can move from step 12 to step 13 since $\boldsymbol{\lambda}_z$ and $\boldsymbol{\lambda}_w$ are not tied. ■

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

Lemma 4 *The FIM matrix* $\mathbf{F}_z$ *is equal to the derivative of the expectation parameter* $\mathbf{m}_z(\boldsymbol{\lambda})$*:*

$$\mathbf{F}_z(\boldsymbol{\lambda}) := \nabla_{\lambda_z} \mathbf{m}_z(\boldsymbol{\lambda}) \tag{16}$$

Proof Recall that

$$\mathbb{E}_{q(z,w|\lambda)}\left[\nabla_{\lambda_z} \log q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda})\right] = \mathbb{E}_{q(z,w|\lambda)}\left[\nabla_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right] \tag{17}$$

$$= \mathbb{E}_{q(z,w|\lambda)}\left[\left[\nabla_{\lambda_z} q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right] / q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right] \tag{18}$$

$$= \mathbb{E}_{q(w|\lambda_w)}\left[\nabla_{\lambda_z} q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right] \tag{19}$$

$$= \nabla_{\lambda_z} \mathbb{E}_{q(z,w|\lambda)}\left[1\right] \tag{20}$$

$$= \mathbf{0} \tag{21}$$

Since $q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) = h_z(\mathbf{z}, \mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle - A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right]$, we have

$$
\begin{aligned}
\mathbf{0} &= \mathbb{E}_{q(z,w|\lambda)} \left[\nabla_{\lambda_z} \log q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda})\right] && (22)\\
&= \mathbb{E}_{q(z,w|\lambda)} \left[\nabla_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right] && (23)\\
&= \mathbb{E}_{q(z,w|\lambda)} \left[\nabla_{\lambda_z} \left(\log h_z(\mathbf{z}, \mathbf{w}) + \langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle - A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right)\right] && (24)\\
&= \mathbb{E}_{q(z,w|\lambda)} \left[\nabla_{\lambda_z} \left(\langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle - A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right)\right] && (25)\\
&= \mathbb{E}_{q(z,w|\lambda)} \left[\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})\right] - \mathbb{E}_{q(z,w|\lambda)} \left[\nabla_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] && (26)\\
&= \mathbf{m}_z - \mathbb{E}_{q(z,w|\lambda)} \left[\nabla_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] && (27)\\
&= \mathbf{m}_z - \mathbb{E}_{q(w|\lambda_w)} \left[\nabla_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] && (28)
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\mathbf{F}_z(\boldsymbol{\lambda}) &= -\mathbb{E}_{q(z,w)} \left[\nabla^2_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right] && (29)\\
&= -\mathbb{E}_{q(z,w|\lambda)} \left[\nabla^2_{\lambda_z} \left(\log h_z(\mathbf{z}, \mathbf{w}) + \langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle - A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right)\right] && (30)\\
&= \mathbb{E}_{q(z,w|\lambda)} \left[\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] && (31)\\
&= \mathbb{E}_{q(w|\lambda_w)} \left[\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] && (32)\\
&= \nabla_{\lambda_z} \mathbb{E}_{q(w|\lambda_w)} \left[\nabla_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] && (33)\\
&= \nabla_{\lambda_z} \mathbf{m}_z && (34)
\end{aligned}
$$

■

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

$$
\mathbf{F}_z(\boldsymbol{\lambda}) \left[\nabla_{m_z} \mathcal{L}(\boldsymbol{\lambda})\right] = \nabla_{\lambda_z} \mathbf{m}_z(\boldsymbol{\lambda}) \left[\nabla_{m_z} \mathcal{L}(\boldsymbol{\lambda})\right] = \nabla_{\lambda_z} \mathcal{L}(\boldsymbol{\lambda}) \tag{35}
$$

If the FIM is invertible, we can express the natural-gradient with respect to $\boldsymbol{\lambda}_z$ as the gradient with respect to $\mathbf{m}_z$. The following lemma establishes the conditions under which the FIM is invertible. This condition is just a sufficient condition. There are cases where FIM is invertible but the condition does not hold.

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

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

$$
q(\mathbf{z}|\mathbf{w}) = h_z(\mathbf{z}, \mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle - A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] \tag{36}
$$

Note that

$$
\begin{aligned}
&\mathbb{E}_{q(z|w)} \left[\nabla^2_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right] && (37)\\
=&\mathbb{E}_{q(z|w)} \left[\nabla_{\lambda_z} \left[\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}) - \nabla_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right]\right] && (38)\\
=&\mathbb{E}_{q(z|w)} \left[-\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] && (39)\\
=&- \nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w}) && (40)
\end{aligned}
$$

Let's denote $\mathbf{e}_z = \frac{\nabla_{\lambda_z} q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)}{q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)}$. Conditioned on $\mathbf{w}$, we have

$$
\begin{aligned}
&\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w}) && (41)\\
=&-\mathbb{E}_{q(z|w)}\left[\nabla^2_{\lambda_z} \log q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)\right] && (42)\\
=&-\mathbb{E}_{q(z|w)}\left[\nabla_{\lambda_z}\left(\frac{\nabla_{\lambda_z} q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)}{q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)}\right)\right] && (43)\\
=&-\mathbb{E}_{q(z|w)}\left[\frac{\nabla^2_{\lambda_z} q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)}{q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)} - \mathbf{e}_z\mathbf{e}_z^T\right] && (44)\\
=&\mathbb{E}_{q(z|w)}\left[-\frac{\nabla^2_{\lambda_z} q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)}{q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)}\right] + \mathbb{E}_{q(z|w)}\left[\mathbf{e}_z\mathbf{e}_z^T\right] && (45)\\
=&\underbrace{\nabla^2_{\lambda_z}\mathbb{E}_{q(z|w)}\left[-1\right]}_{\mathbf{0}} + \mathbb{E}_{q(z|w)}\left[\mathbf{e}_z\mathbf{e}_z^T\right] && (46)\\
=&\mathbb{E}_{q(z|w)}\left[\mathbf{e}_z\mathbf{e}_z^T\right] \succeq \mathbf{0}, && (47)
\end{aligned}
$$

where we use the chain rule to move from step 43 to step 44.
Since $q(\mathbf{z}|\mathbf{w})$ is a minimal exponential-family distribution for all $\mathbf{w} \in S$, by Proposition 3.1 of Wainwright and Jordan (2008), we know that $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ is strictly convex with respect to $\boldsymbol{\lambda}_z$ given $\mathbf{w} \in S$ is known, which implies that $\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w}) \succ \mathbf{0}$.
Using Eq. 32, we have

$$
\begin{aligned}
\mathbf{F}_z(\boldsymbol{\lambda}) &= -\mathbb{E}_{q(z,w)}\left[\nabla^2_{\lambda_z} \log q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)\right] && (48)\\
&= \mathbb{E}_{q(w|\lambda_w)}\left[\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] && (49)
\end{aligned}
$$

Given any nonzero vector $\mathbf{a}$, we have

$$
\begin{aligned}
&\mathbf{a}^T\mathbf{F}_z(\boldsymbol{\lambda})\mathbf{a} && (50)\\
=&\mathbf{a}^T\left[\int_{w\in S} q(\mathbf{w}|\lambda_w)\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z,\mathbf{w})d\mathbf{w} + \int_{w\in(\Omega-S)} q(\mathbf{w}|\lambda_w)\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z,\mathbf{w})d\mathbf{w}\right]\mathbf{a} && (51)\\
=&\int_{w\in S} q(\mathbf{w}|\lambda_w)\left[\mathbf{a}^T\left(\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right)\mathbf{a}\right]d\mathbf{w} + \int_{w\in(\Omega-S)} q(\mathbf{w}|\lambda_w)\underbrace{\left[\mathbf{a}^T\left(\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right)\mathbf{a}\right]}_{\geq 0}d\mathbf{w} && (52)\\
\geq&\int_{w\in S} q(\mathbf{w}|\lambda_w)\underbrace{\left[\mathbf{a}^T\left(\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right)\mathbf{a}\right]}_{>0}d\mathbf{w} && (53)\\
>&0. && (54)
\end{aligned}
$$

We use Eq 47 to move from step 52 to step 53. Furthermore, we use the strict convexity of $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ with respect to $\boldsymbol{\lambda}_z$ when $\mathbf{w} \in S$ and the nonzero measure S to move from step 53 to step 54.
Therefore, we know that $\mathbf{F}_z(\boldsymbol{\lambda}) \succ \mathbf{0}$ and $\mathbf{F}_z(\boldsymbol{\lambda})$ is invertible. ■

Appendix B. Updates for Finite Mixture of Gaussian Distributions

Let's consider $q(w) = \text{Cate}(w|\boldsymbol{\pi})$ and $q(\mathbf{z}|w) = \text{ExpFmy}(\mathbf{z}|\boldsymbol{\lambda}_w)$. We assume $q(\mathbf{z}|w)$ is (implicitly) re-parameterizable.

