Appendices

In the appendices, we will use the index notation and the Einstein summation notation introduced in Section 5.2.

A. A counter-example for Song et al. (2018)

We show that the update suggested by Song et al. (2018) does not stay in the constraint set while ours does.

Let's consider the following univariate Gaussian distribution under a BC parameterization $\boldsymbol{\lambda} = \{\mu, \sigma\}$, where σ denotes the standard deviation[16]. The constraint is $\Omega_1 = \mathbb{R}$ and $\Omega_2 = \mathbb{S}^1_{++}$. $\hat{g}^{(1)}$ and $\hat{g}^{(2)}$ are natural gradients for μ and σ, respectively.

$$q(z|\boldsymbol{\lambda}) = \exp\left\{-\tfrac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2 - \tfrac{1}{2}\log(2\pi) - \log(\sigma)\right\}$$

Recall that the Christoffel symbols of the second kind can be computed as ${\Gamma^c}_{ab} = F^{cd}\Gamma_{d,ab}$ where $\Gamma_{d,ab}$ is the Christoffel symbols of the first kind and F^{cd} is the entry of the inverse the FIM, $\mathbf{F}^{-1}$, at position (c, d).

Under this parameterization, the FIM and the Christoffel symbols of the second kind are given below, where the Christoffel symbols of the first kind are computed by using Eq. (17). The computation of the Christoffel symbols can be difficult since the parameterization is not a BCN parameterization.

$$F_{ab} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}, \; {\Gamma^1}_{ab} = \begin{bmatrix} 0 & -\frac{1}{\sigma} \\ -\frac{1}{\sigma} & 0 \end{bmatrix}, \; {\Gamma^2}_{ab} = \begin{bmatrix} \frac{1}{2\sigma} & 0 \\ 0 & -\frac{1}{\sigma} \end{bmatrix}$$

The update suggested by Song et al. (2018) is

$$\mu \leftarrow \mu - t\hat{g}^{(1)} - t\hat{g}^{(1)} - \frac{t\times t}{2}{\Gamma^1}_{ab}\hat{g}^{(a)}\hat{g}^{(b)} = \mu - t\hat{g}^{(1)} + \frac{t^2}{2}\left(\frac{2\hat{g}^{(1)}\hat{g}^{(2)}}{\sigma}\right)$$
$$\sigma \leftarrow \sigma - t\hat{g}^{(2)} - t\hat{g}^{(2)} - \frac{t\times t}{2}{\Gamma^2}_{ab}\hat{g}^{(a)}\hat{g}^{(b)} = \sigma - t\hat{g}^{(2)} + \frac{t^2}{2}\left(\frac{2(\hat{g}^{(2)})^2 - (\hat{g}^{(1)})^2}{2\sigma}\right)$$

Clearly, the updated σ does not always satisfy the positivity constraint $\mathbb{S}^1_{++}$.

As shown in Eq. (16), our rule can be used in not only a BCN parameterization but also a BC parameterization. Since every block contains only a scalar, we use global indexes such as $\lambda^{(i)} = \lambda^{a_i}$, $\hat{g}^{(i)} = \hat{g}^{[i]}$ and $\Gamma_{i,ii} = \Gamma_{a_i,b_ic_i}$ for notation simplicity. Note that ${\Gamma^1}_{11} = 0$ is the entry at the upper-left corner of ${\Gamma^1}_{ab}$ and ${\Gamma^2}_{22} = -\frac{1}{\sigma}$ is the entry at the lower-right corner of ${\Gamma^2}_{ab}$. In our update (see Eq. (16)), we can see the update automatically satisfies the constraint as shown below.

$$\underbrace{\mu}_{\lambda^{(1)}} \leftarrow \underbrace{\mu}_{\lambda^{(1)}} - \frac{t^2}{2}{\Gamma^1}_{11}\hat{g}^{(1)}\hat{g}^{(1)} = \mu - t\hat{g}^{(1)}$$
$$\underbrace{\sigma}_{\lambda^{(2)}} \leftarrow \underbrace{\sigma}_{\lambda^{(2)}} - \frac{t^2}{2}{\Gamma^2}_{22}\hat{g}^{(2)}\hat{g}^{(2)} = \sigma - t\hat{g}^{(2)} + \frac{t^2}{2}\left(\frac{(\hat{g}^{(2)})^2}{\sigma}\right) = \underbrace{\frac{1}{2\sigma}}_{>0}\left[\underbrace{\sigma^2}_{>0} + \underbrace{\left(\sigma - t\hat{g}^{(2)}\right)^2}_{\geq 0}\right]$$

As we discuss at Section 5.3 of the main text, only the block-wise Christoffel symbol $\Gamma_{i,ii}$ for each block i is required, which becomes essential for multivariate Gaussians and mixture of Gaussians.

[16] It is also used as an unconstrained parameterization of Gaussian distributions for BBVI. Technically, this parameterization has a positivity constraint, which is often ignored in practice. In multivariate cases, the Cholesky factor is used as an unconstrained parameterization, where the positivity constraint in the diagonal elements is often ignored.

Let's consider another BC parameterization $\boldsymbol{\lambda} = \{\mu, v\}$ for the Gaussian distribution, where $v = \sigma^2$ denotes the variance. Note that we consider the parameterization for univariate Gaussian. For multivariate Gaussian, see Appendix E.4. The underlying constraint is $\Omega = \mathbb{R} \times \mathbb{S}^1_{++}$. $\hat{g}^{(1)}$ and $\hat{g}^{(2)}$ are natural gradients for μ and v, respectively.

$$q(z|\boldsymbol{\lambda}) = \exp\left\{-\tfrac{1}{2}\frac{(z-\mu)^2}{v} - \tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log(v)\right\}$$

Under this parameterization, the FIM and the Christoffel symbols of the second kind are given below, where the Christoffel symbols of the first kind are computed by using Eq. (17). The computation of the Christoffel symbols can be difficult since the parameterization is not a BCN parameterization.

$$F_{ab} = \begin{bmatrix} \frac{1}{v} & 0 \\ 0 & \frac{1}{2v^2} \end{bmatrix}, \quad {\Gamma^1}_{ab} = \begin{bmatrix} 0 & -\frac{1}{2v} \\ -\frac{1}{2v} & 0 \end{bmatrix}, \quad {\Gamma^2}_{ab} = \begin{bmatrix} 1 & 0 \\ 0 & -\frac{1}{v} \end{bmatrix}$$

The update suggested by Song et al. (2018) is

$$\mu \leftarrow \mu - t\hat{g}^{(1)} - \frac{t^2}{2}{\Gamma^1}_{ab}\hat{g}^{(a)}\hat{g}^{(b)} = \mu - t\hat{g}^{(1)} + \frac{t^2}{2}\left(\frac{\hat{g}^{(1)}\hat{g}^{(2)}}{v}\right)$$
$$v \leftarrow v - t\hat{g}^{(2)} - \frac{t^2}{2}{\Gamma^2}_{ab}\hat{g}^{(a)}\hat{g}^{(b)} = v - t\hat{g}^{(2)} + \frac{t^2}{2}\left(\frac{(\hat{g}^{(2)})^2}{v} - (\hat{g}^{(1)})^2\right)$$

Obviously, the above updated v does not always satisfy the positivity constraint.

Similarly, we use global indexes such as $\lambda^{(i)} = \lambda^{a_i}$, $\hat{g}^{(i)} = \hat{g}^{[i]}$ and $\Gamma_{i,ii} = \Gamma_{a_i,b_ic_i}$ for notation simplicity since every block contains only a scalar. Note that ${\Gamma^1}_{11} = 0$ is the entry at the upper-left corner of ${\Gamma^1}_{ab}$ and ${\Gamma^2}_{22} = -\frac{1}{v}$ is the entry at the lower-right corner of ${\Gamma^2}_{ab}$. In our update (see Eq. (16)), we can see the update automatically satisfies the constraint as shown below.

$$\mu \leftarrow \mu - t\hat{g}^{(1)} - \frac{t^2}{2}{\Gamma^1}_{11}\hat{g}^{(1)}\hat{g}^{(1)} = \mu - t\hat{g}^{(1)}$$
$$v \leftarrow v - t\hat{g}^{(2)} - \frac{t^2}{2}{\Gamma^2}_{22}\hat{g}^{(2)}\hat{g}^{(2)} = v - t\hat{g}^{(2)} + \frac{t^2}{2}\left(\frac{(\hat{g}^{(2)})^2}{v}\right) = \underbrace{\frac{1}{2v}}_{>0}\left[\underbrace{v^2}_{>0} + \underbrace{\left(v - t\hat{g}^{(2)}\right)^2}_{\geq 0}\right]$$

B. Riemannian Optimization

B.1. Proof of Lemma 1

Let's consider a parameterization $\boldsymbol{\lambda} := \{\boldsymbol{\lambda}^{[1]}, \ldots, \boldsymbol{\lambda}^{[m]}\}$ with m blocks for a statistical manifold with metric $\mathbf{F}$. We first define a BC parameterization $\boldsymbol{\lambda}$ for a general metric $\mathbf{F}$.

Definition 1 Block Coordinate Parameterization: *A parameterization is block coordinate (BC) if the metric* $\mathbf{F}$ *under this parameterization is block-diagonal according to the block structure of the parameterization.*

Recall that we use the following block notation: ${\Gamma^{c_i}}_{a_ib_i}\hat{g}^{a_i}\hat{g}^{b_i} := \sum_{a\in[i]}\sum_{b\in[i]}{\Gamma^{(c_i)}}_{ab}\hat{g}^a\hat{g}^b$ where $[i]$ denotes the index set of block i, (c_i) is the corresponding global index of c_i, and a and b are global indexes.

Now, we prove Lemma 1.

Proof: By the definition of a Riemannian gradient $\hat{\boldsymbol{g}}$, we have

$$\hat{g}^{a_i} = \sum_b F^{(a_i)b}g_b = \sum_{b\in[i]} F^{(a_i)b}g_b + \sum_{b\notin[i]}\underbrace{F^{(a_i)b}}_{0}g_b = \sum_{b\in[i]} F^{(a_i)b}g_b = F^{a_ib_i}g_{b_i},$$

where in the second step, $F^{(a_i)b} = 0$ for any $b \notin [i]$ (see (18) for visualization) since the parameterization is BC, and we use the definition of the block summation notation in the last step.

