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

Wu Lin (UBC), Mohammad Emtiyaz Khan (AIP, RIKEN), Mark Schmidt (UBC)

1 Introduction

Natural gradient method for variational inference can lead to fast convergent algorithms, but its application is usually restricted to exponential-family approximations. We extend the application to a class of distributions, and show fast faster convergence than existing block-box gradient methods.

2 VI using Exponential Family

Given data $\mathcal{D}$ and model $p(\mathcal{D}|\mathbf{z})$ with latent vector $\mathbf{z}$ and prior $p(\mathbf{z})$, our goal is to approximate posterior $p(\mathbf{z}|\mathcal{D})$. Variational inference (VI) approximates the posterior by optimizing the evidence lower bound (ELBO) $\mathcal{L}$ induced by a variational distribution $q(\mathbf{z}|\boldsymbol{\lambda}_z)$.

Black-Box VI and Natural-Gradient VI:

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

Advantages of NGVI:

- NGVI admits a simple update in the exponential family (Khan and Lin, 2017).

  NGVI for Exp-Family: $\boldsymbol{\lambda}_z \leftarrow \boldsymbol{\lambda}_z + \beta\ \nabla_{m_z}\mathcal{L}(\boldsymbol{\lambda}_z)$, where $\mathbf{m}_z$ is the expectation parameter.
- NGVI often results in faster convergence than BBVI.

Challenges of NGVI:

- NGVI could be complicated due to computing $[\mathbf{F}_z(\boldsymbol{\lambda}_z)]^{-1}\nabla_{\lambda_z}\mathcal{L}(\boldsymbol{\lambda}_z)$.
- $\mathbf{F}_z(\boldsymbol{\lambda}_z)$ can be singular.
- Usually, NGVI does not admit a simple update outside the class of exp-family.

3 Simple Natural-gradient VI Update

Many existing works assume q to be exponential family (e.g., Gaussian). In this work, we consider more flexible approximations that are a type of mixture approximations. We consider a new NGVI update which admits a simple update in the following cases.

Structured Approximation: We consider $q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda}) = q(\mathbf{w}|\boldsymbol{\lambda}_w)q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$, where

$$\text{Conditional Exp-family}: q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) := h_z(\mathbf{z}, \mathbf{w}) \exp\left[\langle \phi_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z\rangle - A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right],$$
$$q(\mathbf{w}|\boldsymbol{\lambda}_w) := h_w(\mathbf{w}) \exp\left[\langle \phi_w(\mathbf{w}), \boldsymbol{\lambda}_w\rangle - A_w(\boldsymbol{\lambda}_w)\right].$$

We assume the set of parameters for $q(\mathbf{w})$ and $q(\mathbf{z}|\mathbf{w})$ denoted by Ω_w and Ω_z are open respectively.

For the mixture of exponential family distribution $q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda})$, we define the following

- Expectation parameter:

  $\mathbf{m}_w := \mathbb{E}_{q(w)}[\phi_w(\mathbf{w})] \in \mathcal{M}_w, \quad \mathbf{m}_z := \mathbb{E}_{q(w,z)}[\phi_z(\mathbf{z}, \mathbf{w})] \in \mathcal{M}_z$
- Natural parameter: $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_w, \boldsymbol{\lambda}_z) \in \Omega_w \times \Omega_z$
- Fisher information matrix: $\mathbf{F}_{wz}(\boldsymbol{\lambda}_w, \boldsymbol{\lambda}_z) = -\mathbb{E}_{q(w,z)}\left[\nabla^2 \log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda}_w, \boldsymbol{\lambda}_z)\right]$

The following natural gradient update in natural parameters:

$$\begin{bmatrix}\boldsymbol{\lambda}_w^{t+1}\\ \boldsymbol{\lambda}_z^{t+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\lambda}_w^{t}\\ \boldsymbol{\lambda}_z^{t}\end{bmatrix} + \beta \underbrace{\mathbf{F}_{wz}(\boldsymbol{\lambda}_w^t, \boldsymbol{\lambda}_z^t)^{-1}\begin{bmatrix}\nabla_{\lambda_w}\mathcal{L}^t\\ \nabla_{\lambda_z}\mathcal{L}^t\end{bmatrix}}_{\text{Natural gradient}}$$

When $\mathbf{F}_{wz}(\boldsymbol{\lambda}_w^t, \boldsymbol{\lambda}_z^t)$ is invertible, the update is equivalent to

$$\text{NGVI}: \quad \begin{aligned}\boldsymbol{\lambda}_w^{t+1} &= \boldsymbol{\lambda}_w^{t} + \beta\nabla_{m_w}\mathcal{L}^t\\ \boldsymbol{\lambda}_z^{t+1} &= \boldsymbol{\lambda}_z^{t} + \beta\nabla_{m_z}\mathcal{L}^t\end{aligned}$$

Now, we give a sufficient condition when $\mathbf{F}_{wz}(\boldsymbol{\lambda}_w^t, \boldsymbol{\lambda}_z^t)$ is invertible.