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

Proof

$$q(\mathbf{z}|w) = \sum_{c=1}^{K} \mathbb{I}_c(w)\text{ExpFmy}(\mathbf{z}|\boldsymbol{\lambda}_{z_c}) \tag{55}$$

$$= \sum_{c=1}^{K} \mathbb{I}_c(w) h_z(\mathbf{z}) \exp\left[\boldsymbol{\lambda}_{z_c}^T \boldsymbol{\phi}_z(\mathbf{z}) - A_z(\boldsymbol{\lambda}_{z_c})\right] \tag{56}$$

$$= h_z(\mathbf{z}) \exp\left\{\sum_{c=1}^{K} \langle \mathbb{I}_c(w)\boldsymbol{\phi}_z(\mathbf{z}), \boldsymbol{\lambda}_{z_c}\rangle - \sum_{c'=1}^{K} \mathbb{I}_{c'}(w) A_z(\boldsymbol{\lambda}_{z_{c'}})\right\} \tag{57}$$

■

Firstly, we give the update for $q(\mathbf{z}|w)$. The natural parameter and sufficient statistics $\{\boldsymbol{\lambda}_{z_c}\}_{c=1}^K$ and $\{\mathbb{I}_c(w)\boldsymbol{\phi}_z(\mathbf{z})\}_{c=1}^K$.

The update is

$$\boldsymbol{\lambda}_{z_c} \leftarrow \boldsymbol{\lambda}_{z_c} + \beta_z \nabla_{m_{z_c}} \mathcal{L}(\boldsymbol{\lambda}), \tag{58}$$

where $m_{z_c} = \mathbb{E}_{q(w,z)}\left[\mathbb{I}_c(w)\boldsymbol{\phi}_z(\mathbf{z})\right]$.

We consider the following model.

$$p(\mathcal{D}, \mathbf{z}) = \prod_{n=1}^{N} p(\mathcal{D}_n|\mathbf{z})p(\mathbf{z}) \tag{59}$$

Let denote $h_n(\mathbf{z}) := -(\log p(\mathcal{D}_n|\mathbf{z}) + \log p(\mathbf{z})/N)$. The lower bound can be expressed as

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

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

In general, $\nabla m_{z_c} \mathcal{L}(\boldsymbol{\lambda})$ can be computed

$$\nabla_{m_{z_c}} \mathcal{L}(\boldsymbol{\lambda}) = \left(\nabla_{\lambda_{z_c}} \mathbf{m}_{z_c}\right)^{-1} \nabla_{\lambda_{z_c}} \mathcal{L}(\boldsymbol{\lambda}) \tag{62}$$

$$= \left(\nabla_{\lambda_{z_c}} \mathbb{E}_{q(w,z)}\left[\mathbb{I}_c(w)\boldsymbol{\phi}_z(\mathbf{z})\right]\right)^{-1} \nabla_{\lambda_{z_c}} \mathcal{L}(\boldsymbol{\lambda}), \tag{63}$$

where

$$\nabla_{\lambda_{z_c}} \mathbb{E}_{q(w,z)}\left[\mathbb{I}_c(w)\boldsymbol{\phi}_z(\mathbf{z})\right] = \nabla_{\lambda_{z_c}} \int \pi_c q(\mathbf{z}|w=c)\boldsymbol{\phi}_z(\mathbf{z}) d\mathbf{z} \tag{64}$$

$$= \int \pi_c q(\mathbf{z}|w=c) \nabla_z \left[\boldsymbol{\phi}_z(\mathbf{z})\right] \left[\nabla_{\lambda_{z_c}} \mathbf{z}\right] d\mathbf{z} \tag{65}$$

$\nabla_{\lambda_{z_c}} \mathcal{L}(\boldsymbol{\lambda})$ can be computed as below.

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

$$= \int \pi_c \nabla_{\lambda_{z_c}} q(\mathbf{z}|w=c) \left[-\sum_{n=1}^{N} h_n(\mathbf{z}) - \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(\mathbf{z}|\boldsymbol{\lambda}_{z_{c'}}) \right] d\mathbf{z} + \underbrace{\int q(\mathbf{z}) \left[\nabla_{\lambda_{z_c}} \log q(\mathbf{z}) \right] d\mathbf{z}}_{0} \tag{67}$$

$$= \int \pi_c q(\mathbf{z}|w=c) \left[\nabla_z \left(-\sum_{n=1}^{N} h_n(\mathbf{z}) - \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(\mathbf{z}|\boldsymbol{\lambda}_{z_{c'}}) \right) \right] \left[\nabla_{\lambda_{z_c}} \mathbf{z} \right] d\mathbf{z} \tag{68}$$

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

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

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

Its natural parameter and expectation parameter are

$$\boldsymbol{\lambda}_w = \{\log \frac{\pi_c}{\pi_K}\}_{c=1}^{K-1} \tag{70}$$

$$\mathbf{m}_w = \{\mathbb{E}_{q(w)} [\mathbb{I}_c(w)]\}_{c=1}^{K-1} = \{\pi_c\}_{c=1}^{K-1} \tag{71}$$

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

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

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

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

$$= \int q(\mathbf{z}|w=c) \left[-\sum_{n=1}^{N} h_n(\mathbf{z}) - \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(\mathbf{z}|\boldsymbol{\lambda}_{z_{c'}}) \right] d\mathbf{z} + \underbrace{\int q(\mathbf{z}) \left[\nabla_{\pi_c} \log q(\mathbf{z}) \right] d\mathbf{z}}_{0} \tag{74}$$

$$= \int q(\mathbf{z}|w=c) \left[-\sum_{n=1}^{N} h_n(\mathbf{z}) - \log \sum_{c'=1}^{K} \pi_{c'} \text{ExpFmy}(\mathbf{z}|\boldsymbol{\lambda}_{z_{c'}}) \right] d\mathbf{z} \tag{75}$$

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.

$$\nabla_{\lambda_{z_c}} \mathbb{E}_{q(w,z)} \left[\mathbb{I}(w=c) \boldsymbol{\phi}_z(\mathbf{z}) \right] = \int q(\mathbf{z}) \frac{\pi_c q(\mathbf{z}|w=c)}{q(\mathbf{z})} \left[\nabla_z \boldsymbol{\phi}_z(\mathbf{z}) \right] \left[\nabla_{\lambda_{z_c}} \mathbf{z} \right] d\mathbf{z} \tag{76}$$

$$= \mathbb{E}_{q(z)} \left[q(w=c|\mathbf{z}) \nabla_z \boldsymbol{\phi}_z(\mathbf{z}) \nabla_{\lambda_{z_c}} \mathbf{z} \right] \tag{77}$$

$$\nabla_{\lambda_{z_c}} \mathcal{L}(\boldsymbol{\lambda}) = \int q(\mathbf{z}) \frac{\pi_c q(\mathbf{z}|w=c)}{q(\mathbf{z})} \nabla_z \left[-\sum_{n=1}^{N} h_n(\mathbf{z}) - \log \sum_{c'=1}^{K} \pi_{c'} \mathrm{ExpFmy}(\mathbf{z}|\boldsymbol{\lambda}_{z_{c'}}) \right] \left[\nabla_{\lambda_{z_c}} \mathbf{z} \right] d\mathbf{z} \tag{78}$$

$$= \mathbb{E}_{q(z)} \left[q(w=c|\mathbf{z}) \nabla_z \left(-\sum_{n=1}^{N} h_n(\mathbf{z}) - \log \sum_{c'=1}^{K} \pi_{c'} \mathrm{ExpFmy}(\mathbf{z}|\boldsymbol{\lambda}_{z_{c'}}) \right) \nabla_{\lambda_{z_c}} \mathbf{z} \right], \tag{79}$$

$$\nabla_{\pi_c} \mathcal{L}(\boldsymbol{\lambda}) = \int q(\mathbf{z}) \frac{q(\mathbf{z}|w=c)}{q(\mathbf{z})} \left[-\sum_{n=1}^{N} h_n(\mathbf{z}) - \log \sum_{c'=1}^{K} \pi_{c'} \mathrm{ExpFmy}(\mathbf{z}|\boldsymbol{\lambda}_{z_{c'}}) \right] d\mathbf{z} \tag{80}$$

$$= \mathbb{E}_{q(z)} \left[\frac{q(\mathbf{z}|w=c)}{q(\mathbf{z})} \left(-\sum_{n=1}^{N} h_n(\mathbf{z}) - \log \sum_{c'=1}^{K} \pi_{c'} \mathrm{ExpFmy}(\mathbf{z}|\boldsymbol{\lambda}_{z_{c'}}) \right) \right] \tag{81}$$

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

By the chain rule, we know that

$$\nabla_{m_{z_1^c}} \mathcal{L}(\boldsymbol{\lambda}) = \frac{1}{\pi_c} \left(\nabla_{\mu_c} \mathcal{L}(\boldsymbol{\lambda}) - 2 \nabla_{\Sigma_c} \mathcal{L}(\boldsymbol{\lambda}) \boldsymbol{\mu}_c \right) \tag{82}$$

$$\nabla_{m_{z_2^c}} \mathcal{L}(\boldsymbol{\lambda}) = \frac{1}{\pi_c} \left(\nabla_{\Sigma_c} \mathcal{L}(\boldsymbol{\lambda}) \right) \tag{83}$$

