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 applications are usually restricted to exponential-family approximations.

- We present a new approach to obtain natural-gradient updates for several types of approximations outside the class of exponential-family distributions.
- Our approach enables the derivation of simple updates by introducing a new type of expectation-parameterization.
- Our results demonstrate faster convergence compared to existing block-box gradient methods.

2 VI using Exp-Family Components

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.

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

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

We further consider 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 $f(\mathbf{z},\boldsymbol{\lambda}_1,\dots,\boldsymbol{\lambda}_N)$ is a linear function w.r.t. each block $\boldsymbol{\lambda}_j$ given others.

Black-Box VI and Natural-Gradient VI:

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

Advantages of NGVI:

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

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

Challenges of NGVI: In general, NGVI could be complicated due to the inverse of the Fisher information matrix $[\mathbf{F}_z(\boldsymbol{\lambda}_z)]^{-1}$. Usually, NGVI does not admit a simple update outside the class of exponential family.

3 Simple Natural-gradient VI Update

NGVI can have a simple update in the following cases.

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

- Expectation parameters: $\mathbf{m}_w := \mathbb{E}_{q(w)}\left[\boldsymbol{\phi}_w(\mathbf{w})\right], \mathbf{m}_z := \mathbb{E}_{q(w,z)}\left[\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w})\right]$
- Natural parameters: $\boldsymbol{\lambda}_w, \boldsymbol{\lambda}_z$
- 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:

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

is equivalent to

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

Similarly, for a 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

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

For the j-th block, we define the following

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

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

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

is equivalent to

$$\text{BNGVI}: \boldsymbol{\lambda}_j^{t+1} = \boldsymbol{\lambda}_j^t + \beta\nabla_{\mathbf{m}_j}\mathcal{L}^t$$

The Jacobi Variant:

Instead of updating one block at each iteration, we can update all the blocks at once.

$$\text{BNGVI-J}: \boldsymbol{\lambda}_j^{t+1} = \boldsymbol{\lambda}_j^t + \beta\nabla_{\mathbf{m}_j}\mathcal{L}^t \quad \text{for all } j$$

This is equivalent to the following approximate natural-gradient update in natural parameters:

$$\begin{bmatrix}\boldsymbol{\lambda}_1^{t+1}\\ \cdots\\ \boldsymbol{\lambda}_N^{t+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\lambda}_1^{t}\\ \cdots\\ \boldsymbol{\lambda}_N^{t}\end{bmatrix} + \beta\underbrace{\text{block-diag}\left(\mathbf{F}(\boldsymbol{\lambda}_1^t,\dots,\boldsymbol{\lambda}_N^t)\right)^{-1}\begin{bmatrix}\nabla_{\boldsymbol{\lambda}_1}\mathcal{L}^t\\ \cdots\\ \nabla_{\boldsymbol{\lambda}_N}\mathcal{L}^t\end{bmatrix}}_{\text{Approximate natural gradient}},$$

where $\mathbf{F}(\boldsymbol{\lambda}_1,\dots,\boldsymbol{\lambda}_N) = -\mathbb{E}_{q(z)}\left[\nabla^2\log q(\mathbf{z}|\boldsymbol{\lambda}_1,\dots,\boldsymbol{\lambda}_N)\right]$.

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
We consider a model with a Student's t prior expressed as a scale mixture of Gaussians.

$$p(\mathcal{D},\mathbf{z},w) = \mathrm{InvGam}(w|a_0,a_0)\mathcal{N}(\mathbf{z}|\mathbf{0},w\mathbf{I})\prod_n p(\mathcal{D}_n|\mathbf{z})$$

We use the following mixture of exponential family distributions.

$$q(\mathbf{z},w) = \mathrm{InvGam}(w|a,a)\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma}), \text{ where } \mathbf{z}\in\mathcal{R}^d, w\in\mathcal{R}_{++}$$