Definition 1: Minimal Conditional Exp-family (MCEF)

A conditional exp-family is said to have a minimal representation when $\mathbf{m}_w(\cdot): \Omega_w \to \mathcal{M}_w$ and $\mathbf{m}_z(\cdot, \boldsymbol{\lambda}_w): \Omega_z \to \mathcal{M}_z$ are both one-to-one, $\forall \boldsymbol{\lambda}_w \in \Omega_w$.

Theorem 1

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

We further generalize the following multi-linear exponential family with N blocks.

$$\text{Multi-linear Exp-family}: q(\mathbf{z}|\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_N) = h_z(\mathbf{z}) \exp\left[f(\mathbf{z}, \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_N) - A_z(\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_N)\right],$$

where we assume $f(\mathbf{z}, \boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_N)$ is a linear function w.r.t. each block $\boldsymbol{\lambda}_j$ given others, and the set of parameters for $q(\mathbf{z})$ denoted by Ω_j is open for each block.

Similarly, for the multi-linear exponential family distribution, we propose to optimize $\boldsymbol{\lambda}_j$ given $\boldsymbol{\lambda}_{-j}^t$. The distribution then can be re-expressed as

$$q(\mathbf{z}|\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}) = h_z(\mathbf{z}) \exp\left[\underbrace{\langle \phi_j(\mathbf{z}, \boldsymbol{\lambda}_{-j}), \boldsymbol{\lambda}_j\rangle + r_j(\mathbf{z}, \boldsymbol{\lambda}_{-j})}_{f(\mathbf{z}, \boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j})} - A_z(\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j})\right]$$

For the j-th block, we define the following

- Expectation parameter: $\mathbf{m}_j := \mathbb{E}_{q(z)}\left[\phi_j(\mathbf{z}, \boldsymbol{\lambda}_{-j})\right] \in \mathcal{M}_j$
- Natural parameter: $\boldsymbol{\lambda}_j \in \Omega_j$
- Fisher information matrix: $\mathbf{F}_j(\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j}) = -\mathbb{E}_{q(z)}\left[\nabla^2_{\lambda_j} \log q(\mathbf{z}|\boldsymbol{\lambda}_j, \boldsymbol{\lambda}_{-j})\right]$

The following block natural gradient update in natural parameters at block j:

$$\boldsymbol{\lambda}_j^{t+1} = \boldsymbol{\lambda}_j^{t} + \beta \underbrace{\mathbf{F}_j(\boldsymbol{\lambda}_j^t, \boldsymbol{\lambda}_{-j}^t)^{-1}\nabla_{\lambda_j}\mathcal{L}^t}_{\text{Natural gradient}}$$

When $\mathbf{F}_j(\boldsymbol{\lambda}_j^t, \boldsymbol{\lambda}_{-j}^t)$ is invertible, the update is equivalent to

$$\text{block NGVI}: \boldsymbol{\lambda}_j^{t+1} = \boldsymbol{\lambda}_j^{t} + \beta\nabla_{m_j}\mathcal{L}^t$$

Now, we give a sufficient condition when $\mathbf{F}_j(\boldsymbol{\lambda}_j^t, \boldsymbol{\lambda}_{-j}^t)$ is invertible.

Definition 2: Minimal Multi-linear Exp-family (MMEF)

A conditional exp-family is said to have a minimal representation when $\mathbf{m}_j(\cdot, \boldsymbol{\lambda}_{-j}): \Omega_j \to \mathcal{M}_j$ is one-to-one, $\forall \boldsymbol{\lambda}_{-j} \in \Omega_{-j}$.

Theorem 2

For an MMEF representation, the FIM $\mathbf{F}_j(\boldsymbol{\lambda})$ is positive-definite and invertible for all $\boldsymbol{\lambda} \in \Omega$.