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

$$\nabla_{\mu_c} \mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(z)} \left[q(w=c|\mathbf{z}) \nabla_z \left(-\sum_{n=1}^{N} h_n(\mathbf{z}) - \log \sum_{c'=1}^{K} \pi_{c'} \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{c'}, \boldsymbol{\Sigma}_{c'}) \right) \right] \tag{84}$$

$$\nabla_{\Sigma_c} \mathcal{L}(\boldsymbol{\lambda}) = \frac{1}{2} \mathbb{E}_{q(z)} \left[q(w=c|\mathbf{z}) \nabla_z^2 \left(-\sum_{n=1}^{N} h_n(\mathbf{z}) - \log \sum_{c'=1}^{K} \pi_{c'} \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{c'}, \boldsymbol{\Sigma}_{c'}) \right) \right]. \tag{85}$$

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 *$\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\mathrm{InvGam}(w|a,a)$, where $\mathbf{z} \in \mathcal{R}^d$, is a curved exponential family distribution (Lehmann and Casella, 2006).*

Proof Let $\boldsymbol{\Lambda}_1 = -\frac{1}{2}\boldsymbol{\Sigma}^{-1}, \boldsymbol{\lambda}_2 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \lambda_3 = -\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \lambda_4 = -a$. The distribution on $\mathbf{z} \in \mathcal{R}^d$ and $w \in \mathcal{R}_{++}$ can be expressed as follows.

$$\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\text{InvGam}(w|a,a) \tag{86}$$

$$=\det(2\pi w\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T(w\boldsymbol{\Sigma})^{-1}(\mathbf{z}-\boldsymbol{\mu})\}\frac{a^a}{\Gamma(a)}w^{-a-1}\exp\{-\frac{a}{w}\} \tag{87}$$

$$\begin{aligned}=(2\pi w)^{-d/2} w^{-1} \exp\{&-\tfrac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T(w\boldsymbol{\Sigma})^{-1}(\mathbf{z}-\boldsymbol{\mu}) - \tfrac{1}{2}\log\det(\boldsymbol{\Sigma}) \\ &- \frac{a}{w} - a\log w - (\log\Gamma(a) - a\log a)\}\end{aligned} \tag{88}$$

$$\begin{aligned}=(2\pi w)^{-d/2} w^{-1} \exp\{&\langle -\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}, w\mathbf{z}\mathbf{z}^T\rangle + \langle \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, w\mathbf{z}\rangle + \langle -\tfrac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, w^{-1}\rangle \\ &+ \langle -a, w^{-1} + \log w\rangle - \left[\log\Gamma(a) - a\log a + \tfrac{1}{2}\log\det(\boldsymbol{\Sigma})\right]\}\end{aligned} \tag{89}$$

$$\begin{aligned}=(2\pi w)^{-d/2} w^{-1} \exp\{&\langle \boldsymbol{\Lambda}_1, w\mathbf{z}\mathbf{z}^T\rangle + \langle \boldsymbol{\lambda}_2, w\mathbf{z}\rangle + \langle \lambda_3, w^{-1}\rangle \\ &+ \langle \lambda_4, w^{-1} + \log w\rangle - \left[\log\Gamma(-\lambda_4) + \lambda_4\log(-\lambda_4) - \tfrac{1}{2}\log\det(-2\boldsymbol{\Lambda}_1)\right]\},\end{aligned} \tag{90}$$

It is easy to see that the joint distribution is an exponential family distribution. Note that λ_3 is fully determined by $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\lambda}_2$ as shown below.

$$\begin{aligned}\lambda_3 &= -\tfrac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ &= -\tfrac{1}{2}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)^T\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\ &= -\tfrac{1}{2}\left((-2\boldsymbol{\Lambda}_1)^{-1}\boldsymbol{\lambda}_2\right)^T\boldsymbol{\lambda}_2\end{aligned} \tag{91}$$

We know that $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\text{InvGam}(w|a,a)$ is a curved exponential family distribution since λ_3 is fully determined by $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\lambda}_2$.

■

We give the update for the following model where the prior over $p(\mathbf{z}, w)$ 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 $q(\mathbf{z}, w) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\text{InvGam}(w|a,a)$ as the variational distribution. Note that when $a \leq 1$, the variance of the marginal distribution $q(\mathbf{z})$ does not exist. We assume that $a > 1$ such that the variance exists.

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

$$p(\mathcal{D}, \mathbf{z}, w) = \prod_{n=1}^{N} p(\mathcal{D}_n|\mathbf{z})\ \mathcal{N}(\mathbf{z}|\mathbf{0}, w\mathbf{I})\ \text{InvGam}(w|a_0, a_0) \tag{92}$$

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 $(\mathbf{z}_*, w_*) \sim q(\mathbf{z}, w|\boldsymbol{\lambda})$.
2. Sample a random data example $p(\mathcal{D}_n|\mathbf{z})$ and compute its gradient at $\mathbf{z}_*$.

$$\mathbf{g} \leftarrow \nabla_z \log p(\mathcal{D}_n|\mathbf{z}) \tag{93}$$

3. Compute a *learning-rate* modifier:

$$u \leftarrow \frac{a}{(a+d/2-1)}\left[1+\tfrac{1}{2}\left(\mathbf{z}_*-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{z}_*-\boldsymbol{\mu}\right)/a\right] \tag{94}$$

4. Update a for the Inverse-Gamma distribution:

$$a \leftarrow (1-\beta_2)a + \beta_2\left[a_0 - \frac{Nw_*^2}{2(1-w_*)}\mathbf{g}^\top\boldsymbol{\Sigma}\mathbf{g}\right], \tag{95}$$

where we choose β_2 such that $a > 1$.

5. Take a Newton-like step to update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

$$\mathbf{S} \leftarrow (1-\beta_1)\mathbf{S} + \beta_1(u\mathbf{g}\mathbf{g}^\top) \tag{96}$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta_1(\mathbf{S}+\mathbf{I}/N)^{-1}\left(\mathbf{g}+\boldsymbol{\mu}/N\right) \tag{97}$$

6. Repeat until convergence.

The update of $\boldsymbol{\mu}$ and $\mathbf{S}$ is very similar to the Gauss-Newton update of Khan et al. (2018), but here the learning rate β_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 $q(\mathbf{z}|w) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})$. The natural parameter and the sufficient statistics of $q(\mathbf{z}|w)$ are $\boldsymbol{\lambda}_z = \{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\}$ and $\phi_z(\mathbf{z}, w) = \{w^{-1}\mathbf{z}, w^{-1}\mathbf{z}\mathbf{z}^T\}$. The expectation parameters are

$$\mathbf{m}_{z_1} = \mathbb{E}_{q(z,w)}\left[w^{-1}\mathbf{z}\right] = \boldsymbol{\mu} \tag{98}$$

$$\mathbf{m}_{z_2} = \mathbb{E}_{q(z,w)}\left[w^{-1}\mathbf{z}\mathbf{z}^T\right] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma} \tag{99}$$

By the chain rule, we know that

$$\nabla_{m_{z_1}}\mathcal{L}(\boldsymbol{\lambda}) = \nabla_\mu\mathcal{L}(\boldsymbol{\lambda}) - 2\nabla_\Sigma\mathcal{L}(\boldsymbol{\lambda})\boldsymbol{\mu} \tag{100}$$

$$\nabla_{m_{z_2}}\mathcal{L}(\boldsymbol{\lambda}) = \nabla_\Sigma\mathcal{L}(\boldsymbol{\lambda}) \tag{101}$$

Let's denote $h_n(\mathbf{z}) := -\log p(\mathcal{D}_n|\mathbf{z})$. In the model, the lower bound $\mathcal{L}(\boldsymbol{\lambda})$ can be expressed as

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(z,w)}\left[-\sum_{n=1}^N h_n(\mathbf{z}) + \log\frac{\mathcal{N}(\mathbf{z}|\mathbf{0}, w\mathbf{I})}{\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})} + \log\frac{\mathrm{InvGam}(w|a_0, a_0)}{\mathrm{InvGam}(w|a, a)}\right], \tag{102}$$

The gradients of the lower bound are

$$\nabla_\mu\mathcal{L}(\boldsymbol{\lambda}) = \nabla_\mu\mathbb{E}_{q(z,w)}\left[-\sum_{n=1}^N h_n(\mathbf{z})\right] - \boldsymbol{\mu} \tag{103}$$

$$\nabla_\Sigma\mathcal{L}(\boldsymbol{\lambda}) = \nabla_\Sigma\mathbb{E}_{q(z,w)}\left[-\sum_{n=1}^N h_n(\mathbf{z})\right] - \tfrac{1}{2}\mathbf{I} + \tfrac{1}{2}\boldsymbol{\Sigma}^{-1} \tag{104}$$

The natural gradient update is

$$\boldsymbol{\lambda}_{z_1} \leftarrow \boldsymbol{\lambda}_{z_1} + \beta_1 \nabla_{m_{z_1}} \mathcal{L}(\boldsymbol{\lambda}) \tag{105}$$

$$\boldsymbol{\lambda}_{z_2} \leftarrow \boldsymbol{\lambda}_{z_2} + \beta_1 \nabla_{m_{z_2}} \mathcal{L}(\boldsymbol{\lambda}) \tag{106}$$