Similarly, we have

$$\Gamma^{c_i}_{a_i b_i} = \sum_d F^{(c_i)d}\Gamma_{d,(a_i)(b_i)} = \sum_{d\in[i]} F^{(c_i)d}\Gamma_{d,(a_i)(b_i)} + \sum_{d\notin[i]} \underbrace{F^{(c_i)d}}_{0}\Gamma_{d,(a_i)(b_i)} = \sum_{d\in[i]} F^{(c_i)d}\Gamma_{d,(a_i)(b_i)} = F^{c_i d_i}\Gamma_{d_i,a_i b_i}$$

□

B.2. NGD is a First-order Approximation of $\mathbf{R}(t)$

Now, we assume parameterization $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^{[1]}, \dots, \boldsymbol{\lambda}^{[m]}\}$ is a BC parameterization with m blocks. Recall that we define the curve $\mathbf{R}(t)$ as $\mathbf{R}(t) := \{\mathbf{R}^{[1]}(t), \dots, \mathbf{R}^{[m]}(t)\}$, where $\mathbf{R}^{[i]}(t)$ is the solution of following ODE for block i.

$$\dot{R}^{\,c_i}(0) = -F^{c_i a_i} g_{a_i} \;;\; R^{\,c_i}(0) = \lambda^{c_i}$$
$$\ddot{R}^{\,c_i}(t) = -\Gamma^{c_i}_{a_i b_i}(t)\dot{R}^{\,a_i}(t)\dot{R}^{\,b_i}(t)$$

where $R^{\,c_i}(0)$, $\dot{R}^{\,c_i}(0)$, $\ddot{R}^{\,c_i}(t)$ respectively denote the c-th entry of $\mathbf{R}^{[i]}(0)$, $\dot{\mathbf{R}}^{[i]}(0)$, and $\ddot{\mathbf{R}}^{[i]}(t)$; $\Gamma^{c_i}_{a_i b_i}(t) := \Gamma^{c_i}_{a_i b_i}\big|^{\lambda^{[-i]}=R^{[-i]}(0)}_{\lambda^{[i]}=R^{[i]}(t)}$.

Recall that $F^{c_i a_i}$ is the entry of $(\mathbf{F}^{[i]})^{-1}$ at position (c, a), where $\mathbf{F}^{[i]}$ is the i-th block of $\mathbf{F}$. Note that $\mathbf{F}$ and $\hat{\boldsymbol{g}}$ are computed at $\boldsymbol{\lambda} = \mathbf{R}(0)$. Since $\boldsymbol{\lambda}$ is a BC parameterization, by Lemma 1, we have $F^{c_i a_i} g_{a_i} = \hat{g}^{c_i}$.

Therefore, when $\mathbf{F}$ is the FIM, the first-order approximation of $\mathbf{R}(t)$ at $t_0 = 0$ is also a NGD update as shown below.

$$\begin{aligned}\lambda^{c_i} &\leftarrow R^{c_i}(t_0) + \dot{R}^{c_i}(t_0)(t - t_0)\\ &= \lambda^{c_i} - t\hat{g}^{c_i}\end{aligned}$$

C. Summary of Approximations Considered in This Work

Recall that we give Assumption 1-3 for exponential family distributions in Section 3. We also extend Assumption 1-3 to exponential family mixtures as shown in Appendix I.

In Appendix H, F, G, E, J, K, we show that Assumption 1-3 are satisfied and the additional term for each approximation is simplified. In the corresponding appendix, we also show how to compute natural gradients with the (implicit) reparameterization trick for each approximation listed in Table 2.

D. Exponential Family (EF) Approximation

D.1. Christoffel Symbols

We first show how to simplify the Christoffel symbols of the first kind. The FIM and the corresponding Christoffel symbols of the first kind are defined as follows.

$$F_{ab} := -\mathbb{E}_{q(z|\lambda)}\left[\partial_a \partial_b \log q(\mathbf{z}|\boldsymbol{\lambda})\right]; \quad \Gamma_{d,ab} := \tfrac{1}{2}\left[\partial_a F_{bd} + \partial_b F_{ad} - \partial_d F_{ab}\right]$$

where we denote $\partial_a = \partial_{\lambda^a}$ for notation simplicity.

Since $\partial_a F_{bd} = -\mathbb{E}_{q(z|\lambda)}\left[\partial_b \partial_d \log q(\mathbf{z}|\boldsymbol{\lambda})\partial_a \log q(\mathbf{z}|\boldsymbol{\lambda})\right] - \mathbb{E}_{q(z|\lambda)}\left[\partial_a \partial_b \partial_d \log q(\mathbf{z}|\boldsymbol{\lambda})\right]$, the Christoffel symbols of the first kind induced by the FIM can be computed as follows , where $\boldsymbol{\lambda}$ can be any parameterization.

$$\begin{aligned}\Gamma_{d,ab} = \tfrac{1}{2}\Big[&\mathbb{E}_{q(z|\lambda)}\left[\partial_a \partial_b \log q(\mathbf{z}|\boldsymbol{\lambda})\partial_d \log q(\mathbf{z}|\boldsymbol{\lambda})\right] - \mathbb{E}_{q(z|\lambda)}\left[\partial_b \partial_d \log q(\mathbf{z}|\boldsymbol{\lambda})\partial_a \log q(\mathbf{z}|\boldsymbol{\lambda})\right]\\ &- \mathbb{E}_{q(z|\lambda)}\left[\partial_a \partial_d \log q(\mathbf{z}|\boldsymbol{\lambda})\partial_b \log q(\mathbf{z}|\boldsymbol{\lambda})\right] - \mathbb{E}_{q(z|\lambda)}\left[\partial_a \partial_b \partial_d \log q(\mathbf{z}|\boldsymbol{\lambda})\right]\Big]\end{aligned} \tag{17}$$

Note that Eq 17 is also applied to a general distribution beyond exponential family. However, the Christoffel symbol is not easy to compute due to extra integrations in Eq 17 and the FIM can be singular in general. The Christoffel symbol could be easy to compute for an exponential family distribution under a BCN parameterization since we compute the symbol via differentiation without the extra integrations. Moreover, the FIM is always positive-definite under a BCN parameterization. Theorem 3 show this.

[17] We do not compute the additional term in MOG since $\boldsymbol{\lambda}_w \in \mathbb{R}^{K-1}$ is unconstrained.

Table 2. Summary of the Proposed Updates Induced by Our Rule in Various Approximations

Approximation	Parameterization ($\boldsymbol{\lambda}$)	Constraints	Additional Term
Inverse Gaussian (Appendix H)	$\lambda^{(1)} = \beta^2$	$\lambda^{(1)} \in \mathbb{S}^1_{++}$	$\frac{t^2}{2}\left(\frac{3}{4\lambda^{(1)}}\right)\left(\hat{g}^{(1)}\right)^2$
	$\lambda^{(2)} = \alpha$	$\lambda^{(2)} \in \mathbb{S}^1_{++}$	$\frac{t^2}{2}\left(\frac{1}{\lambda^{(2)}}\right)\left(\hat{g}^{(2)}\right)^2$
Gamma (Appendix F)	$\lambda^{(1)} = \alpha$	$\lambda^{(1)} \in \mathbb{S}^1_{++}$	$-\frac{t^2}{2}\frac{\partial^2_{\lambda^{(1)}}\psi(\lambda^{(1)}) + \frac{1}{\left(\lambda^{(1)}\right)^2}}{2\left(\partial_{\lambda^{(1)}}\psi(\lambda^{(1)}) - \frac{1}{\lambda^{(1)}}\right)}\left(\hat{g}^{(1)}\right)^2$
	$\lambda^{(2)} = \frac{\beta}{\alpha}$	$\lambda^{(2)} \in \mathbb{S}^1_{++}$	$\frac{t^2}{2}\left(\frac{1}{\lambda^{(2)}}\right)\left(\hat{g}^{(2)}\right)^2$
Exponential (Appendix G)	$\lambda^{(1)} = \lambda$	$\lambda^{(1)} \in \mathbb{S}^1_{++}$	$\frac{t^2}{2}\left(\frac{1}{\lambda^{(1)}}\right)\left(\hat{g}^{(1)}\right)^2$
Multivariate Gaussian (Appendix E)	$\boldsymbol{\lambda}^{[1]} = \boldsymbol{\mu}$	$\boldsymbol{\lambda}^{[1]} \in \mathbb{R}^d$	$\mathbf{0}$
	$\boldsymbol{\lambda}^{[2]} = \boldsymbol{\Sigma}^{-1}$	$\boldsymbol{\lambda}^{[2]} \in \mathbb{S}^{d\times d}_{++}$	$\frac{t^2}{2}\hat{\boldsymbol{g}}^{[2]}\left(\boldsymbol{\lambda}^{[2]}\right)^{-1}\hat{\boldsymbol{g}}^{[2]}$
Mixture of Gaussians (Appendix J)	$\{\boldsymbol{\lambda}_c^{[1]}\}_{c=1}^K = \{\boldsymbol{\mu}_c\}_{c=1}^K$	$\boldsymbol{\lambda}_c^{[1]} \in \mathbb{R}^d$	$\mathbf{0}$
	$\{\boldsymbol{\lambda}_c^{[2]}\}_{c=1}^K = \{\boldsymbol{\Sigma}_c^{-1}\}_{c=1}^K$	$\boldsymbol{\lambda}_c^{[2]} \in \mathbb{S}^{d\times d}_{++}$	$\frac{t^2}{2}\hat{\boldsymbol{g}}_c^{[2]}\left(\boldsymbol{\lambda}_c^{[2]}\right)^{-1}\hat{\boldsymbol{g}}_c^{[2]}$
	$\boldsymbol{\lambda}_w = \{\log(\pi_c/(1-\sum_{k=1}^{K-1}\pi_k))\}_{c=1}^{K-1}$	$\boldsymbol{\lambda}_w \in \mathbb{R}^{K-1}$	$\mathbf{0}$[17]
Skew Gaussian (Appendix K)	$\boldsymbol{\lambda}^{[1]} = \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{bmatrix}$	$\boldsymbol{\lambda}^{[1]} \in \mathbb{R}^{2d}$	$\mathbf{0}$
	$\boldsymbol{\lambda}^{[2]} = \boldsymbol{\Sigma}^{-1}$	$\boldsymbol{\lambda}^{[2]} \in \mathbb{S}^{d\times d}_{++}$	$\frac{t^2}{2}\hat{\boldsymbol{g}}^{[2]}\left(\boldsymbol{\lambda}^{[2]}\right)^{-1}\hat{\boldsymbol{g}}^{[2]}$

D.2. Proof of Theorem 3

In this case, $q(\mathbf{z}|\boldsymbol{\lambda})$ is an EF distribution. Since $\boldsymbol{\lambda}$ is a BCN parameterization, given that $\boldsymbol{\lambda}^{[-i]}$ is known, $q(\mathbf{z}|\boldsymbol{\lambda})$ is a one-parameter EF distribution as

$$q(\mathbf{z}|\boldsymbol{\lambda}) = h_i(\mathbf{z}, \boldsymbol{\lambda}^{[-i]}) \exp\left[\langle \boldsymbol{\phi}_i(\mathbf{z}, \boldsymbol{\lambda}^{[-i]}), \boldsymbol{\lambda}^{[i]}\rangle - A(\boldsymbol{\lambda})\right]$$

Therefore, we have the following identities given $\boldsymbol{\lambda}^{[-i]}$ is known.

$$\partial_{a_i}\partial_{b_i}\log q(\mathbf{z}|\boldsymbol{\lambda}) = -\partial_{a_i}\partial_{b_i}A(\boldsymbol{\lambda}); \quad \mathbb{E}_{q(z|\lambda)}\left[\partial_{a_i}\log q(\mathbf{z}|\boldsymbol{\lambda})\right] = 0$$

where $\partial_{a_i} = \partial_{\lambda^{a_i}}$ for notation simplicity.