The natural parameter and expectation parameter are

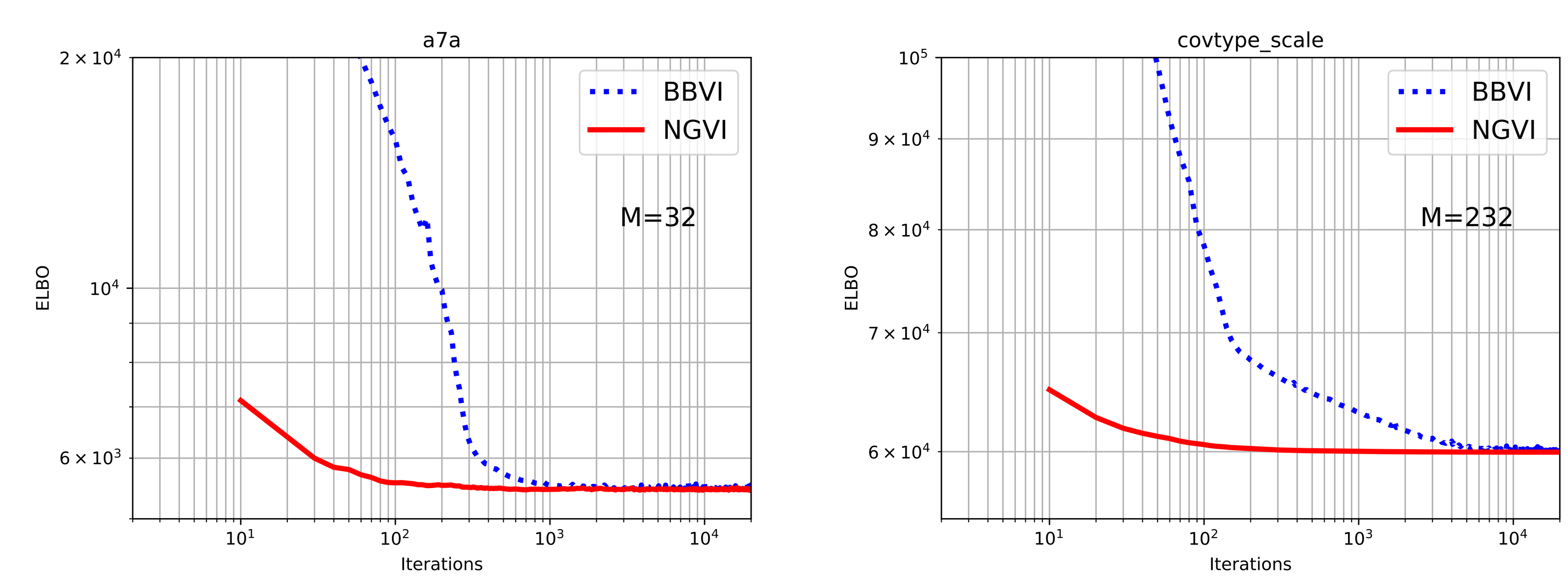
$$\boldsymbol{\lambda}_z = \left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\right\}, \qquad \mathbf{m}_z = \{\boldsymbol{\mu},\boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\Sigma}\}$$
$$\lambda_w = a, \qquad m_w = -1-(\log a - \psi(a))$$

The ELBO is $\mathcal{L} = \mathbb{E}_{q(z,w)}\left[\sum_n\log p(\mathcal{D}_n|\mathbf{z}) + \log\frac{\mathcal{N}(\mathbf{z}|\mathbf{0},w\mathbf{I})}{\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},w\boldsymbol{\Sigma})} + \log\frac{\mathrm{InvGam}(w|a_0,a_0)}{\mathrm{InvGam}(w|a,a)}\right]$.

We can re-express the update in terms of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}^{-1}$, and a.

$$\text{NGVI}: \quad a^{t+1} = (1-\beta)a^t + \beta\left(a_0 + \frac{\sum_n\nabla_a\mathbb{E}_{q^t(z,w)}\left[\log p_n(\mathcal{D}_n|\mathbf{z})\right]}{\nabla_a\mathbb{E}_{q^t(w)}\left[\phi_w(w)\right]}\right)$$
$$\left(\boldsymbol{\Sigma}^{t+1}\right)^{-1} = (1-\beta)\left(\boldsymbol{\Sigma}^t\right)^{-1} - 2\beta\nabla_{\Sigma}\mathbb{E}_{q^t(z,w)}\left[\sum_n\log p_n(\mathcal{D}_n|\mathbf{z})\right] + \beta\mathbf{I}$$
$$\boldsymbol{\mu}^{t+1} = \boldsymbol{\mu}^t + \beta\boldsymbol{\Sigma}^{t+1}\left(\nabla_{\mu}\mathbb{E}_{q^t(z,w)}\left[\sum_n\log p_n(\mathcal{D}_n|\mathbf{z})\right] - \boldsymbol{\mu}^t\right)$$

Bayesian logistic regression (t prior) with NGVI and BBVI.



Example of Mixture of Exponential Family:
We consider a model with a Gaussian prior.

$$p(\mathcal{D},\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0},\mathbf{S}_0)\prod_n p(\mathcal{D}_n|\mathbf{z})$$

We use the following distribution, which is the marginal distribution of a K-mixture of Gaussians shown below. Let $\pi_K = 1-\sum_{c=1}^{K-1}\pi_c$.