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

$$\boldsymbol{\Sigma}^{-1} \leftarrow (1-\beta_1)\boldsymbol{\Sigma}^{-1} + 2\beta_1 \nabla_{\Sigma} \mathbb{E}_{q(z,w)} \left[\sum_{n=1}^{N} h_n(\mathbf{z}) \right] + \beta_1 \mathbf{I} \tag{107}$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta_1 \boldsymbol{\Sigma} \left(\nabla_{\mu} \mathbb{E}_{q(z,w)} \left[\sum_{n=1}^{N} h_n(\mathbf{z}) \right] + \boldsymbol{\mu} \right) \tag{108}$$

By defining $\mathbf{S} = \left[\boldsymbol{\Sigma}^{-1} - \mathbf{I}\right] \Big/ N$, we have

$$\mathbf{S} \leftarrow (1-\beta_1)\mathbf{S} + 2\beta_1 \nabla_{\Sigma} \mathbb{E}_{q(z,w)} \left[\sum_{n=1}^{N} h_n(\mathbf{z})/N \right] \tag{109}$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta_1 \left(\mathbf{S} + \mathbf{I}/N\right)^{-1} \left(\nabla_{\mu} \mathbb{E}_{q(z,w)} \left[\sum_{n=1}^{N} h_n(\mathbf{z})/N \right] + \boldsymbol{\mu}/N \right) \tag{110}$$

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

$$\nabla_{m_w} \mathcal{L}(\boldsymbol{\lambda}) = \nabla_{m_w} \mathbb{E}_{q(z,w)} \left[-\sum_{n=1}^{N} h_n(\mathbf{z}) \right] + a_0 - a \tag{111}$$

The update can be expressed in terms of a as

$$a \leftarrow (1-\beta_2)a + \beta_2 \left(a_0 - \sum_{n=1}^{N} \nabla_{m_w} \mathbb{E}_{q(z,w)} \left[h_n(\mathbf{z}) \right] \right) \tag{112}$$

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 $\nabla_{m_w} \mathbb{E}_{q(z,w)} \left[h_n(\mathbf{z}) \right]$ can be computed as

$$\nabla_{m_w} \mathbb{E}_{q(z,w)} \left[h(\mathbf{z}) \right] = \left(\nabla_{\lambda_w} m_w \right)^{-1} \nabla_{\lambda_w} \mathbb{E}_{q(z,w)} \left[h_n(\mathbf{z}) \right] \tag{113}$$

$$= \left(\nabla_{\lambda_w} \mathbb{E}_{q(w)} \left[\phi_w(w) \right] \right)^{-1} \nabla_{\lambda_w} \mathbb{E}_{q(z,w)} \left[h_n(\mathbf{z}) \right] \tag{114}$$

$$= \left(\nabla_{a} \mathbb{E}_{q(w)} \left[\phi_w(w) \right] \right)^{-1} \nabla_{a} \mathbb{E}_{q(z,w)} \left[h_n(\mathbf{z}) \right] \tag{115}$$

Note that $\nabla_a \mathbb{E}_{q(w)} \left[\phi_w(w) \right] = \nabla_a m_w$ 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

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

$$\nabla_a \mathbb{E}_{q(z,w)}\left[h_n(\mathbf{z})\right] = \int\int q(w)\left(\nabla_w q(\mathbf{z}|w)\right)\left(\nabla_a w\right)h_n(\mathbf{z})dwd\mathbf{z} \tag{117}$$

$$= \int\int \text{InvGam}(w|a,a)\left(\nabla_w \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\right)\left(\nabla_a w\right)h_n(\mathbf{z})dwd\mathbf{z} \tag{118}$$

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

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

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

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

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

$$\nabla_\mu \mathbb{E}_{q(w)\mathcal{N}(z|\mu,w\Sigma)}\left[h(\mathbf{z})\right] = \int q(\mathbf{z})\nabla_z h(\mathbf{z})d\mathbf{z} \tag{121}$$

$$\nabla_\Sigma \mathbb{E}_{q(w)\mathcal{N}(z|\mu,w\Sigma)}\left[h(\mathbf{z})\right] = \tfrac{1}{2}\int u(\mathbf{z})\nabla_z^2 h(\mathbf{z})d\mathbf{z} \tag{122}$$

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

Proof

$$\nabla_{\mu_j}\mathbb{E}_{q(w)\mathcal{N}(z|\boldsymbol{\mu},w\boldsymbol{\Sigma})}\left[h(\mathbf{z})\right] = \int\int q(w)\left[\nabla_{\mu_j}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\right]h(\mathbf{z})dwd\mathbf{z} \tag{123}$$

$$= -\int q(w)\left\{\int\left[\nabla_{z_j}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\right]h(\mathbf{z})d\mathbf{z}\right\}dw \tag{124}$$

$$= -\int q(w)\left\{\int\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})h(\mathbf{z})d\mathbf{z}_{-j}\Big|_{z_j=-\infty}^{z_j=+\infty} - \int\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\nabla_{z_j}h(\mathbf{z})d\mathbf{z}\right\}dw \tag{125}$$

$$= -\underbrace{\int q(z_j,\mathbf{z}_{-j})h(z_j,\mathbf{z}_{-j})d\mathbf{z}_{-j}\Big|_{z_j=-\infty}^{z_j=+\infty}}_{0} + \int q(w)\int\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\nabla_{z_j}h(\mathbf{z})d\mathbf{z}dw \tag{126}$$

$$= \int q(\mathbf{z})\nabla_{z_j}h(\mathbf{z})d\mathbf{z} \tag{127}$$

We use integration by part 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.

$$\nabla_{\Sigma_{j,k}} \mathbb{E}_{q(w)\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})}\left[h(\mathbf{z})\right] = \int q(w)\left[\nabla_{\Sigma_{j,k}}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})\right]h(\mathbf{z})dwd\mathbf{z} \tag{128}$$

$$= \tfrac{1}{2}\int\int q(w)\left[w\nabla_{z_j}\nabla_{z_k}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})\right]h(\mathbf{z})dwd\mathbf{z} \tag{129}$$

$$= \tfrac{1}{2}\underbrace{\int q(w)w\int\left[\nabla_{z_k}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})\right]h(\mathbf{z})d\mathbf{z}_{-j}dw\Big|_{z_j=-\infty}^{z_j=+\infty}}_{0} - \tfrac{1}{2}\int\int q(w)w\nabla_{z_k}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})\nabla_{z_j}h(\mathbf{z})d\mathbf{z}dw \tag{130}$$

$$= -\tfrac{1}{2}\int q(w)w\int\nabla_{z_k}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})\nabla_{z_j}h(\mathbf{z})d\mathbf{z}dw \tag{131}$$

$$= -\tfrac{1}{2}\int q(w)w\left\{\int\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})\nabla_{z_j}h(\mathbf{z})d\mathbf{z}_{-k}\Big|_{z_k=-\infty}^{z_k=+\infty} - \int\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})\nabla_{z_k}\nabla_{z_j}h(\mathbf{z})d\mathbf{z}\right\}dw \tag{132}$$

$$= -\tfrac{1}{2}\underbrace{\int u(z_k,\mathbf{z}_{-k})\nabla_{z_j}h(\mathbf{z})d\mathbf{z}_{-k}\Big|_{z_k=-\infty}^{z_k=+\infty}}_{0} + \tfrac{1}{2}\int q(w)w\int\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})\nabla_{z_k}\nabla_{z_j}h(\mathbf{z})d\mathbf{z}dw \tag{133}$$

$$= \tfrac{1}{2}\int u(\mathbf{z})\nabla_{z_k}\nabla_{z_j}h(\mathbf{z})d\mathbf{z} \tag{134}$$

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

$$\mathbf{S} \leftarrow (1-\beta_1)\mathbf{S} + \beta_1\mathbb{E}_{q(z)}\left[\frac{a}{(a+d/2-1)}\left(1+(2a)^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\right)\left(\sum_{n=1}^{N}\nabla_z^2 h_n(\mathbf{z})/N\right)\right] \tag{135}$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta_1\left(\mathbf{S}+\mathbf{I}/N\right)^{-1}\left(\mathbb{E}_{q(z)}\left[\sum_{n=1}^{N}\nabla_z h_n(\mathbf{z})/N\right] + \boldsymbol{\mu}/N\right) \tag{136}$$

Using MC approximation, we have the following update

$$\mathbf{S} \leftarrow (1-\beta_1)\mathbf{S} + \beta_1\frac{a}{(a+d/2-1)}\left(1+(2a)^{-1}\left(\mathbf{z}_*-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}_*-\boldsymbol{\mu}\right)\right)\left(\sum_{n=1}^{N}\nabla_{z_*}^2 h_n(\mathbf{z})/N\right) \tag{137}$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta_1\left(\mathbf{S}+\mathbf{I}/N\right)^{-1}\left(\sum_{n=1}^{N}\nabla_{z_*} h_n(\mathbf{z})/N + \boldsymbol{\mu}/N\right) \tag{138}$$