References:

- Khan and Lin. Conjugate-computation variational inference. *AIStats*, 2017.
- Gupta et al. Shampoo: Preconditioned Stochastic Tensor Optimization *ICML*, 2018.
- Zhang et al. Noisy natural gradient as variational inference. *ICML*, 2018.

4 Examples

Example of Mixture of Exponential Family

Consider a model with a Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})$.

$$p(\mathcal{D}, \mathbf{z}) = \prod_{n=1}^{N} p(\mathcal{D}_n|\mathbf{z})\mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})$$

We use the following mixture of exponential family distributions (skew-Gaussian distribution).

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

The natural parameter and expectation parameter are shown below, where $c = \sqrt{2/\pi}$.

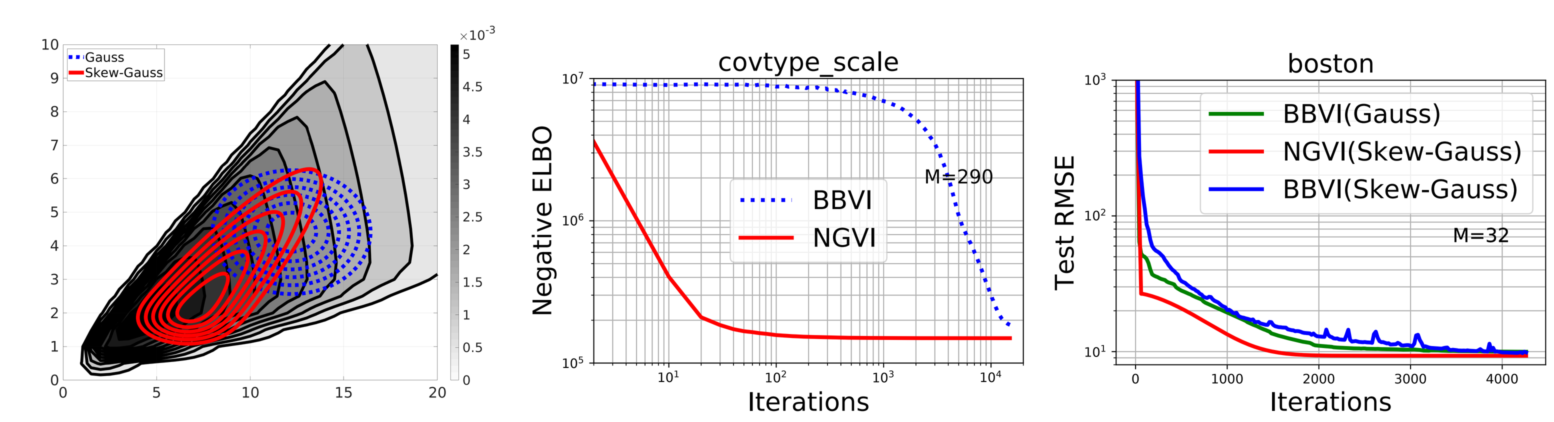
$$\boldsymbol{\lambda}_{z_1} = \boldsymbol{\Sigma}^{-1}\mu, \qquad \boldsymbol{\lambda}_{z_2} = \boldsymbol{\Sigma}^{-1}\alpha, \qquad \boldsymbol{\lambda}_{z_3} = -\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}$$
$$\mathbf{m}_{z_1} = \mu + c\alpha, \qquad \mathbf{m}_{z_2} = c\mu + \alpha, \qquad \mathbf{m}_{z_3} = \mu\mu^T + \alpha\alpha^T + c\left(\mu\alpha^T + \alpha\mu^T\right) + \boldsymbol{\Sigma}$$

The ELBO: $\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(z)}\left[\sum_{n=1}^N \log p(\mathcal{D}_n|\mathbf{z}) + \overbrace{\log\mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})}^{\text{prior}}\overbrace{-\log q(\mathbf{z})}^{\text{entropy}}\right]$

We can re-express the update in terms of μ, $\boldsymbol{\Sigma}^{-1}$, and α, where $\mathbf{g}^n_\mu$, $\mathbf{g}^n_\alpha$, and $\mathbf{g}^n_\Sigma$ are defined in the paper. The derivatives about the prior and the entropy can be computed almost exactly.