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

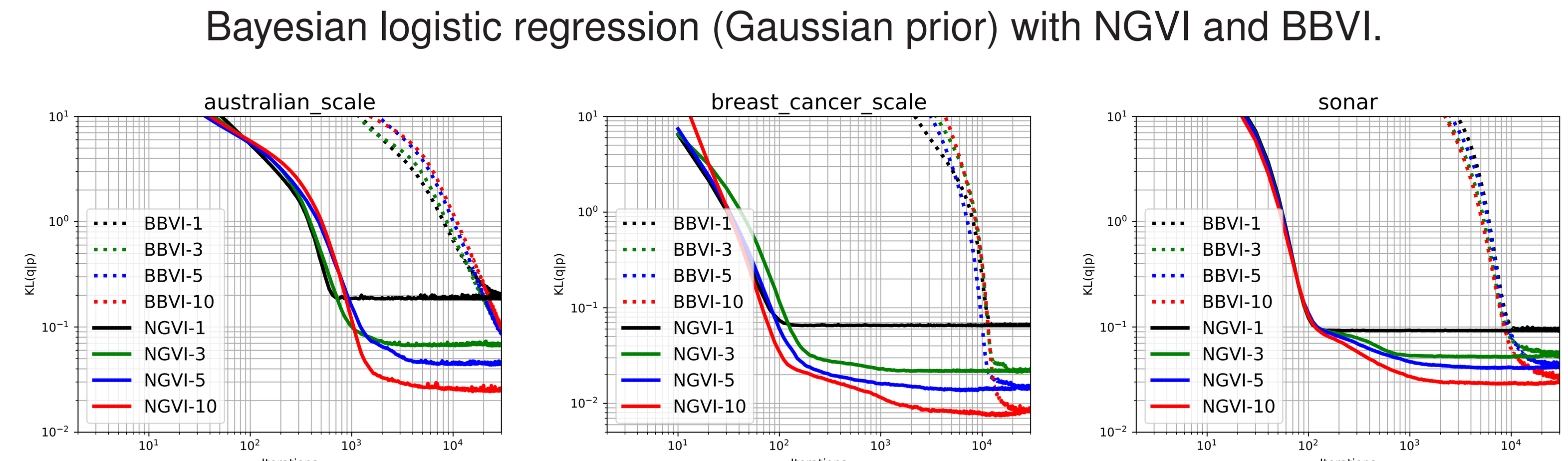
The natural parameter and expectation parameter are

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

The ELBO is $\mathcal{L} = \mathbb{E}_{q(z)}\left[\log p(\mathbf{z}) + \sum_n\log p(\mathcal{D}_n|\mathbf{z}) - \log q(\mathbf{z})\right]$.
We can re-express the update in terms of $\{\boldsymbol{\mu}_c\}_{c=1}^K$, $\{\boldsymbol{\Sigma}_c\}_{c=1}^K$, and $\{\pi_c\}_{c=1}^K$.

$$\text{NGVI}: \quad \log\frac{\pi_c^{t+1}}{\pi_K^{t+1}} = \log\frac{\pi_c^t}{\pi_K^t} + \beta\nabla_{\pi_c}\mathcal{L}^t \text{ for } c=1,\dots,K-1$$
$$\left(\boldsymbol{\Sigma}_c^{t+1}\right)^{-1} = \left(\boldsymbol{\Sigma}_c^t\right)^{-1} - \frac{2\beta}{\pi_c^t}\nabla_{\Sigma_c}\mathcal{L}^t \text{ for } c=1,\dots,K$$
$$\boldsymbol{\mu}_c^{t+1} = \boldsymbol{\mu}_c^t + \frac{\beta}{\pi_c^t}\boldsymbol{\Sigma}_c^{t+1}\nabla_{\mu_c}\mathcal{L}^t \text{ for } c=1,\dots,K$$

Bayesian logistic regression (Gaussian prior) with NGVI and BBVI.



Example of Multi-linear Exponential Family Approximation:
We consider a Bayesian model $p(\mathcal{D},\mathbf{Z})$. We use the following multi-linear exponential family distribution $\mathbf{Z}\in\mathcal{R}^{d\times p}$.

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

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

Using the Gauss-Newton approximation to the Hessian matrix, we obtain the update:

$$\text{BNGVI-J}: \quad \mathbf{W}^{t+1} = \mathbf{W}^t + \beta\mathbf{U}^t\mathbb{E}_{q^t(Z)}\left[\nabla_Z\log p(\mathcal{D},\mathbf{Z})\right]\mathbf{V}^t$$
$$\left(\mathbf{U}^{t+1}\right)^{-1} = \left(\mathbf{U}^t\right)^{-1} + \beta\mathbb{E}_{q^t(Z)}\left[\nabla_Z\log p(\mathcal{D},\mathbf{Z})\mathbf{V}^t\nabla_Z\log p(\mathcal{D},\mathbf{Z})^T\right]$$
$$\left(\mathbf{V}^{t+1}\right)^{-1} = \left(\mathbf{V}^t\right)^{-1} + \beta\mathbb{E}_{q^t(Z)}\left[\nabla_Z\log p(\mathcal{D},\mathbf{Z})^T\mathbf{U}^t\nabla_Z\log p(\mathcal{D},\mathbf{Z})\right]$$

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