Similarly, the update for $q(w)$ can be expressed as

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

where

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

$$\nabla_a \mathbb{E}_{q(z,w)}\left[h_n(\mathbf{z})\right] = \int\int \mathrm{InvGam}(w|a,a)\left(\nabla_w \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\right)(\nabla_a w)\, h_n(\mathbf{z}) dw d\mathbf{z} \tag{141}$$

Using one MC sample, we have

$$\nabla_a \mathbb{E}_{q(w)}\left[\phi_w(w)\right] \approx (w_*^{-2} - w_*^{-1})\nabla_a w_* \tag{142}$$

$$\nabla_a \mathbb{E}_{q(z,w)}\left[h_n(\mathbf{z})\right] \approx \int \left(\nabla_{w_*}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\right)(\nabla_a w_*)\, h_n(\mathbf{z}) d\mathbf{z} \tag{143}$$

$$= \int \mathrm{Tr}\left(\boldsymbol{\Sigma}\nabla_{\hat{\Sigma}}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}})\right)(\nabla_a w_*)\, h_n(\mathbf{z}) d\mathbf{z} \tag{144}$$

$$= \nabla_a w_* \mathrm{Tr}\left(\boldsymbol{\Sigma}\nabla_{\hat{\Sigma}}\mathbb{E}_{\mathcal{N}(z|\mu,\hat{\Sigma})}\left[h_n(\mathbf{z})\right]\right) \tag{145}$$

$$= \frac{\nabla_a w_*}{2} \mathrm{Tr}\left(\boldsymbol{\Sigma}\mathbb{E}_{\mathcal{N}(z|\mu,\hat{\Sigma})}\left[\nabla_z^2 h_n(\mathbf{z})\right]\right) \tag{146}$$

$$\approx \frac{\nabla_a w_*}{2} \mathrm{Tr}\left(\boldsymbol{\Sigma}\nabla_{z_*}^2 h_n(\mathbf{z})\right), \tag{147}$$

where $\hat{\boldsymbol{\Sigma}} = w_*\boldsymbol{\Sigma}$ and we use the Price's theorem $\nabla_{\hat{\Sigma}}\mathbb{E}_{\mathcal{N}(z|\mu,\hat{\Sigma})}\left[h_n(\mathbf{z})\right] = \frac{1}{2}\mathbb{E}_{\mathcal{N}(z|\mu,\hat{\Sigma})}\left[\nabla_z^2 h_n(\mathbf{z})\right]$.

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 $\boldsymbol{\Lambda}_1 = \mathbf{W}$, $\boldsymbol{\Lambda}_2 = \mathbf{U}^{-1}$, and $\boldsymbol{\Lambda}_3 = \mathbf{V}^{-1}$. The distribution on $\mathbf{Z} \in \mathcal{R}^{d\times p}$ can be expressed as follows.

$$\mathcal{MN}(\mathbf{Z}|\mathbf{W},\mathbf{U},\mathbf{V}) \tag{148}$$

$$= (2\pi)^{-dp/2}\exp\left[-\tfrac{1}{2}\mathrm{Tr}\left(\mathbf{V}^{-1}(\mathbf{Z}-\mathbf{W})^T\mathbf{U}^{-1}(\mathbf{Z}-\mathbf{W})\right) - (d/2\log\mathrm{Det}(\mathbf{V}) + p/2\log\mathrm{Det}(\mathbf{U}))\right]$$

$$= (2\pi)^{-dp/2}\exp\Big\{\mathrm{Tr}\left(\boldsymbol{\Lambda}_3\left(-\tfrac{1}{2}\mathbf{Z}+\boldsymbol{\Lambda}_1\right)^T\boldsymbol{\Lambda}_2\mathbf{Z}\right)$$

$$- \tfrac{1}{2}\Big[\mathrm{Tr}\left(\boldsymbol{\Lambda}_3\boldsymbol{\Lambda}_1^T\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_1\right) + d\log\mathrm{Det}(\boldsymbol{\Lambda}_3) + p\log\mathrm{Det}(\boldsymbol{\Lambda}_2)\Big]\Big\}. \tag{149}$$

The function $\mathrm{Tr}\left(\boldsymbol{\Lambda}_3\left(-\frac{1}{2}\mathbf{Z}+\boldsymbol{\Lambda}_1\right)^T\boldsymbol{\Lambda}_2\mathbf{Z}\right)$ is linear with respect each $\boldsymbol{\Lambda}_j$ given others. ■

We now derive the NGVI update using our new expectation parameterization. We can obtain function ϕ_1, ϕ_2, and ϕ_3 from the multi-linear function

$$f(\mathbf{Z}, \mathbf{\Lambda}) := \text{Tr}\left(\mathbf{\Lambda}_3 \left(-\tfrac{1}{2}\mathbf{Z} + \mathbf{\Lambda}_1\right)^T \mathbf{\Lambda}_2 \mathbf{Z}\right). \tag{150}$$

For example, we can obtain function ϕ_1 from $f(\mathbf{Z}, \mathbf{\Lambda})$ as shown below:

$$f(\mathbf{Z}, \mathbf{\Lambda}) = \langle \mathbf{\Lambda}_1, \underbrace{\mathbf{\Lambda}_2 \mathbf{Z} \mathbf{\Lambda}_3}_{\phi_1(\mathbf{Z}, \mathbf{\Lambda}_{-1})} \rangle \underbrace{-\tfrac{1}{2}\text{Tr}\left(\mathbf{\Lambda}_3 \mathbf{Z}^T \mathbf{\Lambda}_2 \mathbf{Z}\right)}_{r_1(\mathbf{Z}, \mathbf{\Lambda}_{-1})}. \tag{151}$$

Similarly, we can obtain functions ϕ_2 and ϕ_3. The corresponding expectation parameters of the Matrix Gaussian distribution can then be derived as below:

$$\mathbf{M}_1 = \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\mathbf{\Lambda}_2 \mathbf{Z} \mathbf{\Lambda}_3\right] = \mathbf{\Lambda}_2 \mathbf{\Lambda}_1 \mathbf{\Lambda}_3 \tag{152}$$

$$\mathbf{M}_2 = \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[-\tfrac{1}{2}\mathbf{Z}\mathbf{\Lambda}_3\mathbf{Z}^T + \mathbf{Z}\mathbf{\Lambda}_3\mathbf{\Lambda}_1^T\right] = \tfrac{1}{2}\left(\mathbf{\Lambda}_1\mathbf{\Lambda}_3\mathbf{\Lambda}_1^T - p\mathbf{\Lambda}_2^{-1}\right) \tag{153}$$

$$\mathbf{M}_3 = \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[-\tfrac{1}{2}\mathbf{Z}^T\mathbf{\Lambda}_2\mathbf{Z} + \mathbf{\Lambda}_1^T\mathbf{\Lambda}_2\mathbf{Z}\right] = \tfrac{1}{2}\left(\mathbf{\Lambda}_1^T\mathbf{\Lambda}_2\mathbf{\Lambda}_1 - d\mathbf{\Lambda}_3^{-1}\right) \tag{154}$$

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

$$\nabla_{M_1}\mathbb{E}_{q(Z|\lambda)}\left[h(\mathbf{Z})\right] = \left(\mathbf{\Lambda}_2\right)^{-1}\nabla_W \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right]\left(\mathbf{\Lambda}_3\right)^{-1} \tag{155}$$

$$\nabla_{M_2}\mathbb{E}_{q(Z|\lambda)}\left[h(\mathbf{Z})\right] = \frac{-2}{p}\nabla_U \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right] \tag{156}$$

$$\nabla_{M_3}\mathbb{E}_{q(Z|\lambda)}\left[h(\mathbf{Z})\right] = \frac{-2}{d}\nabla_V \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right] \tag{157}$$

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

$$\nabla_\mu \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\hat{h}(\mathbf{z})\right] = \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z})\right] \tag{158}$$

$$\nabla_\Sigma \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z})\right] = \tfrac{1}{2}\mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z^2 \hat{h}(\mathbf{z})\right] \tag{159}$$

These identities can be used to express the gradient with respect to the expectation parameters in terms of the gradient with respect to $\mathbf{Z}$:

$$\nabla_W \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right] = \text{Mat}\left(\mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z})\right]\right) \tag{160}$$

$$= \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})\right] \tag{161}$$

$$\nabla_U \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right] = \left(\nabla_U \mathbf{\Sigma}\right) \nabla_\Sigma \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z})\right] \tag{162}$$

$$= \tfrac{1}{2}\left(\nabla_U \mathbf{\Sigma}\right) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z^2 \hat{h}(\mathbf{z})\right] \tag{163}$$

$$\approx \tfrac{1}{2}\left(\nabla_U \mathbf{\Sigma}\right) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z}) \nabla_z \hat{h}(\mathbf{z})^T\right] \tag{164}$$

$$= \tfrac{1}{2}\mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z}) \mathbf{V} \nabla_Z h(\mathbf{Z})^T\right] \tag{165}$$

$$\nabla_V \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right] = \left(\nabla_V \mathbf{\Sigma}\right) \nabla_\Sigma \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z})\right] \tag{166}$$

$$= \tfrac{1}{2}\left(\nabla_V \mathbf{\Sigma}\right) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z^2 \hat{h}(\mathbf{z})\right] \tag{167}$$

$$\approx \tfrac{1}{2}\left(\nabla_V \mathbf{\Sigma}\right) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z}) \nabla_z \hat{h}(\mathbf{z})^T\right] \tag{168}$$

$$= \tfrac{1}{2}\mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})^T \mathbf{U} \nabla_Z h(\mathbf{Z})\right]. \tag{169}$$