Using the above identities, we have

$$\mathbb{E}_{q(z|\lambda)}\left[\partial_{a_i}\partial_{b_i}\log q(\mathbf{z}|\boldsymbol{\lambda})\partial_{d_i}\log q(\mathbf{z}|\boldsymbol{\lambda})\right] = -\partial_{a_i}\partial_{b_i}A(\boldsymbol{\lambda})\underbrace{\mathbb{E}_{q(z|\lambda)}\left[\partial_{d_i}\log q(\mathbf{z}|\boldsymbol{\lambda})\right]}_{0} = 0$$

Therefore, by Eq. (17), Γ_{d_i,a_ib_i} can be computed as follows

$$\Gamma_{d_i,a_ib_i} = -\tfrac{1}{2}\mathbb{E}_{q(z|\lambda)}\left[\partial_{a_i}\partial_{b_i}\partial_{d_i}\log q(\mathbf{z}|\boldsymbol{\lambda})\right] = \tfrac{1}{2}\partial_{a_i}\partial_{b_i}\partial_{d_i}A(\boldsymbol{\lambda})$$

Let $\mathbf{m}_{[i]} = \mathbb{E}_{q(z|\lambda)}\left[\boldsymbol{\phi}_i(\mathbf{z})\right]$ denote the block coordinate expectation (BCE) parameter. We have

$$0 = \mathbb{E}_{q(z|\lambda)}\left[\partial_{a_i}\log q(\mathbf{z}|\boldsymbol{\lambda})\right] = m_{a_i} - \partial_{a_i}A(\boldsymbol{\lambda})$$

where m_{a_i} denotes the a-th element of $\mathbf{m}_{[i]}$.

Therefore, we know that $m_{a_i} = \partial_{a_i}A(\boldsymbol{\lambda})$

Recall that the i-th block of $\mathbf{F}$ denoted by $\mathbf{F}^{[i]}$, can be computed as

$$F_{a_ib_i} = -\mathbb{E}_{q(z|\lambda)}\left[\partial_{b_i}\partial_{a_i}\log q(\mathbf{z}|\boldsymbol{\lambda})\right] = \partial_{b_i}\partial_{a_i}A(\boldsymbol{\lambda}) = \partial_{b_i}\left[\partial_{a_i}A(\boldsymbol{\lambda})\right] = \partial_{\lambda^{b_i}}m_{a_i}$$

where $\partial_{b_i} = \partial_{\lambda^{b_i}}$ is for notation simplicity.

Recall that $\boldsymbol{\lambda}$ is a BC parameterization with n blocks and $\mathbf{F}$ is block diagonal as shown below.

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}^{[1]} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{F}^{[n]} \end{bmatrix} \tag{18}$$

Recall that F^{ab} denotes the element of $\mathbf{F}^{-1}$ with global index (a, b) and $F^{a_i b_i}$ denotes the element of $\left(\mathbf{F}^{[i]}\right)^{-1}$ with local index (a, b) in block i.

If $\mathbf{F}^{[i]}$ is positive definite everywhere, we have

$$F^{a_i b_i} = \partial_{m_{a_i}} \lambda^{b_i}$$

Note that $\mathbf{F}^{[i]}$ is positive definite everywhere when $q(\mathbf{z}|\boldsymbol{\lambda}^{[i]}, \boldsymbol{\lambda}^{[-i]})$ is a one-parameter minimal EF distribution given $\boldsymbol{\lambda}^{[-i]}$ is known (See Theorem 1 of Lin et al. (2019a)).

By Lemma 1, Riemannian gradient $\hat{g}^{a_i}$ can be computed as

$$\hat{g}^{a_i} = F^{a_i b_i} g_{b_i} = \left[\partial_{m_{a_i}} \lambda^{b_i}\right] \left[\partial_{\lambda^{b_i}} \mathcal{L}\right] = \partial_{m_{a_i}} \mathcal{L}$$

where $g_{b_i} = \partial_{\lambda^{b_i}} \mathcal{L}$ is a Euclidean gradient.

E. Example: Gaussian Approximation

We consider the following parameterization $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \mathbf{S}\}$, where $\boldsymbol{\mu}$ is the mean and $\mathbf{S}$ is the precision. The open-set constraint is $\Omega_1 = \mathbb{R}^d$ and $\Omega_2 = \mathbb{S}_{++}^{d \times d}$. Under this parameterization, the distribution can be expressed as below.

$$q(z|\boldsymbol{\lambda}) = \exp\Big(-\tfrac{1}{2}\mathbf{z}^T\mathbf{S}\mathbf{z} + \mathbf{z}^T\mathbf{S}\boldsymbol{\mu} - A(\boldsymbol{\lambda})\Big)$$

where $A(\boldsymbol{\lambda}) = \frac{1}{2}\left[\boldsymbol{\mu}^T\mathbf{S}\boldsymbol{\mu} - \log|\mathbf{S}/(2\pi)|\right]$

Lemma 2 *The Fisher information matrix under this parameterization is block diagonal with two blocks*

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_\mu & \underbrace{\mathbf{0}}_{\mathbf{F}_{\mu S}} \\ & \mathbf{F}_S \end{bmatrix},$$

where $\mathbf{F}_{\mu S} = -\mathbb{E}_{q(z)}\left[\partial_{\text{vec}(S)}\partial_\mu \log q(\mathbf{z}|\boldsymbol{\mu}, \mathbf{S})\right]$ *and* $\mathbf{F}_S = -\mathbb{E}_{q(z)}\left[\partial^2_{\text{vec}(S)} \log q(\mathbf{z}|\boldsymbol{\mu}, \mathbf{S})\right]$.

Therefore, $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \mathbf{S}\}$ *is a BC parameterization.*

Proof: We denote the i-th element of $\boldsymbol{\mu}$ using μ^i. Similarly, we denote the element of $\mathbf{S}$ at position (j, k) using S^{jk}. We prove this statement by showing cross terms in the Fisher information matrix denoted by $\mathbf{F}_{\mu S}$ are all zeros. To show $\mathbf{F}_{\mu S} = \mathbf{0}$, it is equivalent to show $-\mathbb{E}_{q(z|\lambda)}\left[\partial_{S^{jk}}\partial_{\mu^i} \log q(\mathbf{z}|\boldsymbol{\lambda})\right] = 0$ each μ^i and S^{jk}.

Notice that $\mathbb{E}_{q(z|\lambda)}[\mathbf{z}] = \boldsymbol{\mu}$. We can obtain the above expression since

$$\begin{aligned} \mathbb{E}_{q(z|\lambda)}\left[\partial_{S^{jk}}\partial_{\mu^i} \log q(\mathbf{z}|\boldsymbol{\lambda})\right] &= \mathbb{E}_{q(z|\lambda)}\left[\partial_{S^{jk}}\left(\mathbf{z}^T\mathbf{S}\mathbf{e}_i - \mathbf{e}_i^T\mathbf{S}\boldsymbol{\mu}\right)\right] \\ &= \mathbb{E}_{q(z|\lambda)}\left[\left(\mathbf{z}^T\mathbf{I}_{jk}\mathbf{e}_i - \mathbf{e}_i^T\mathbf{I}_{jk}\boldsymbol{\mu}\right)\right] \\ &= \mathbb{E}_{q(z|\lambda)}\left[\left(\mathbf{e}_i^T\mathbf{I}_{jk}\left(\mathbf{z} - \boldsymbol{\mu}\right)\right)\right] \\ &= \mathbf{e}_i^T\mathbf{I}_{jk}\underbrace{\mathbb{E}_{q(z|\lambda)}\left[\mathbf{z} - \boldsymbol{\mu}\right]}_{\mathbf{0}} = 0 \end{aligned}$$

where $\mathbf{e}_i$ denotes an one-hot vector where all entries are zeros except the i-th entry with value 1, and $\mathbf{I}_{jk}$ denotes an one-hot matrix where all entries are zeros except the entry at position (j, k) with value 1.

The above expression also implies that $\mathbb{E}_{q(z|\lambda)}\left[\partial_S \partial_{\mu^i} \log q(\mathbf{z}|\boldsymbol{\lambda})\right] = \mathbf{0}$. □

Now, we show that $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ is also a BC parameterization. Note that

$$-\mathbb{E}_{q(z|\lambda)}\left[\partial_{\Sigma^{jk}} \partial_{\mu^i} \log q(\mathbf{z}|\boldsymbol{\lambda})\right] = -\mathbb{E}_{q(z|\lambda)}\left[\mathrm{Tr}\left\{(\partial_{\Sigma^{jk}} \mathbf{S}) \partial_S \partial_{\mu^i} \log q(\mathbf{z}|\boldsymbol{\lambda})\right\}\right] = -\mathrm{Tr}\Big\{(\partial_{\Sigma^{jk}} \mathbf{S}) \underbrace{\mathbb{E}_{q(z|\lambda)}\left[\partial_S \partial_{\mu^i} \log q(\mathbf{z}|\boldsymbol{\lambda})\right]}_{\mathbf{0}}\Big\} = 0.$$

Since $\mathbf{F}_{\mu\Sigma} = -\mathbb{E}_{q(z|\lambda)}\left[\partial_{\mathrm{vec}(\Sigma)} \partial_\mu \log q(\mathbf{z}|\boldsymbol{\lambda})\right]$ and $-\mathbb{E}_{q(z|\lambda)}\left[\partial_{\Sigma^{jk}} \partial_{\mu^i} \log q(\mathbf{z}|\boldsymbol{\lambda})\right] = 0$ from above expression for any i, j, and k, we have $\mathbf{F}_{\mu\Sigma} = \mathbf{0}$. Therefore, $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ is also a BC parameterization since the cross terms of FIM under this new parameterization denoted by $\mathbf{F}_{\mu\Sigma}$ are zeros.