$$\text{NGVI}: \quad \begin{aligned}\boldsymbol{\Sigma}^{-1} &\leftarrow (1-\beta)\boldsymbol{\Sigma}^{-1} + \beta(\delta\mathbf{I} + N\mathbf{g}^n_\Sigma)\\ \mu &\leftarrow \mu - \beta\boldsymbol{\Sigma}\left(\frac{N}{1-c^2}(\mathbf{g}^n_\mu - c\mathbf{g}^n_\alpha) + \delta\mu\right)\\ \alpha &\leftarrow \alpha - \beta\boldsymbol{\Sigma}\left(\frac{N}{1-c^2}(\mathbf{g}^n_\alpha - c\mathbf{g}^n_\mu) + \delta\alpha\right)\end{aligned}$$

VI using skew Gaussian distribution



Example of Mixture of Exponential Family:

Consider a model with a Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})$.

$$p(\mathcal{D}, \mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})\prod_n p(\mathcal{D}_n|\mathbf{z})$$

We use a K-mixture of Gaussians shown below, where $\pi_K = 1 - \sum_{c=1}^{K-1}\pi_c$.

$$q(\mathbf{z}) = \sum_{w=1}^{K} \underbrace{\text{Cate}_K(w|\boldsymbol{\pi})}_{\pi_w}\mathcal{N}(\mathbf{z}|\mu_w, \boldsymbol{\Sigma}_w), \text{ where } \text{Cate}_K(w|\boldsymbol{\pi}) = \exp\left(\sum_{c=1}^{K-1}\mathbb{I}_c(w)\log\frac{\pi_c}{\pi_K} + \log\pi_K\right)$$

The natural parameter and expectation parameter are

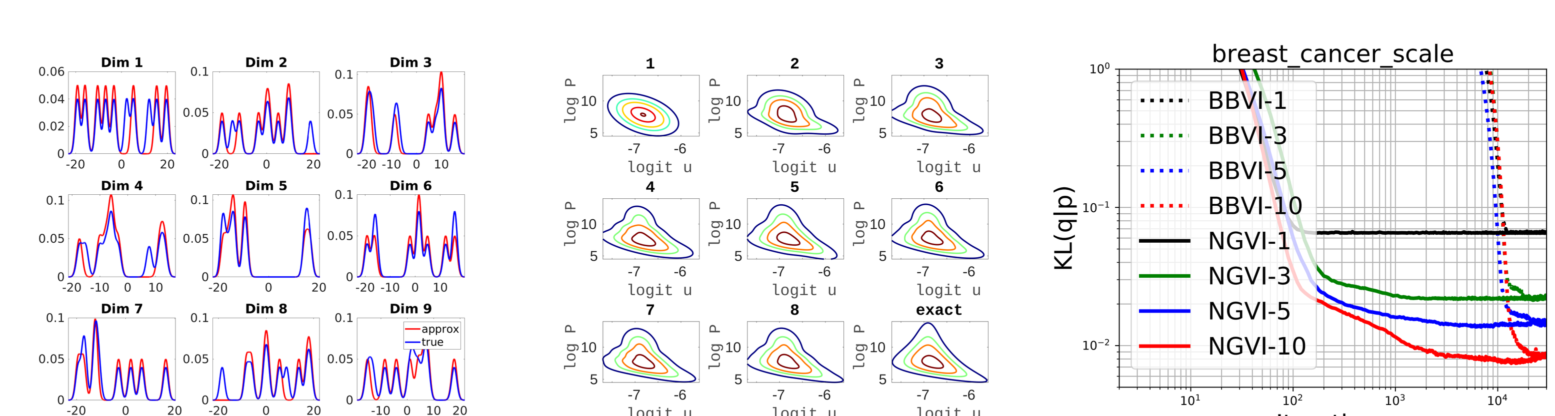
$$\boldsymbol{\lambda}_z = \left\{\boldsymbol{\Sigma}_c^{-1}\mu_c, -\tfrac{1}{2}\boldsymbol{\Sigma}_c^{-1}\right\}_{c=1}^{K}, \qquad \mathbf{m}_z = \left\{\pi_c\mu_c, \pi_c(\mu_c\mu_c^T + \boldsymbol{\Sigma}_c)\right\}_{c=1}^{K}$$
$$\boldsymbol{\lambda}_w = \left\{\log\frac{\pi_c}{\pi_K}\right\}_{c=1}^{K-1}, \qquad \mathbf{m}_w = \{\pi_c\}_{c=1}^{K-1}$$

The ELBO: $\mathcal{L} = \mathbb{E}_{q(z)}[-h(\mathbf{z})]$, where $h(\mathbf{z}) := \log[q(\mathbf{z})/p(\mathbf{z})] - \sum_n \log p(\mathcal{D}_n|\mathbf{z})$.