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

$$\mathbf{\Lambda}_1 \leftarrow \mathbf{\Lambda}_1 - \beta_1 \left(\mathbf{\Lambda}_2\right)^{-1} \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})\right] \left(\mathbf{\Lambda}_3\right)^{-1} \tag{170}$$

$$\mathbf{\Lambda}_2 \leftarrow \mathbf{\Lambda}_2 + \beta_2 \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z}) \mathbf{V} \nabla_Z h(\mathbf{Z})^T\right] \tag{171}$$

$$\mathbf{\Lambda}_3 \leftarrow \mathbf{\Lambda}_3 + \beta_2 \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})^T \mathbf{U} \nabla_Z h(\mathbf{Z})\right] \tag{172}$$

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

$$\mathbf{W} \leftarrow \mathbf{W} - \beta_1 \mathbf{U} \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})\right] \mathbf{V} \tag{173}$$

$$\left(\mathbf{U}\right)^{-1} \leftarrow \left(\mathbf{U}\right)^{-1} + \beta_2 \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z}) \mathbf{V} \nabla_Z h(\mathbf{Z})^T\right] \tag{174}$$

$$\left(\mathbf{V}\right)^{-1} \leftarrow \left(\mathbf{V}\right)^{-1} + \beta_2 \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})^T \mathbf{U} \nabla_Z h(\mathbf{Z})\right] \tag{175}$$

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 $\Psi_j(\boldsymbol{\lambda}_j) = \mathbb{E}_{q(\cdot|\lambda_j,\lambda_{-j})}\left[A_j(\cdot)\right]$, where $A_j(\cdot)$ is the log-partition function of the distribution at

block j. As shown in the following examples, $\Psi_j(\boldsymbol{\lambda}_j)$ is convex w.r.t. $\boldsymbol{\lambda}_j$ given $\boldsymbol{\lambda}_{-j}$ is known. If $\Psi_j(\boldsymbol{\lambda}_j)$ is strictly convex w.r.t. $\boldsymbol{\lambda}_j$, we say the minimality holds. We then define the convex conjugate of function Ψ_j as below.

$$\Psi_j^*(\mathbf{x}) := \sup_{\eta} \upsilon_j(\mathbf{x}, \boldsymbol{\eta}), \tag{176}$$

where $\upsilon_j(\boldsymbol{\eta}) = \langle \mathbf{x}, \boldsymbol{\eta} \rangle - \Psi_j(\boldsymbol{\eta})$.

Example 4 (Mixture of exponential family distribution) *The variational distribution is a mixture of exponential family distribution: $q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w)$. When $j = w$, the log-partition function is $A_j(\boldsymbol{\lambda}_j) = A_w(\boldsymbol{\lambda}_w)$. The expectation parameter is $\mathbf{m}_j = \mathbb{E}_q[\boldsymbol{\phi}_w(\mathbf{w})]$. Since $q(\mathbf{w})$ is an exponential family distribution when $j = w$, we know that $A_w(\boldsymbol{\lambda}_w)$ is convex w.r.t. $\boldsymbol{\lambda}_w$. Furthermore, $\Psi_j(\boldsymbol{\lambda}_j) = A_w(\boldsymbol{\lambda}_w)$ is convex w.r.t. $\boldsymbol{\lambda}_j$ given $\boldsymbol{\lambda}_{-j}$ is known. The minimal condition implies that $q(\mathbf{w})$ is a minimal exponential family distribution. When $j = z$, the log-partition function is $A_j(\boldsymbol{\lambda}_j) = A_z(\boldsymbol{\lambda}_z, \mathbf{w})$. The expectation parameter is $\mathbf{m}_j = \mathbb{E}_q[\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})]$. According Eq. 47, we know that $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ is convex w.r.t. $\boldsymbol{\lambda}_j$ for any valid $\mathbf{w}$, which implies that $\Psi_j(\boldsymbol{\lambda}_j) = \mathbb{E}_{q(w)}[A_z(\boldsymbol{\lambda}_z, \mathbf{w})]$ is convex w.r.t. $\boldsymbol{\lambda}_j$ given $\boldsymbol{\lambda}_{-j}$ 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: $q(\mathbf{z}|\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N)$. When $j \in \{1, \ldots, N\}$, the log-partition function is $A_j(\boldsymbol{\lambda}_j) = A_z(\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j})$. The expectation parameter is $\mathbf{m}_j = \mathbb{E}_q[\phi_j(\mathbf{z}, \boldsymbol{\lambda}_{-j})]$. Since $\boldsymbol{\lambda}_{-j}$ is known, we know that $q(\mathbf{z}|\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}) = q_j(\mathbf{z}|\boldsymbol{\lambda}_j)$ becomes a one-parameter exponential family distribution. We know that $\Psi_j(\boldsymbol{\lambda}_j) = A_z(\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j})$ is convex w.r.t. $\boldsymbol{\lambda}_j$ given that $\boldsymbol{\lambda}_{-j}$ is known. The minimal condition implies that for all $j \in \{1, \ldots, N\}$, $q_j(\mathbf{z}|\boldsymbol{\lambda}_j)$ is a minimal exponential family distribution.*

Lemma 11 *If the minimality holds, we have*

$$\Psi_j^*(\mathbf{m}_j) = \sup_{\eta} \upsilon_j(\mathbf{m}_j, \boldsymbol{\eta}) \tag{177}$$

$$= \langle \mathbf{m}_j, \boldsymbol{\lambda}_j \rangle - \Psi_j(\boldsymbol{\lambda}_j), \tag{178}$$

where $\boldsymbol{\lambda}_j$ and $\mathbf{m}_j$ 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 $j = \mathbf{w}$, since $q(\mathbf{w})$ is an exponential family distribution, we know that

$$\mathbf{m}_w = \nabla_{\lambda_w} A_w(\boldsymbol{\lambda}_w), \tag{179}$$

which implies that $\mathbf{m}_w - \nabla_{\lambda_w} A_w(\boldsymbol{\lambda}_w) = \mathbf{0}$.

Note that the gradient of $\upsilon_j(\mathbf{m}_w, \boldsymbol{\eta})$ is

$$\nabla_\eta \upsilon_j(\mathbf{m}_w, \boldsymbol{\eta}) = \nabla_\eta \langle \mathbf{m}_w, \boldsymbol{\eta} \rangle - \Psi_j(\boldsymbol{\eta}) = \mathbf{m}_w - \nabla_\eta A_w(\boldsymbol{\eta}). \tag{180}$$

When $\boldsymbol{\eta} = \boldsymbol{\lambda}_w$, the gradient is $\mathbf{0}$. Since $\Psi_j(\boldsymbol{\lambda}_w)$ is convex w.r.t. $\boldsymbol{\lambda}_w$, we know that $\boldsymbol{\lambda}_w$ is an optimal solution. Since the minimality holds, we know that $\Psi_j(\boldsymbol{\lambda}_w)$ is strictly convex, which implies that the solution is unique. Therefore, $\underbrace{\Psi_w^*(\mathbf{m}_w)}_{\Psi_j^*(\mathbf{m}_j)} = \underbrace{\langle \mathbf{m}_w, \boldsymbol{\lambda}_w \rangle - A_w(\boldsymbol{\lambda}_w)}_{\langle \mathbf{m}_j, \boldsymbol{\lambda}_j \rangle - \Psi_j(\boldsymbol{\lambda}_j)}$.

When $j = \mathbf{z}$, by Eq. 21, we know that

$$\mathbf{m}_z = \nabla_{\lambda_z} \mathbb{E}_{q(w)} \left[A_z(\boldsymbol{\lambda}_z, \mathbf{w}) \right], \tag{181}$$

which implies that $\mathbf{m}_z - \nabla_{\lambda_z} \mathbb{E}_{q(w)} \left[A_z(\boldsymbol{\lambda}_z, \mathbf{w}) \right] = \mathbf{0}$. Similarly, since the minimality holds, when $\boldsymbol{\lambda}_w$ is known, we know $\underbrace{\Psi_z^*(\mathbf{m}_z)}_{\Psi_j^*(\mathbf{m}_j)} = \underbrace{\langle \mathbf{m}_z, \boldsymbol{\lambda}_z \rangle - \mathbb{E}_q \left[A_z(\boldsymbol{\lambda}_z, \mathbf{w}) \right]}_{\langle \mathbf{m}_j, \boldsymbol{\lambda}_j \rangle - \Psi_j(\boldsymbol{\lambda}_j)}$. ■