We denote the Christoffel symbols of the first kind and the second kind for $\boldsymbol{\mu}$ as $\Gamma_{a_1, b_1 c_1}$ and $\Gamma^{a_1}{}_{b_1 c_1}$, respectively.

Lemma 3 *All entries of* $\Gamma^{a_1}{}_{b_1 c_1}$ *are zeros.*

Proof: We will prove this by showing that all entries of $\Gamma_{a_1, b_1 c_1}$ are zeros. For notation simplicity, we use $\Gamma_{a,bc}$ to denote $\Gamma_{a_1, b_1 c_1}$ in the proof. Let μ^a denote the a-th element of $\boldsymbol{\mu}$. The following expression holds for any valid a, b, and c.

$$\Gamma_{a,bc} = \tfrac{1}{2}\mathbb{E}_{q(z|\lambda)}\left[\partial_{\mu^b} \partial_{\mu^c} \partial_{\mu^a} A(\boldsymbol{\lambda})\right] = 0$$

We can obtain the above expression since

$$\mathbb{E}_{q(z|\lambda)}\left[\partial_{\mu^b} \partial_{\mu^c} \partial_{\mu^a} A(\boldsymbol{\lambda})\right] = \mathbb{E}_{q(z|\lambda)}\left[\partial_{\mu^b} \partial_{\mu^c} \left(\mathbf{e}_a^T \mathbf{S} \boldsymbol{\mu}\right)\right] = \mathbb{E}_{q(z|\lambda)}\left[\partial_{\mu^b} \left(\mathbf{e}_a^T \mathbf{S} \mathbf{e}_c\right)\right] = 0$$

where in the last step we use the fact that $\mathbf{S}$, $\mathbf{e}_a$, and $\mathbf{e}_c$ do not depend on $\boldsymbol{\mu}$. □

Similarly, we denote the Christoffel symbols of the second kind for $\mathrm{vec}(\mathbf{S})$ as $\Gamma^{a_2}{}_{b_2 c_2}$. Note that $\mathbf{S}$ is now a matrix. It is possible but tedious to directly compute the Christoffel symbol and element-wisely validate the expression of the additional term for $\mathbf{S}$. Below, we give an alternative approach to identify the additional term for $\mathbf{S}$ as shown in the proof of Lemma 4.

Recall that $\mathbf{R}^{[2]}(t)$ is the solution of the following ODE for block $\mathrm{vec}(\mathbf{S})$:

$$\dot{R}^{\,a_2}(0) = -\hat{g}^{a_2} \;;\; R^{\,a_2}(0) = S^{a_2}$$
$$\ddot{R}^{\,a_2}(t) = -\Gamma^{a_2}{}_{b_2 c_2}(t) \dot{R}^{\,b_2}(t) \dot{R}^{\,c_2}(t),$$

where $R^{\,a_2}(t)$ denotes the a-th element of $\mathbf{R}^{[2]}(t)$ and S^{a_2} denotes the a-th entry of $\mathrm{vec}(\mathbf{S})$.

Lemma 4 *The additional term for* $\mathbf{S}$ *is* $\mathrm{Mat}(\Gamma^{a_2}{}_{b_2 c_2} \hat{g}^{b_2} \hat{g}^{c_2}) = -\hat{\boldsymbol{g}}^{[2]} \mathbf{S}^{-1} \hat{\boldsymbol{g}}^{[2]}$ *where* $\hat{g}^{a_2}$ *denotes the* a*-th element of* $\mathrm{vec}(\hat{\boldsymbol{g}}^{[2]})$.

Proof: As discussed in Sec 5, $\mathbf{R}^{[i]}(t)$ is a (block coordinate) geodesic given $\boldsymbol{\lambda}^{[-i]}$ is known. In this case, given that $\boldsymbol{\mu}$ is known, $\mathbf{R}^{[2]}(t)$ has the following closed-form expression (Pennec et al., 2006; Fletcher & Joshi, 2004; Minh & Murino, 2017).

$$\mathrm{Mat}(\mathbf{R}^{[2]}(t)) = \mathbf{U}\mathrm{Exp}(t\mathbf{U}^{-1}\hat{\boldsymbol{g}}^{[2]}\mathbf{U}^{-1})\mathbf{U}$$

where $\mathbf{U} = \mathbf{S}^{\frac{1}{2}}$ denotes the matrix square root and $\mathrm{Exp}(\mathbf{X}) := \mathbf{I} + \sum_{n=1}^{\infty} \frac{\mathbf{X}^n}{n!}$ denotes the matrix exponential function.[18]

[18] The function is well-defined since the matrix series is absolutely convergent element-wisely.

The additional term for $\mathbf{S}$ can be obtained as follows.

$$\begin{aligned}
-\mathrm{Mat}(\Gamma^{a_2}{}_{b_2 c_2}\hat{g}^{b_2}\hat{g}^{c_2}) &= \mathrm{Mat}(\ddot{\mathbf{R}}^{[2]}(0)) \\
&= \mathrm{Mat}(\nabla_t^2 \mathbf{R}^{[2]}(t)\big|_{t=0}) \\
&= \nabla_t^2 \mathrm{Mat}(\mathbf{R}^{[2]}(t))\big|_{t=0} \\
&= \nabla_t^2 \Big(\mathbf{U}\mathrm{Exp}(\mathbf{U}^{-1} t\hat{\boldsymbol{g}}^{[2]}\mathbf{U}^{-1})\mathbf{U}\Big)\big|_{t=0} \\
&= \mathbf{U}\nabla_t^2 \Big(\mathrm{Exp}(\mathbf{U}^{-1} t\hat{\boldsymbol{g}}^{[2]}\mathbf{U}^{-1})\Big)\big|_{t=0}\mathbf{U} \\
&= \mathbf{U}(\mathbf{U}^{-1}\hat{\boldsymbol{g}}^{[2]}\mathbf{U}^{-1})(\mathbf{U}^{-1}\hat{\boldsymbol{g}}^{[2]}\mathbf{U}^{-1})\mathbf{U} \\
&= \mathbf{U}(\mathbf{U}^{-1}\hat{\boldsymbol{g}}^{[2]}\mathbf{S}^{-1}\hat{\boldsymbol{g}}^{[2]}\mathbf{U}^{-1})\mathbf{U} \\
&= \hat{\boldsymbol{g}}^{[2]}\mathbf{S}^{-1}\hat{\boldsymbol{g}}^{[2]}
\end{aligned}$$

where we use the following expression to move from step 5 to step 6.

$$\nabla_t^2 \mathrm{Exp}(t\mathbf{X})\big|_{t=0} = \nabla_t^2 \left(\mathbf{I} + \sum_{n=1}^{\infty} \frac{(t\mathbf{X})^n}{n!}\right)\big|_{t=0} = \mathbf{X}^2$$

□

Finally, by Lemma 3 and 4, the update induced by the proposed rule is

$$\begin{aligned}
\mu^c &\leftarrow \mu^c - t\hat{g}^{c_1} - \frac{t\times t}{2}\overbrace{\Gamma^{c_1}_{a_1 b_1}}^{0}\hat{g}^{a_1}\hat{g}^{b_1} \\
s^c &\leftarrow s^c - t\hat{g}^{c_2} - \frac{t\times t}{2}\Gamma^{c_2}_{a_2 b_2}\hat{g}^{a_2}\hat{g}^{b_2}
\end{aligned}$$

where s^c is the c-th element of $\mathrm{vec}(\mathbf{S})$.

Therefore, we have

$$\begin{aligned}
\boldsymbol{\mu} &\leftarrow \overbrace{\boldsymbol{\mu}}^{\mathrm{vec}(\mu^c)} - t\ \overbrace{\hat{\boldsymbol{g}}^{[1]}}^{\mathrm{vec}(\hat{g}^{c_1})} \\
\mathbf{S} &\leftarrow \underbrace{\mathbf{S}}_{\mathrm{Mat}(s^c)} - t\ \underbrace{\hat{\boldsymbol{g}}^{[2]}}_{\mathrm{Mat}(\hat{g}^{c_2})} + \frac{t\times t}{2}\ \underbrace{\hat{\boldsymbol{g}}^{[2]}\mathbf{S}^{-1}\hat{\boldsymbol{g}}^{[2]}}_{-\mathrm{Mat}(\Gamma^{c_2}_{a_2 b_2}\hat{g}^{a_2}\hat{g}^{b_2})}
\end{aligned}$$

E.1. Proof of Theorem 1

Now, we give a proof of Theorem 1.

Proof: First note that $\hat{\mathbf{G}} = \mathbf{S} - \mathbb{E}_q\big[\nabla_z^2 \bar{\ell}(\mathbf{z})\big]$ is a symmetric matrix. Let $\mathbf{L}$ be the Cholesky of the current $\mathbf{S} = \mathbf{L}\mathbf{L}^T$. We can simplify the right hand side of (9) as follows:

$$(1-t)\mathbf{S} + t\mathbb{E}_q\big[\nabla_z^2 \bar{\ell}(\mathbf{z})\big] + \frac{t^2}{2}\hat{\mathbf{G}}\mathbf{S}^{-1}\hat{\mathbf{G}} = \mathbf{S} - t\hat{\mathbf{G}} + \frac{t^2}{2}\hat{\mathbf{G}}\mathbf{S}^{-1}\hat{\mathbf{G}} = \tfrac{1}{2}\left(\mathbf{S} + \left(\mathbf{L} - t\hat{\mathbf{G}}\mathbf{L}^{-T}\right)\left(\mathbf{L}^T - t\mathbf{L}^{-1}\hat{\mathbf{G}}\right)\right) = \tfrac{1}{2}\Big(\mathbf{S} + \mathbf{U}^T\mathbf{U}\Big),$$

where $\mathbf{U} := \mathbf{L}^T - t\mathbf{L}^{-1}\hat{\mathbf{G}}$. Since the current $\mathbf{S}$ is positive-definite, and $\mathbf{U}^T\mathbf{U}$ is positive semi-definite, we know that the update for $\mathbf{S}$ is positive-definite. □

E.2. Natural Gradients and the Reparameterization Trick

Since $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \mathbf{S}\}$ is a BCN parameterization of a exponential family distribution, gradients w.r.t. BC expectation parameters are natural gradients for BC natural parameters as shown in Theorem 3.