We can re-express the update in terms of $\{\mu_c\}_{c=1}^K$, $\{\boldsymbol{\Sigma}_c\}_{c=1}^K$, and $\{\pi_c\}_{c=1}^K$, where $\delta_c := \mathcal{N}(\mathbf{z}|\mu_c, \boldsymbol{\Sigma}_c)/\sum_{k=1}^K \pi_k\mathcal{N}(\mathbf{z}|\mu_k, \boldsymbol{\Sigma}_k)$ and $\mathbf{z}$ is generated from $q(\mathbf{z})$.

$$\text{NGVI}: \quad \begin{aligned}\boldsymbol{\Sigma}_c^{-1} &\leftarrow \boldsymbol{\Sigma}_c^{-1} + \beta\delta_c\left[\nabla_z^2 h(\mathbf{z})\right] \text{ for } c = 1, \dots, K\\ \mu_c &\leftarrow \mu_c - \beta\delta_c\boldsymbol{\Sigma}_c\left[\nabla_z h(\mathbf{z})\right] \text{ for } c = 1, \dots, K\\ \log(\pi_c/\pi_K) &\leftarrow \log(\pi_c/\pi_K) - \beta(\delta_c - \delta_K)h(\mathbf{z}) \text{ for } c = 1, \dots, K-1\end{aligned}$$

VI using finite mixture of Gaussians.



Example of Multi-linear Exponential Family Approximation:

Consider a Bayesian model $p(\mathcal{D}, \mathbf{Z})$. We use a matrix Gaussian distribution $\mathbf{Z} \in \mathcal{R}^{d\times p}$.

$$q(\mathbf{Z}) = \mathcal{MN}(\mathbf{Z}|\mathbf{W}, \mathbf{U}, \mathbf{V}), \text{ where } f(\mathbf{Z}, \mathbf{W}, \mathbf{U}^{-1}, \mathbf{V}^{-1}) = \text{Tr}\left(\mathbf{V}^{-1}(-\tfrac{1}{2}\mathbf{Z} + \mathbf{W})^T\mathbf{U}^{-1}\mathbf{Z}\right)$$

The natural parameter and expectation parameter are

$$\boldsymbol{\lambda}_1 = \mathbf{W}, \qquad \boldsymbol{\lambda}_2 = \mathbf{U}^{-1}, \qquad \boldsymbol{\lambda}_3 = \mathbf{V}^{-1}$$
$$\mathbf{m}_1 = \mathbf{U}^{-1}\mathbf{W}\mathbf{V}^{-1}, \qquad \mathbf{m}_2 = \tfrac{1}{2}\left(\mathbf{W}\mathbf{V}^{-1}\mathbf{W}^T - p\mathbf{U}\right), \qquad \mathbf{m}_3 = \tfrac{1}{2}\left(\mathbf{W}^T\mathbf{U}^{-1}\mathbf{W} - d\mathbf{V}\right)$$

Using the Gauss-Newton approximation to the Hessian matrix, we obtain the following update, where the gradient is $\mathbf{G} := \nabla_Z[-\log p(\mathcal{D}, \mathbf{Z}) + \log q(\mathbf{Z})]$ and $\mathbf{Z}$ is sampled from $q(\mathbf{Z})$.

$$\text{block NGVI}: \quad \begin{aligned}\mathbf{W} &\leftarrow \mathbf{W} - \beta_1\mathbf{U}\mathbf{G}\mathbf{V},\\ \mathbf{U}^{-1} &\leftarrow \mathbf{U}^{-1} + \beta_2\mathbf{G}\mathbf{V}\mathbf{G}^\top,\\ \mathbf{V}^{-1} &\leftarrow \mathbf{V}^{-1} + \beta_2\mathbf{G}^\top\mathbf{U}\mathbf{G},\end{aligned}$$

The update is similar to Shampoo (Gupta et al., 2018). If the prior $p(\mathbf{Z})$ is also a matrix-variate Gaussian distribution, the update resembles noisy K-FAC (Zhang et al. 2018).

More Examples: For Birnbaum-Saunders distribution, exponentially modified Gaussian, Student's t, symmetric normal inverse-Gaussian, please see the appendix of the paper.

Conclusion:

We propose a new type of simple and faster natural-gradient updates for several kinds of approximations outside the existing class of exponential-family distributions.