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

Lemma 12 *If the minimality holds, given that $\boldsymbol{\lambda}_{-j} = \boldsymbol{\lambda}_{-j}^k$, the following identity is true.*

$$\mathbb{B}_{\Psi_j^*}(\mathbf{m}_j \| \mathbf{m}_j^k) = \mathbb{D}_{KL}(q(\cdot|\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}^k) \,\|\, q(\cdot|\boldsymbol{\lambda}_j^k, \boldsymbol{\lambda}_{-j}^k)) \tag{182}$$

$$= \Psi_j(\boldsymbol{\lambda}_j^k) - \Psi_j(\boldsymbol{\lambda}_j) - \left\langle \nabla_{\lambda_j} \Psi_j(\boldsymbol{\lambda}_j), \left(\boldsymbol{\lambda}_j^k - \boldsymbol{\lambda}_j \right) \right\rangle, \tag{183}$$

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

$$\mathbb{B}_{\Psi_j^*}(\mathbf{m}_j \| \mathbf{m}_j^k) = \mathbb{B}_{\Psi_j}(\boldsymbol{\lambda}_j^k \| \boldsymbol{\lambda}_j) \tag{184}$$

When $j = w$, $\Psi_w(\boldsymbol{\lambda}_w) = A_w(\boldsymbol{\lambda}_w)$. Therefore, we have the following results.

$$\mathbb{D}_{KL}(q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda}_w, \boldsymbol{\lambda}_z^k) \,\|\, q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda}_w^k, \boldsymbol{\lambda}_z^k)) \tag{185}$$

$$= \mathbb{D}_{KL}(q(\mathbf{w}|\boldsymbol{\lambda}_w) \,\|\, q(\mathbf{w}|\boldsymbol{\lambda}_w^k)) \tag{186}$$

$$= \left\langle \mathbb{E}_{q(w|\lambda_w)} \left[\boldsymbol{\phi}_w(\mathbf{w}) \right], \left(\boldsymbol{\lambda}_w - \boldsymbol{\lambda}_w^k \right) \right\rangle - A_w(\boldsymbol{\lambda}_w) + A_w(\boldsymbol{\lambda}_w^k) \tag{187}$$

$$= \left\langle \mathbf{m}_w, \left(\boldsymbol{\lambda}_w - \boldsymbol{\lambda}_w^k \right) \right\rangle - A_w(\boldsymbol{\lambda}_w) + A_w(\boldsymbol{\lambda}_w^k) \tag{188}$$

$$= \Psi_w(\boldsymbol{\lambda}_w^k) - \Psi_w(\boldsymbol{\lambda}_w) - \left\langle \nabla_{\lambda_w} \Psi_w(\boldsymbol{\lambda}_w), \left(\boldsymbol{\lambda}_w^k - \boldsymbol{\lambda}_w \right) \right\rangle \tag{189}$$

$$= \mathbb{B}_{\Psi_w}(\boldsymbol{\lambda}_w^k \| \boldsymbol{\lambda}_w), \tag{190}$$

where $\mathbf{m}_w = \nabla_{\lambda_w} A_w(\boldsymbol{\lambda}_w) = \nabla_{\lambda_w} \Psi_w(\boldsymbol{\lambda}_w)$.

When $j = z$, $\Psi_z(\boldsymbol{\lambda}_z) = \mathbb{E}_{q(w|\lambda_w^k)}\left[A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right]$. Therefore, we have the following results.

$$\begin{aligned}
&\mathbb{D}_{KL}(q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w^k) \,\|\, q(\mathbf{z}, \mathbf{w}|\boldsymbol{\lambda}_z^k, \boldsymbol{\lambda}_w^k)) && (191)\\
=&\mathbb{D}_{KL}(q(\mathbf{w}|\boldsymbol{\lambda}_w^k)q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) \,\|\, q(\mathbf{w}|\boldsymbol{\lambda}_w^k)q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z^k)) && (192)\\
=&\left\langle \mathbb{E}_{q(z,w|\lambda_z,\lambda_w^k)}\left[\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})\right], \left(\boldsymbol{\lambda}_z - \boldsymbol{\lambda}_z^k\right)\right\rangle - \mathbb{E}_{q(w|\lambda_w^k)}\left[A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] + \mathbb{E}_{q(w|\lambda_w^k)}\left[A_z(\boldsymbol{\lambda}_z^k, \mathbf{w})\right] && (193)\\
=&\left\langle \mathbf{m}_z, \left(\boldsymbol{\lambda}_z - \boldsymbol{\lambda}_z^k\right)\right\rangle - \Psi_z(\boldsymbol{\lambda}_z) + \Psi_z(\boldsymbol{\lambda}_z^k) && (194)\\
=&\Psi_z(\boldsymbol{\lambda}_z^k) - \Psi_z(\boldsymbol{\lambda}_z) - \left\langle \nabla_{\lambda_z}\Psi_z(\boldsymbol{\lambda}_z), \left(\boldsymbol{\lambda}_z^k - \boldsymbol{\lambda}_z\right)\right\rangle && (195)\\
=&\mathbb{B}_{\Psi_z}(\boldsymbol{\lambda}_z^k \| \boldsymbol{\lambda}_z), && (196)
\end{aligned}$$

where $\mathbf{m}_z = \nabla_{\lambda_z}\Psi_z(\boldsymbol{\lambda}_z)$ due to Eq. 181. ■

Lemma 13 *If the minimality holds, given $\boldsymbol{\lambda}_{-j} = \boldsymbol{\lambda}_{-j}^k$, a valid step-size β_j^k, and the following optimization problem*

$$\max_{m_j} \left\langle \mathbf{m}_j, \widehat{\nabla}_{m_j^k}\mathcal{L}(\boldsymbol{\lambda})\right\rangle - \frac{1}{\beta_j^k}\mathbb{B}_{\Psi_j^*}(\mathbf{m}_j \| \mathbf{m}_j^k), \tag{197}$$

the solution of the optimization problem is

$$\boldsymbol{\lambda}_j^{k+1} = \boldsymbol{\lambda}_j^k + \beta_j^k \widehat{\nabla}_{m_j^k}\mathcal{L}(\boldsymbol{\lambda}) \tag{198}$$

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

$$\begin{aligned}
&\left\langle \mathbf{m}_j, \widehat{\nabla}_{m_j^k}\mathcal{L}(\boldsymbol{\lambda})\right\rangle - \frac{1}{\beta_j^k}\mathbb{B}_{\Psi_j^*}(\mathbf{m}_j \| \mathbf{m}_j^k) && (199)\\
=&\left\langle \mathbf{m}_j, \widehat{\nabla}_{m_j^k}\mathcal{L}(\boldsymbol{\lambda})\right\rangle - \frac{1}{\beta_j^k}\mathbb{D}_{KL}(q(\cdot|\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}^k) \,\|\, q(\cdot|\boldsymbol{\lambda}_j^k, \boldsymbol{\lambda}_{-j}^k)) && (200)\\
=&\left\langle \mathbf{m}_j, \widehat{\nabla}_{m_j^k}\mathcal{L}(\boldsymbol{\lambda})\right\rangle - \frac{1}{\beta_j^k}\left\{\left\langle \mathbf{m}_j, \left(\boldsymbol{\lambda}_j - \boldsymbol{\lambda}_j^k\right)\right\rangle - \Psi_j(\boldsymbol{\lambda}_j) + \Psi_j(\boldsymbol{\lambda}_j^k)\right\} && (201)
\end{aligned}$$

Similarly, by taking the gradient of Eq. 201 w.r.t. $\mathbf{m}_j$ to be $\mathbf{0}$, we have

$$\begin{aligned}
\mathbf{0} =&\nabla_{m_j}\left\{\left\langle \mathbf{m}_j, \widehat{\nabla}_{m_j^k}\mathcal{L}(\boldsymbol{\lambda})\right\rangle - \frac{1}{\beta_j^k}\left\{\left\langle \mathbf{m}_j, \left(\boldsymbol{\lambda}_j - \boldsymbol{\lambda}_j^k\right)\right\rangle - \Psi_j(\boldsymbol{\lambda}_j) + \Psi_j(\boldsymbol{\lambda}_j^k)\right\}\right\}\\
=&\widehat{\nabla}_{m_j^k}\mathcal{L}(\boldsymbol{\lambda}) - \frac{1}{\beta_j^k}\left[\left(\boldsymbol{\lambda}_j - \boldsymbol{\lambda}_j^k\right) + \left(\nabla_{m_j}\boldsymbol{\lambda}_j\right)\mathbf{m}_j - \nabla_{m_j}\Psi_j(\boldsymbol{\lambda}_j)\right] && (202)\\
=&\widehat{\nabla}_{m_j^k}\mathcal{L}(\boldsymbol{\lambda}) - \frac{1}{\beta_j^k}\left[\left(\boldsymbol{\lambda}_j - \boldsymbol{\lambda}_j^k\right) + \left(\nabla_{m_j}\boldsymbol{\lambda}_j\right)\mathbf{m}_j - \left(\nabla_{m_j}\boldsymbol{\lambda}_j\right)\underbrace{\nabla_{\lambda_j}\Psi_j(\boldsymbol{\lambda}_j)}_{\mathbf{m}_j}\right] && (203)\\
=&\widehat{\nabla}_{m_j^k}\mathcal{L}(\boldsymbol{\lambda}) - \frac{1}{\beta_j^k}\left(\boldsymbol{\lambda}_j - \boldsymbol{\lambda}_j^k\right) && (204)
\end{aligned}$$