Given that $\mathbf{S}$ is known, the BC expectation parameter is $\mathbf{m}_{[1]} = \mathbb{E}_{q(z)}[\mathbf{S}\mathbf{z}] = \mathbf{S}\boldsymbol{\mu}$. In this case, we know that $\partial_\mu \mathcal{L} = \mathbf{S}\partial_{m_{[1]}}\mathcal{L}$. Therefore, the natural gradient w.r.t. $\boldsymbol{\mu}$ is $\hat{\boldsymbol{g}}^{[1]} = \partial_{m_{[1]}}\mathcal{L} = \mathbf{S}^{-1}\partial_\mu \mathcal{L} = \boldsymbol{\Sigma}\partial_\mu \mathcal{L}$.

Likewise, given that $\boldsymbol{\mu}$ is known, the BC expectation parameter is $\mathbf{m}_{[2]} = \mathbb{E}_{q(z)}\left[-\frac{1}{2}\mathbf{z}\mathbf{z}^T + \boldsymbol{\mu}\mathbf{z}^T\right] = \frac{1}{2}\left(\boldsymbol{\mu}\boldsymbol{\mu}^T - \mathbf{S}^{-1}\right)$. Therefore, the natural gradient w.r.t. $\mathbf{S}$ is $\hat{\boldsymbol{g}}^{[2]} = \partial_{m_{[2]}}\mathcal{L} = -2\partial_{S^{-1}}\mathcal{L} = -2\partial_\Sigma \mathcal{L}$.

Recall that $\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(z|\lambda)}[\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z}) + \log q(\mathbf{z}|\boldsymbol{\lambda})]$, by the Gaussian identities (Opper & Archambeau, 2009; Särkkä, 2013) (see Lin et al. (2019b) for a derivation of these identities), we have

$$\begin{aligned}
\partial_\mu \mathcal{L}(\boldsymbol{\lambda}) &= \partial_\mu \left[\mathbb{E}_{q(z|\lambda)}[\,\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z})] - \tfrac{1}{2}\log|2\pi e\boldsymbol{\Sigma}|\right] \\
&= \partial_\mu \left[\mathbb{E}_{q(z|\lambda)}[\,\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z})]\right] \\
&= \mathbb{E}_{q(z|\lambda)}\left[\nabla_z\left[\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z})\right]\right] \qquad (19) \\
\partial_\Sigma \mathcal{L}(\boldsymbol{\lambda}) &= \partial_\Sigma \left[\mathbb{E}_{q(z|\lambda)}[\,\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z})] - \tfrac{1}{2}\log|2\pi e\boldsymbol{\Sigma}|\right] \\
&= \partial_\Sigma \left[\mathbb{E}_{q(z|\lambda)}[\,\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z})]\right] - \tfrac{1}{2}\boldsymbol{\Sigma}^{-1} \\
&= \tfrac{1}{2}\mathbb{E}_{q(z|\lambda)}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\nabla_z^T\left[\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z})\right]\right] - \tfrac{1}{2}\boldsymbol{\Sigma}^{-1} \qquad (20) \\
&= \tfrac{1}{2}\mathbb{E}_{q(z|\lambda)}\left[\nabla_z^2\left[\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z})\right]\right] - \tfrac{1}{2}\boldsymbol{\Sigma}^{-1} \qquad (21)
\end{aligned}$$

where (19) is also known as the reparameterization trick for the mean, (20) is also known as the reparameterization trick for the covariance, and we call (21) the Hessian trick.

Using Monte Carlo approximation, we have

$$\begin{aligned}
&\partial_\mu \mathcal{L} \approx \nabla_z\left[\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z})\right] \\
&\partial_\Sigma \mathcal{L} \approx \frac{1}{4}\left[\bar{\mathbf{S}} + \bar{\mathbf{S}}^T\right] - \tfrac{1}{2}\boldsymbol{\Sigma}^{-1} && \text{referred to as "-rep"} \\
&\partial_\Sigma \mathcal{L} \approx \tfrac{1}{2}\left[\nabla_z^2\left[\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z})\right]\right] - \tfrac{1}{2}\boldsymbol{\Sigma}^{-1} && \text{referred to as "-hess"}
\end{aligned}$$

where $\bar{\mathbf{S}} := \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\nabla_z^T\left[\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z})\right]$ and $\mathbf{z} \sim q(\mathbf{z}|\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

E.3. Adam-like Update

We consider to solve the following problem, where we use a diagonal Gaussian approximation $q(\mathbf{z}|\boldsymbol{\mu}, \mathbf{s}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \mathbf{s})$ and $\mathbf{s} = \boldsymbol{\sigma}^{-2}$.

$$\min_{\mu, s} \mathcal{L}(\boldsymbol{\mu}, \mathbf{s}) = \mathrm{E}_{q(z|\mu,s)}\left[\left(\sum_{i=1}^N \ell_i(\mathbf{z})\right) - \log \mathcal{N}(\mathbf{z}|\mathbf{0}, \lambda^{-1}\mathbf{I}) + \log q(\mathbf{z}|\boldsymbol{\mu}, \mathbf{s})\right]$$

Note that

$$\begin{aligned}
\partial_\mu \mathcal{L}(\boldsymbol{\mu}, \mathbf{s}) &:= \sum_{i=1}^N \partial_\mu \mathbb{E}_{q(z|\mu,s)}[\ell_i(\mathbf{z})] + \lambda\boldsymbol{\mu} \\
\partial_{\sigma^2} \mathcal{L}(\boldsymbol{\mu}, \mathbf{s}) &:= \sum_{i=1}^N \partial_{\sigma^2} \mathbb{E}_{q(z|\mu,s)}[\ell_i(\mathbf{z})] + \tfrac{1}{2}\lambda - \tfrac{1}{2}\mathbf{s}
\end{aligned}$$

where $\partial_\mu \mathbb{E}_{q(z|\mu,s)}[\ell_i(\mathbf{z})]$ and $\partial_{\sigma^2}\mathbb{E}_{q(z|\mu,s)}[\ell_i(\mathbf{z})]$ can be computed by the reparameterization trick with MC approximations where $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \mathbf{s})$.

$$\begin{aligned}
\partial_\mu \mathbb{E}_{q(z|\mu,s)}[\ell_i(\mathbf{z})] &= \mathbb{E}_{q(z|\mu,s)}[\nabla_z \ell_i(\mathbf{z})] \approx \nabla_z \ell_i(\mathbf{z}) \\
\partial_{\sigma^2}\mathbb{E}_{q(z|\mu,s)}[\ell_i(\mathbf{z})] &= \tfrac{1}{2}\mathbb{E}_{q(z|\mu,s)}[\mathbf{s} \odot (\mathbf{z}-\boldsymbol{\mu}) \odot \nabla_z \ell_i(\mathbf{z})] \approx \tfrac{1}{2}[\mathbf{s} \odot (\mathbf{z}-\boldsymbol{\mu})] \odot \nabla_z \ell_i(\mathbf{z})
\end{aligned}$$

The natural gradients can be computed as follows.

$$\begin{aligned}
\hat{\boldsymbol{g}}_k^{[1]} &= \boldsymbol{\sigma}_k^2\left(\partial_\mu \mathcal{L}(\boldsymbol{\mu}, \mathbf{s})\big|_{\mu=\mu_k, s=s_k}\right) \\
\hat{\boldsymbol{g}}_k^{[2]} &= -2\partial_{\sigma^2}\mathcal{L}(\boldsymbol{\mu}, \mathbf{s})\big|_{\mu=\mu_k, s=s_k}
\end{aligned}$$

The update induced by our rule with exponential decaying step-sizes and the natural momentum (Khan et al., 2018) shown in blue is given as follows.

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - t_1 \hat{\boldsymbol{g}}_k^{[1]} + t_2 \boldsymbol{\sigma}_k^2 \odot \boldsymbol{\sigma}_{k-1}^{-2} \odot \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k-1}\right)$$
$$\boldsymbol{\sigma}_{k+1}^{-2} = \boldsymbol{\sigma}_k^{-2} - t_3 \hat{\boldsymbol{g}}_k^{[2]} + \frac{t_3^2}{2} \hat{\boldsymbol{g}}_k^{[2]} \odot \boldsymbol{\sigma}_k^2 \odot \hat{\boldsymbol{g}}_k^{[2]}$$

where $t_1 = t(1-r_1)\frac{1-r_2^k}{1-r_1^k}$, $t_2 = r_1 \frac{1-r_2^k}{1-r_1^k}\frac{1-r_1^{k-1}}{1-r_2^{k-1}}$, and $t_3 = (1-r_2)$.

Recall that $\mathbf{s} = \boldsymbol{\sigma}^{-2}$. The proposed update can be expressed as

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - t(1-r_1)\frac{1-r_2^k}{1-r_1^k}\hat{\mathbf{s}}_k^{-1} \odot \mathbf{g}_k + r_1 \frac{1-r_2^k}{1-r_1^k}\frac{1-r_1^{k-1}}{1-r_2^{k-1}}\hat{\mathbf{s}}_k^{-1} \odot \hat{\mathbf{s}}_{k-1} \odot \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k-1}\right)$$
$$\hat{\mathbf{s}}_{k+1} = \hat{\mathbf{s}}_k + (1-r_2)\mathbf{h}_k + \frac{(1-r_2)^2}{2}\mathbf{h}_k \odot \hat{\mathbf{s}}_k^{-1} \odot \mathbf{h}_k$$
$$\mathbf{s}_{k+1} = N\hat{\mathbf{s}}_{k+1}$$

where $\mathbf{g}_k := \frac{1}{N}\sum_{i=1}^N \partial_\mu \mathbb{E}_{q(z|\mu,s)}\left[\ell_i(\mathbf{z})\right]\big|_{\mu=\mu_k, s=s_k} + \frac{\lambda}{N}\boldsymbol{\mu}_k$ and $\mathbf{h}_k := \frac{2}{N}\sum_{i=1}^N \partial_{\sigma^2} \mathbb{E}_{q(z|\mu,s)}\left[\ell_i(\mathbf{z})\right]\big|_{\mu=\mu_k, s=s_k} + \frac{\lambda}{N} - \hat{\mathbf{s}}_k$.