Finally, we obtain the following solution by solving Eq. 204

$$\boldsymbol{\lambda}_j^{k+1} = \boldsymbol{\lambda}_j^k + \beta_j^k \widehat{\nabla}_{m_j} \mathcal{L}(\boldsymbol{\lambda}) \tag{205}$$

■

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

$$\boldsymbol{\lambda}_w^{k+1} = \boldsymbol{\lambda}_w^k + \beta \widehat{\nabla}_{m_w^k} \mathcal{L}(\boldsymbol{\lambda}) \tag{206}$$

$$\boldsymbol{\lambda}_z^{k+1} = \boldsymbol{\lambda}_z^k + \beta \widehat{\nabla}_{m_z^k} \mathcal{L}(\boldsymbol{\lambda}), \tag{207}$$

where $\beta_w^k = \beta_z^k = \beta$. 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

$$\boldsymbol{\lambda}_j^{k+1} = \boldsymbol{\lambda}_j^k + \beta \widehat{\nabla}_{m_j^k} \mathcal{L}(\boldsymbol{\lambda}) \;\; j \in \{1, \ldots, N\}. \tag{208}$$

Since $\boldsymbol{\lambda}_{-j}^k$ is known, by the chain rule, we have

$$\widehat{\nabla}_{m_j^k} \widetilde{\mathcal{L}}(\boldsymbol{\lambda}) = \left(\widehat{\nabla}_{\lambda_j^k} \mathbf{m}_j\right)^{-1} \widehat{\nabla}_{\lambda_j^k} \mathcal{L}(\boldsymbol{\lambda}) \tag{209}$$

$$= \left(\widehat{\nabla}_{\lambda_j^k} \mathbb{E}_{q(z|\lambda_j, \lambda_{-j}^k)} \left[\phi_j(\mathbf{z}, \boldsymbol{\lambda}_{-j}^k)\right]\right)^{-1} \widehat{\nabla}_{\lambda_j^k} \mathcal{L}(\boldsymbol{\lambda}) \tag{210}$$

Recall that $q(\mathbf{z}|\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}^k)$ is an one-parameter exponential family distribution given that $\boldsymbol{\lambda}_{-j}$ is known. We have

$$\mathbf{0} = \mathbb{E}_{q(z|\lambda_j, \lambda_{-j}^k)} \left[\nabla_{\lambda_j} \log q(\mathbf{z}|\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}^k)\right] \tag{211}$$

$$= \mathbb{E}_{q(z|\lambda_j, \lambda_{-j}^k)} \left[\nabla_{\lambda_j} \left[\langle \phi_j(\mathbf{z}, \boldsymbol{\lambda}_{-j}^k), \boldsymbol{\lambda}_j \rangle + r_j(\boldsymbol{\lambda}_{-j}^k) - A_z(\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}^k)\right]\right] \tag{212}$$

$$= \mathbb{E}_{q(z|\lambda_j, \lambda_{-j}^k)} \left[\phi_j(\mathbf{z}, \boldsymbol{\lambda}_{-j}^k) - \nabla_{\lambda_j} A_z(\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}^k)\right] \tag{213}$$

$$= \mathbb{E}_{q(z|\lambda_j, \lambda_{-j}^k)} \left[\phi_j(\mathbf{z}, \boldsymbol{\lambda}_{-j}^k)\right] - \nabla_{\lambda_j} A_z(\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}^k) \tag{214}$$

Let's define

$$\mathbf{F}_{\lambda_j} = \mathbb{E}_{q(z|\lambda_j, \lambda_{-j}^k)} \left[\nabla_{\lambda_j}^2 A_z(\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}^k)\right] = \nabla_{\lambda_j}^2 A_z(\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}^k). \tag{215}$$

Due to Eq 214, we know that

$$\widehat{\nabla}_{\lambda_j^k} \mathbb{E}_{q(z|\lambda_j, \lambda_{-j}^k)} \left[\phi_j(\mathbf{z}, \boldsymbol{\lambda}_{-j}^k)\right] = \mathbf{F}_{\lambda_j^k}. \tag{216}$$

Therefore, the Jacobi update is

$$\boldsymbol{\lambda}_j^{k+1} = \boldsymbol{\lambda}_j^k + \beta \left(\mathbf{F}_{\lambda_j^k}\right)^{-1} \widehat{\nabla}_{\lambda_j^k} \mathcal{L}(\boldsymbol{\lambda}) \;\; j \in \{1, \ldots, N\}. \tag{217}$$

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

$$\mathbf{F}_{\lambda^k} = -\mathbb{E}_{q(z|\lambda^k)} \left[\widehat{\nabla}_{\lambda^k}^2 \log q(\mathbf{z}|\boldsymbol{\lambda})\right] \tag{218}$$

$$= -\mathbb{E}_{q(z|\lambda^k)} \begin{bmatrix} \widehat{\nabla}_{\lambda_1^k}^2 \log q(\mathbf{z}|\boldsymbol{\lambda}) & \cdots & \widehat{\nabla}_{\lambda_1^k} \widehat{\nabla}_{\lambda_N^k} \log q(\mathbf{z}|\boldsymbol{\lambda}) \\ \cdots & \cdots & \cdots \\ \widehat{\nabla}_{\lambda_N^k} \widehat{\nabla}_{\lambda_1^k} \log q(\mathbf{z}|\boldsymbol{\lambda}) & \cdots & \widehat{\nabla}_{\lambda_N^k}^2 \log q(\mathbf{z}|\boldsymbol{\lambda}) \end{bmatrix} \tag{219}$$

$$= -\mathbb{E}_{q(z|\lambda^k)} \begin{bmatrix} -\widehat{\nabla}_{\lambda_1^k}^2 A_z(\boldsymbol{\lambda}) & \cdots & \widehat{\nabla}_{\lambda_1^k} \widehat{\nabla}_{\lambda_N^k} \log q(\mathbf{z}|\boldsymbol{\lambda}) \\ \cdots & \cdots & \cdots \\ \widehat{\nabla}_{\lambda_N^k} \widehat{\nabla}_{\lambda_1^k} \log q(\mathbf{z}|\boldsymbol{\lambda}) & \cdots & -\widehat{\nabla}_{\lambda_N^k}^2 A_z(\boldsymbol{\lambda}) \end{bmatrix} \tag{220}$$

$$= \begin{bmatrix} \mathbf{F}_{\lambda_1^k} & \cdots & -\widehat{\nabla}_{\lambda_1^k} \widehat{\nabla}_{\lambda_{z_N}^k} \log q(\mathbf{z}|\boldsymbol{\lambda}) \\ \cdots & \cdots & \cdots \\ -\widehat{\nabla}_{\lambda_N^k} \widehat{\nabla}_{\lambda_1^k} \log q(\mathbf{z}|\boldsymbol{\lambda}) & \cdots & \mathbf{F}_{\lambda_N^k} \end{bmatrix} \tag{221}$$

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

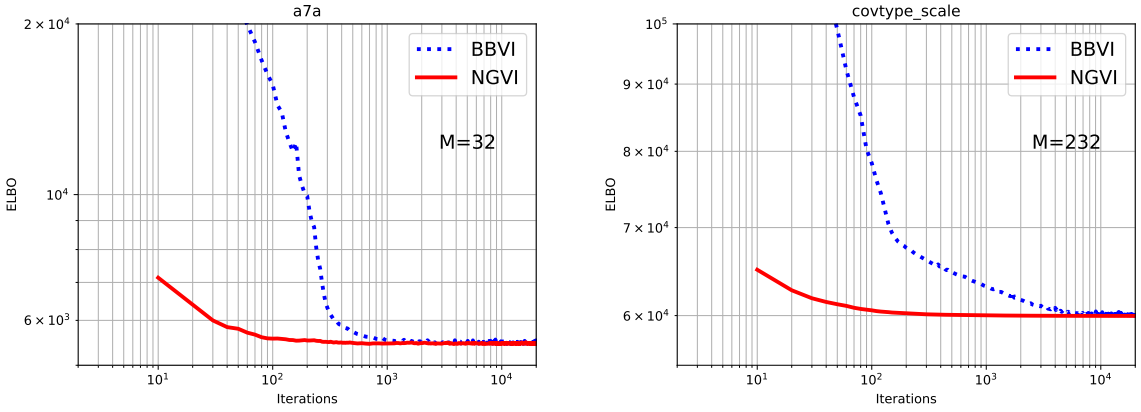

Figure 2: This figure demonstrates a fast convergence of NGVI over BBVI to approximate the posterior distribution of Bayesian logistic regression using a scale mixture of Gaussians with full covariance matrix. The plot shows the ELBO of the whole training set obtained using 5×10^4 MC samples. For both algorithms, we used mini-batches by using 10 MC samples to compute stochastic approximations. M denotes that 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.

matrix as shown below.

$$\widehat{\mathbf{F}}_{\lambda^k} := \begin{bmatrix} \mathbf{F}_{\lambda_1^k} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots \\ \mathbf{0} & \cdots & \mathbf{F}_{\lambda_N^k} \end{bmatrix} \tag{222}$$

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

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.