Let's define $\mathbf{m}_k := \frac{1-r_1^{k-1}}{t(1-r_2^{k-1})}\hat{\mathbf{s}}_{k-1} \odot \left(\boldsymbol{\mu}_{k-1} - \boldsymbol{\mu}_k\right)$. We can further simplify the above update as shown below.

$$\begin{aligned}
\boldsymbol{\mu}_{k+1} &= \boldsymbol{\mu}_k - t(1-r_1)\frac{1-r_2^k}{1-r_1^k}\hat{\mathbf{s}}_k^{-1} \odot \mathbf{g}_k + t r_1 \frac{1-r_2^k}{1-r_1^k}\hat{\mathbf{s}}_k^{-1} \odot \left(\frac{1-r_1^{k-1}}{t(1-r_2^{k-1})}\hat{\mathbf{s}}_{k-1} \odot \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k-1}\right)\right) \\
&= \boldsymbol{\mu}_k - t\frac{1-r_2^k}{1-r_1^k}\hat{\mathbf{s}}_k^{-1} \odot \left[(1-r_1)\mathbf{g}_k + r_1 \mathbf{m}_k\right] \\
\mathbf{m}_{k+1} &= \frac{1-r_1^k}{t(1-r_2^k)}\hat{\mathbf{s}}_k \odot \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k+1}\right) \\
&= \frac{1-r_1^k}{t(1-r_2^k)} t \frac{1-r_2^k}{1-r_1^k}\left[(1-r_1)\mathbf{g}_k + r_1\mathbf{m}_k\right] \\
&= (1-r_1)\mathbf{g}_k + r_1 \mathbf{m}_k \\
\hat{\mathbf{s}}_{k+1} &= \hat{\mathbf{s}}_k + (1-r_2)\mathbf{h}_k + \frac{(1-r_2)^2}{2}\mathbf{h}_k \odot \hat{\mathbf{s}}_k^{-1} \odot \mathbf{h}_k \\
&= \tfrac{1}{2}\left[\hat{\mathbf{s}}_k + (\hat{\mathbf{s}}_k + (1-r_2)\mathbf{h}_k) \odot \hat{\mathbf{s}}_k^{-1} \odot (\hat{\mathbf{s}}_k + (1-r_2)\mathbf{h}_k)\right] \\
\mathbf{s}_{k+1} &= N\hat{\mathbf{s}}_{k+1}
\end{aligned}$$

where $\mathbf{z} \sim q(\mathbf{z}|\boldsymbol{\mu}_k, \mathbf{s}_k)$, $\mathbf{g}_k \approx \nabla_z \ell_i(\mathbf{z}) + \frac{\lambda}{N}\boldsymbol{\mu}_k$, and $\mathbf{h}_k \approx \left[(N\hat{\mathbf{s}}_k) \odot (\mathbf{z} - \boldsymbol{\mu})\right] \odot \nabla_z \ell_i(\mathbf{z}) + \frac{\lambda}{N} - \hat{\mathbf{s}}_k$.

E.4. Tran et al. (2019) is a special case of our update

In the Gaussian case, Tran et al. (2019) consider the following update by using parameterization $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, where $\boldsymbol{\Sigma}$ is the covariance matrix.

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - t\boldsymbol{\Sigma}(\partial_\mu \mathcal{L}) \tag{22}$$
$$\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma} - t\hat{\boldsymbol{g}}^{[2]} + \frac{t \times t}{2}\hat{\boldsymbol{g}}^{[2]}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{g}}^{[2]} = Ret(\boldsymbol{\Sigma}, -t\hat{\boldsymbol{g}}^{[2]}). \tag{23}$$

where the natural gradient[19] for $\boldsymbol{\Sigma}$ is $\hat{\boldsymbol{g}}^{[2]} := 2\boldsymbol{\Sigma}(\partial_\Sigma \mathcal{L})\boldsymbol{\Sigma}$ and the retraction map is $Ret(\boldsymbol{\Sigma}, \mathbf{b}) := \boldsymbol{\Sigma} + \mathbf{b} + \frac{1}{2}\mathbf{b}\boldsymbol{\Sigma}^{-1}\mathbf{b}$.

However, Tran et al. (2019) do not justify the use of the retraction map, which is just one of retraction maps developed for positive definite matrices. In this section, we show that how to derive this update from our rule.

[19] There is a typo in Algorithm 2 of Tran et al. (2019). The natural gradient for $\boldsymbol{\Sigma}$ should be $2\boldsymbol{\Sigma}(\partial_\Sigma \mathcal{L})\boldsymbol{\Sigma}$ instead of $\boldsymbol{\Sigma}(\partial_\Sigma \mathcal{L})\boldsymbol{\Sigma}$.

As shown in Eq. (16), our rule can be used under not only a BCN parameterization but also a BC parameterization. Now, we show that our rule can recover the above update using the parameterization $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. Recall that this parameterization is a BC parameterization. It only requires us to show that natural gradients and the additional terms are described in Eq. (23).

Given that $\boldsymbol{\Sigma}$ is known, $\boldsymbol{\mu}$ is the natural parameter and the expectation parameter is $\mathbf{m}_{[1]} = \mathbb{E}_{q(z)}\left[\boldsymbol{\Sigma}^{-1}\mathbf{z}\right] = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ as shown in Appendix E.2. Therefore, the natural gradient w.r.t. $\boldsymbol{\mu}$ is $\hat{\boldsymbol{g}}^{[1]} = \partial_{m_{[1]}}\mathcal{L} = \boldsymbol{\Sigma}\partial_{\mu}\mathcal{L}$.

Now, we show that the natural gradients w.r.t. $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{g}}^{[2]} = 2\boldsymbol{\Sigma}(\partial_{\Sigma}\mathcal{L})\boldsymbol{\Sigma}$$

A proof using matrix calculus is provided below. See Malagò & Pistone (2015) for alternative proofs. By matrix calculus, we have

$$\begin{aligned}
&-\mathbb{E}_{q(z)}\left[\partial_{\Sigma^{ij}}\partial_{\Sigma}\left[\log q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\right]\right] \\
=&\mathbb{E}_{q(z)}\left[\partial_{\Sigma^{ij}}\partial_{\Sigma}\left[\tfrac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}/(2\pi)|\right]\right] \\
=&\tfrac{1}{2}\mathbb{E}_{q(z)}\left[\partial_{\Sigma^{ij}}\left[-\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\right]\right] \\
=&\tfrac{1}{2}\mathbb{E}_{q(z)}\left[-\partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right](\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^T\partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right] + \partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right]\right] \\
=&\tfrac{1}{2}\mathbb{E}_{q(z)}\left[-\partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right](\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^T\partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right] + \partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right]\right] \\
=&-\tfrac{1}{2}\partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right]\underbrace{\mathbb{E}_{q(z)}\left[(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^T\right]}_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\underbrace{\mathbb{E}_{q(z)}\left[(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^T\right]}_{\boldsymbol{\Sigma}}\partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right] + \tfrac{1}{2}\partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right] \\
=&\tfrac{1}{2}\left[-\partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right]\mathbf{I} - \mathbf{I}\partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right] + \partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right]\right] \\
=&-\tfrac{1}{2}\partial_{\Sigma^{ij}}\left[\boldsymbol{\Sigma}^{-1}\right]
\end{aligned}$$

Therefore, the block matrix of the FIM related to $\boldsymbol{\Sigma}$ is $\mathbf{F}_{\Sigma} := -\mathbb{E}_{q(z)}\left[\partial^2_{\mathrm{vec}(\Sigma)}\left[\log q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\right]\right] = -\frac{1}{2}\partial_{\mathrm{vec}(\Sigma)}\left[\mathrm{vec}(\boldsymbol{\Sigma}^{-1})\right]$ due to the above expression. Note that $\mathbf{F}_{\Sigma}^{-1} = -2\partial_{\mathrm{vec}(\Sigma^{-1})}\left[\mathrm{vec}(\boldsymbol{\Sigma})\right]$.

Note that $\hat{\boldsymbol{g}}^{[2]}$ is the natural gradient for $\boldsymbol{\Sigma}$. Since $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ is a BC parameterization, by Lemma 1, the natural gradient w.r.t. $\mathrm{vec}(\boldsymbol{\Sigma})$ is

$$\begin{aligned}
\mathrm{vec}(\hat{\boldsymbol{g}}^{[2]}) &:= \mathbf{F}_{\Sigma}^{-1}\mathrm{vec}(\partial_{\Sigma}\mathcal{L}) \\
&= -2\partial_{\mathrm{vec}(\Sigma^{-1})}\left[\mathrm{vec}(\boldsymbol{\Sigma})\right]\mathrm{vec}(\partial_{\Sigma}\mathcal{L}) \\
&= -2\partial_{\mathrm{vec}(\Sigma^{-1})}\left[\mathrm{vec}(\boldsymbol{\Sigma})\right]\partial_{\mathrm{vec}(\Sigma)}\mathcal{L} \\
&= -2\partial_{\mathrm{vec}(\Sigma^{-1})}\mathcal{L} \\
&= -2\mathrm{vec}(\partial_{\Sigma^{-1}}\mathcal{L})
\end{aligned}$$

where we obtain the fourth step using the chain rule.

Therefore, we have $\hat{\boldsymbol{g}}^{[2]} = -2\partial_{\Sigma^{-1}}\mathcal{L}$. By matrix calculus, we have

$$\partial_{\Sigma^{-1}}\mathcal{L} = -\boldsymbol{\Sigma}(\partial_{\Sigma}\mathcal{L})\boldsymbol{\Sigma}$$

Finally, we have

$$\hat{\boldsymbol{g}}^{[2]} = 2\boldsymbol{\Sigma}(\partial_{\Sigma}\mathcal{L})\boldsymbol{\Sigma}$$

Now, we show that the additional term for $\boldsymbol{\mu}$ is $\mathbf{0}$ under parameterization $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. Since $\boldsymbol{\lambda}$ is a BC parameterization, by Lemma 3, all entries of $\Gamma^{a_1}_{b_1c_1}$ for $\boldsymbol{\mu}$ are zeros. Therefore, the additional term for $\boldsymbol{\mu}$ is $\mathbf{0}$.

We denote the Christoffel symbol of the second kind for $\mathrm{vec}(\boldsymbol{\Sigma})$ as $\Gamma^{a_2}{}_{b_2c_2}$. Now, we show that the additional term for $\boldsymbol{\Sigma}$ is $\frac{t\times t}{2}\hat{\boldsymbol{g}}^{[2]}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{g}}^{[2]}$. It is equivalent to show $\mathrm{Mat}(\Gamma^{a_2}_{b_2c_2}\hat{g}^{b_2}\hat{g}^{c_2}) = -\hat{\boldsymbol{g}}^{[2]}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{g}}^{[2]}$.

Recall that the natural gradient for $\mathbf{S} = \mathbf{\Sigma}^{-1}$ is $\mathbf{G} = -2\partial_{\Sigma}\mathcal{L}$. Under parameterization $\bar{\boldsymbol{\lambda}} = \{\boldsymbol{\mu}, \mathbf{S}\}$, $\bar{\mathbf{R}}^{[2]}(t)$ has the following closed-form expression, which is used in the proof of Lemma 4.

$$\mathrm{Mat}(\bar{\mathbf{R}}^{[2]}(t)) = \mathbf{U}\mathrm{Exp}(t\mathbf{U}^{-1}\mathbf{G}\mathbf{U}^{-1})\mathbf{U}$$

where $\mathbf{U} = \mathbf{S}^{\frac{1}{2}}$ and $\mathrm{Exp}(\mathbf{X}) := \mathbf{I} + \sum_{n=1}^{\infty}\frac{\mathbf{X}^n}{n!}$.

Note that $\mathbf{\Sigma} = \mathbf{S}^{-1}$. Therefore, under parameterization $\boldsymbol{\lambda} = \{\boldsymbol{\mu}, \mathbf{\Sigma}\}$, we have

$$\begin{aligned}
\overbrace{\mathrm{Mat}(\mathbf{R}^{[2]}(t))}^{\mathbf{\Sigma}_{\text{new}}} &= \overbrace{\left[\ \mathrm{Mat}(\bar{\mathbf{R}}^{[2]}(t))\ \right]^{-1}}^{\mathbf{S}_{\text{new}}} \\
&= (\mathbf{U}\mathrm{Exp}(t\mathbf{U}^{-1}\mathbf{G}\mathbf{U}^{-1})\mathbf{U})^{-1} \\
&= \mathbf{U}^{-1}\mathrm{Exp}(-t\mathbf{U}^{-1}\mathbf{G}\mathbf{U}^{-1})\mathbf{U}^{-1} \\
&= \mathbf{\Sigma}^{1/2}\mathrm{Exp}(-t\mathbf{\Sigma}^{1/2}\mathbf{G}\mathbf{\Sigma}^{1/2})\mathbf{\Sigma}^{1/2} \\
&= \mathbf{\Sigma}^{1/2}\mathrm{Exp}(t\mathbf{\Sigma}^{1/2}(2\partial_{\Sigma}\mathcal{L})\mathbf{\Sigma}^{1/2})\mathbf{\Sigma}^{1/2} \\
&= \mathbf{\Sigma}^{1/2}\mathrm{Exp}(t\mathbf{\Sigma}^{-1/2}\underbrace{[2\mathbf{\Sigma}(\partial_{\Sigma}\mathcal{L})\mathbf{\Sigma}]}_{\hat{\boldsymbol{g}}^{[2]}}\mathbf{\Sigma}^{-1/2})\mathbf{\Sigma}^{1/2} \\
&= \mathbf{\Sigma}^{1/2}\mathrm{Exp}(t\mathbf{\Sigma}^{-1/2}\hat{\boldsymbol{g}}^{[2]}\mathbf{\Sigma}^{-1/2})\mathbf{\Sigma}^{1/2},
\end{aligned}$$

where we use the identity $(\mathrm{Exp}(t\mathbf{U}^{-1}\mathbf{G}\mathbf{U}^{-1}))^{-1} = \mathrm{Exp}(-t\mathbf{U}^{-1}\mathbf{G}\mathbf{U}^{-1})$.

Note that a geodesic is invariant under parameterization. Alternatively, we can obtain the above equation by using the fact that $\mathbf{R}^{[2]}(t)$ is a geodesic of Gaussian distribution with a constant mean.

Using a similar proof as shown in Lemma 4, the additional term for $\mathbf{\Sigma}$ is

$$\mathrm{Mat}(\Gamma^{a_2}_{b_2c_2}\hat{g}^{b_2}\hat{g}^{c_2}) = -\hat{\boldsymbol{g}}^{[2]}\mathbf{\Sigma}^{-1}\hat{\boldsymbol{g}}^{[2]}$$

where $\Gamma^{a_2}_{b_2c_2}$ is the Christoffel symbol of the second kind for $\mathrm{vec}(\mathbf{\Sigma})$ and $\hat{g}^{a_2}$ denotes the a-th element of $\mathrm{vec}(\hat{\boldsymbol{g}}^{[2]})$.

F. Example: Gamma Approximation

We consider the gamma distribution under the parameterization $\boldsymbol{\lambda} = \{\lambda^{[1]}, \lambda^{[2]}\}$, where $\lambda^{[1]} = \alpha$ and $\lambda^{[2]} = \frac{\beta}{\alpha}$.

Since every block contains only a scalar, we use global indexes such as $\lambda^{(i)} = \lambda^{[i]}$, $\lambda^{(i)} = \lambda^{a_i}$ and $\Gamma_{i,ii} = \Gamma_{a_i,b_ic_i}$ for notation simplicity. The open-set constraint is $\Omega_1 = \mathbb{S}^1_{++}$ and $\Omega_2 = \mathbb{S}^1_{++}$. Under this parameterization, we can express the distribution as below.

$$q(z|\boldsymbol{\lambda}) = z^{-1}\exp\left(\lambda^{(1)}\log z - z\lambda^{(1)}\lambda^{(2)} - A(\boldsymbol{\lambda})\right)$$

where $A(\boldsymbol{\lambda}) = \log\mathrm{Ga}(\lambda^{(1)}) - \lambda^{(1)}\left(\log\lambda^{(1)} + \log\lambda^{(2)}\right)$ and $\mathrm{Ga}(\cdot)$ is the gamma function.

Lemma 5 *The Fisher information matrix is diagonal under this parameterization. It implies that this parameterization is a BC parameterization.*

Proof: Notice that $\mathbb{E}_{q(z|\lambda)}[z] = \frac{1}{\lambda^{(2)}}$. The Fisher information matrix is diagonal as shown below.

$$\begin{aligned}
\mathbf{F} &= -\mathbb{E}_{q(z|\lambda)}\left[\partial^2_{\lambda}\log q(z|\boldsymbol{\lambda})\right] \\
&= -\mathbb{E}_{q(z|\lambda)}\begin{bmatrix}-\partial^2_{\lambda^{(1)}}A(\boldsymbol{\lambda}) & \left(-z + \frac{1}{\lambda^{(2)}}\right) \\ \left(-z + \frac{1}{\lambda^{(2)}}\right) & -\partial^2_{\lambda^{(2)}}A(\boldsymbol{\lambda})\end{bmatrix} \\
&= \mathbb{E}_{q(z|\lambda)}\begin{bmatrix}\partial^2_{\lambda^{(1)}}A(\boldsymbol{\lambda}) & 0 \\ 0 & \partial^2_{\lambda^{(2)}}A(\boldsymbol{\lambda})\end{bmatrix} \\
&= \begin{bmatrix}\partial_{\lambda^{(1)}}\psi(\lambda^{(1)}) - \frac{1}{\lambda^{(1)}} & 0 \\ 0 & \frac{\lambda^{(1)}}{\left(\lambda^{(2)}\right)^2}\end{bmatrix}
\end{aligned}$$

where $\psi(\cdot)$ denotes the digamma function. □

Lemma 6 *$\boldsymbol{\lambda}$ is a BCN parameterization.*

Proof: By Lemma 5, we know that $\boldsymbol{\lambda}$ is a BC parameterization. Now, we show that $\boldsymbol{\lambda} = \{\lambda^{(1)}, \lambda^{(2)}\}$ is a BCN parameterization. Clearly, each $\lambda^{(i)} \in \mathbb{S}^1_{++}$ has all degrees of freedom.

The gamma distribution which can be written as following exponential form:

$$q(z|\lambda^{(1)}, \lambda^{(2)}) = z^{-1} \exp\left(\lambda^{(1)} \log z - z\lambda^{(1)}\lambda^{(2)} - A(\boldsymbol{\lambda})\right)$$

Considering two blocks with $\lambda^{(1)}$ and $\lambda^{(2)}$ respectively, we can express this distribution in the following two ways where the first equation is for the $\lambda^{(1)}$ block while the second equation is for the $\lambda^{(2)}$ block:

$$\begin{aligned} q(z|\lambda^{(1)}, \lambda^{(2)}) &= \underbrace{z^{-1}}_{h_1(z,\lambda^{(2)})} \exp\left(\langle \underbrace{\log z - z\lambda^{(2)}}_{\phi_1(z,\lambda^{(2)})}, \lambda^{(1)}\rangle - A(\boldsymbol{\lambda})\right) \\ &= \underbrace{z^{-1}\exp(\lambda^{(1)}\log z)}_{h_2(z,\lambda^{(1)})} \exp\left(\langle \underbrace{-z\lambda^{(1)}}_{\phi_2(z,\lambda^{(1)})}, \lambda^{(2)}\rangle - A(\boldsymbol{\lambda})\right) \end{aligned}$$

Therefore, by the definition of BCN, we know that $\boldsymbol{\lambda}$ is a BCN parameterization.

□

Using this BCN parameterization, the Christoffel symbols can be readily computed as below.

$$\Gamma_{1,11} = \tfrac{1}{2}\partial^3_{\lambda^{(1)}} A(\boldsymbol{\lambda}) = \tfrac{1}{2}\left(\partial^2_{\lambda^{(1)}}\psi(\lambda^{(1)}) + \frac{1}{\left(\lambda^{(1)}\right)^2}\right), \quad \Gamma_{2,22} = \tfrac{1}{2}\partial^3_{\lambda^{(2)}} A(\boldsymbol{\lambda}) = -\frac{\lambda^{(1)}}{\left(\lambda^{(2)}\right)^3}$$

$$\Gamma^1{}_{11} = \frac{\Gamma_{1,11}}{F_{11}} = \frac{\partial^2_{\lambda^{(1)}}\psi(\lambda^{(1)}) + \frac{1}{\left(\lambda^{(1)}\right)^2}}{2\left(\partial_{\lambda^{(1)}}\psi(\lambda^{(1)}) - \frac{1}{\lambda^{(1)}}\right)}, \quad \Gamma^2{}_{22} = \frac{\Gamma_{2,22}}{F_{22}} = -\frac{1}{\lambda^{(2)}}$$

F.1. Proof of Theorem 2

We first prove the following lemma.

Lemma 7 $\Gamma^1{}_{11} < -\frac{1}{\lambda^{(1)}}$ *when* $\lambda^{(1)} > 0$.

Proof: By Eq 1.4 at Batir (2005) and the last inequality at page 13 of Koumandos (2008), we have the following inequalities when $\lambda^{(1)} > 0$.

$$\partial_{\lambda^{(1)}}\psi(\lambda^{(1)}) - \frac{1}{\lambda^{(1)}} > \frac{1}{2\left(\lambda^{(1)}\right)^2} > 0 \qquad \text{Batir (2005)} \tag{24}$$

$$\partial^2_{\lambda^{(1)}}\psi(\lambda^{(1)}) < \frac{1}{\left(\lambda^{(1)}\right)^2} - \frac{2\partial_{\lambda^{(1)}}\psi(\lambda^{(1)})}{\lambda^{(1)}} \qquad \text{Koumandos (2008)} \tag{25}$$

By (25), we have

$$\partial^2_{\lambda^{(1)}}\psi(\lambda^{(1)}) + \frac{1}{\left(\lambda^{(1)}\right)^2} < \frac{2}{\left(\lambda^{(1)}\right)^2} - \frac{2\partial_{\lambda^{(1)}}\psi(\lambda^{(1)})}{\lambda^{(1)}} = \frac{2}{\lambda^{(1)}}\left(\frac{1}{\lambda^{(1)}} - \partial_{\lambda^{(1)}}\psi(\lambda^{(1)})\right)$$

Since $\partial_{\lambda^{(1)}}\psi(\lambda^{(1)}) - \frac{1}{\lambda^{(1)}} > 0$, we have

$$2\Gamma^1{}_{11} = \frac{\partial^2_{\lambda^{(1)}}\psi(\lambda^{(1)}) + \frac{1}{\left(\lambda^{(1)}\right)^2}}{\partial_{\lambda^{(1)}}\psi(\lambda^{(1)}) - \frac{1}{\lambda^{(1)}}} < -\frac{2}{\lambda^{(1)}}$$

which shows $\Gamma^1{}_{11} < -\frac{1}{\lambda^{(1)}}$. □

Now, We give a proof for Theorem 2.

Proof: The proposed update for $\lambda^{(1)}$ with step-size t is given below.

$$\begin{aligned}
\lambda^{(1)} &\leftarrow \lambda^{(1)} - t\hat{g}^{(1)} - \frac{t^2}{2}\left(\Gamma^1{}_{11}\right)\left(\hat{g}^{(1)}\right)^2 \\
&> \lambda^{(1)} - t\hat{g}^{(1)} + \frac{t^2}{2}\left(\frac{1}{\lambda^{(1)}}\right)\left(\hat{g}^{(1)}\right)^2 \\
&= \frac{1}{2\lambda^{(1)}}\left[2\left(\lambda^{(1)}\right)^2 - 2t\hat{g}^{(1)}\lambda^{(1)} + \left(t\hat{g}^{(1)}\right)^2\right] \\
&= \underbrace{\frac{1}{2\lambda^{(1)}}}_{>0}\left[\underbrace{\left(\lambda^{(1)}\right)^2}_{>0} + \underbrace{\left(\lambda^{(1)} - t\hat{g}^{(1)}\right)^2}_{\geq 0}\right]
\end{aligned}$$

where in the second step we use the inequality $\Gamma^1{}_{11} < -\frac{1}{\lambda^{(1)}}$ shown in Lemma 7 since the current/old $\lambda^{(1)} > 0$.

Similarly, we can show the update for $\lambda^{(2)}$ also satisfies the constraint.

$$\begin{aligned}
\lambda^{(2)} &\leftarrow \lambda^{(2)} - t\hat{g}^{(2)} + \frac{t^2}{2}\left(\frac{1}{\lambda^{(2)}}\right)\left(\hat{g}^{(2)}\right)^2 \\
&= \underbrace{\frac{1}{2\lambda^{(2)}}}_{>0}\left[\underbrace{\left(\lambda^{(2)}\right)^2}_{>0} + \underbrace{\left(\lambda^{(2)} - t\hat{g}^{(2)}\right)^2}_{\geq 0}\right]
\end{aligned}$$

It is obvious to see that the proposed update satisfies the underlying constraint. □

F.2. Natural Gradients

Recall that $\hat{g}$ are the natural-gradients, which can be computed as shown below.

$$\hat{g}^{(1)} = \frac{\partial_{\lambda^{(1)}}\mathcal{L}}{\partial_{\lambda^{(1)}}\psi(\lambda^{(1)}) - \frac{1}{\lambda^{(1)}}},\ \hat{g}^{(2)} = \frac{\left(\lambda^{(2)}\right)^2}{\lambda^{(1)}}\partial_{\lambda^{(2)}}\mathcal{L}$$

Recall that $\lambda^{(1)} = \alpha$ and $\lambda^{(2)} = \frac{\beta}{\alpha}$. Using the chain rule, we know that

$$\partial_{\lambda^{(1)}}\mathcal{L} = \partial_\alpha\mathcal{L} + \frac{\beta}{\alpha}\partial_\beta\mathcal{L},\ \ \partial_{\lambda^{(2)}}\mathcal{L} = \alpha\partial_\beta\mathcal{L}$$

$\partial_\alpha\mathcal{L}$ and $\partial_\beta\mathcal{L}$ can be computed by the implicit reparameterization trick (Salimans & Knowles, 2013; Figurnov et al., 2018).

G. Example: Exponential Approximation

In this case, there is only one block with a scalar. We use global indexes such as $\lambda^{(1)} = \lambda^{[1]}$ and $\Gamma_{1,11} = \Gamma_{a_1,b_1c_1}$ for notation simplicity. We consider an exponential distribution under the natural parameterization $\lambda = \lambda^{(1)}$ with the open-set constraint $\Omega = \mathbb{S}^1_{++}$:

$$q(z|\lambda) = \exp\left(-\lambda^{(1)}z - A(\lambda)\right)$$

where $A(\lambda) = -\log\lambda^{(1)}$. The FIM is a scalar $F_{11} = \frac{1}{\left(\lambda^{(1)}\right)^2}$. It is obvious that λ is a BCN parameterization. the Christoffel symbols can be readily computed as below.

$$\Gamma_{1,11} = \tfrac{1}{2}\partial^3_{\lambda^{(1)}}A(\lambda) = -\frac{1}{\left(\lambda^{(1)}\right)^3},\quad \Gamma^1{}_{11} = \frac{\Gamma_{1,11}}{F_{11}} = -\frac{1}{\lambda^{(1)}}$$

The proposed natural-gradient update with step-size t is

$$\lambda^{(1)} = \lambda^{(1)} - t\hat{g}^{(1)} + \frac{t^2}{2}\left(\frac{1}{\lambda^{(1)}}\right)\left(\hat{g}^{(1)}\right)^2$$

where $\hat{g}^{(1)}$ is the natural-gradient. Note that $\hat{g}^{(1)}$ is the natural-gradient, which can be computed as shown below.

$$\hat{g}^{(1)} = \left(\lambda^{(1)}\right)^2 \partial_{\lambda^{(1)}}\mathcal{L}.$$

where $\partial_{\lambda^{(1)}}\mathcal{L}$ can be computed by the implicit reparameterization trick as $\partial_{\lambda^{(1)}}\mathcal{L} \approx [\partial_\lambda z]\,[\partial_z b(z)]$. where $z \sim q(z|\lambda^{(1)})$ and $b(z) := \bar{\ell}(z) + \log q(z|\lambda^{(1)})$

Lemma 8 *The proposed update satisfies the underlying constraint.*

Proof: The proposed natural-gradient update with step-size t is given below.

$$\begin{aligned}
\lambda^{(1)} &\leftarrow \lambda^{(1)} - t\hat{g}^{(1)} + \frac{t^2}{2}\left(\frac{1}{\lambda^{(1)}}\right)\left(\hat{g}^{(1)}\right)^2 \\
&= \frac{1}{2\lambda^{(1)}}\left[2\left(\lambda^{(1)}\right)^2 - 2t\hat{g}^{(1)}\lambda^{(1)} + \left(t\hat{g}^{(1)}\right)^2\right] \\
&= \frac{1}{2\lambda^{(1)}}\left[\left(\lambda^{(1)}\right)^2 + \left(\lambda^{(1)} - t\hat{g}^{(1)}\right)^2\right]
\end{aligned}$$

It is obvious to see that the proposed update satisfies the underlying constraint. □

G.1. Implicit reparameterization gradient

Now, we discuss how to compute the gradients w.r.t. λ using the implicit reparameterization trick. To use the implicit reparameterization trick, we have to compute the following term.

$$\partial_\lambda z = -\frac{\partial_\lambda Q(z|\lambda)}{q(z|\lambda)} = -\frac{\partial_\lambda\left(1 - \exp(-\lambda z)\right)}{\lambda\exp(-\lambda z)} = -\frac{z\exp(-\lambda z)}{\lambda\exp(-\lambda z)} = -\frac{z}{\lambda}$$

where $Q(z|\lambda)$ is the C.D.F. of $q(z|\lambda)$.

H. Example: Inverse Gaussian Approximation

We consider the following distribution.

$$q(z|\alpha,\beta) = \sqrt{\frac{1}{2\pi z^3}}\exp\left(-\frac{z\alpha\beta^2}{2} - \frac{\alpha}{2z} + \frac{\log\alpha}{2} + \alpha\beta\right)$$

where $\{\frac{1}{\beta}, \alpha\}$ is a BC parameterization.

We consider a BCN parameterization $\boldsymbol{\lambda} = \{\lambda^{[1]}, \lambda^{[2]}\}$, where $\lambda^{[1]} = \beta^2$ and $\lambda^{[2]} = \alpha$ and the open-set constraint is $\Omega_1 = \mathbb{S}^1_{++}$ and $\Omega_2 = \mathbb{S}^1_{++}$. Since every block contains only a scalar, we use global indexes such as $\lambda^{(i)} = \lambda^{[i]}$ and $\Gamma_{i,ii} = \Gamma_{a_i,b_ic_i}$ for notation simplicity. Under this parameterization, we can re-express the distribution as

$$q(z|\boldsymbol{\lambda}) = \sqrt{\frac{1}{2\pi z^3}}\exp\left(-\frac{z}{2}\lambda^{(1)}\lambda^{(2)} - \frac{\lambda^{(2)}}{2z} - A(\boldsymbol{\lambda})\right)$$

where $A(\boldsymbol{\lambda}) = -\frac{\log\lambda^{(2)}}{2} - \lambda^{(2)}\sqrt{\lambda^{(1)}}$.

Lemma 9 *The FIM is (block) diagonal under this parameterization.*

Proof: Notice that $\mathbb{E}_{q(z|\lambda)}[z] = \frac{1}{\sqrt{\lambda^{(1)}}}$. The FIM is (block) diagonal as shown below.

$$\begin{aligned}
\mathbf{F} &= -\mathbb{E}_{q(z|\lambda)}\left[\partial^2_{\boldsymbol{\lambda}} \log q(z|\boldsymbol{\lambda})\right] \\
&= -\mathbb{E}_{q(z|\lambda)}\begin{bmatrix} -\partial^2_{\lambda^{(1)}} A(\boldsymbol{\lambda}) & \frac{1}{2}\left(-z + \frac{1}{\sqrt{\lambda^{(1)}}}\right) \\ \frac{1}{2}\left(-z + \frac{1}{\sqrt{\lambda^{(1)}}}\right) & -\partial^2_{\lambda^{(2)}} A(\boldsymbol{\lambda}) \end{bmatrix} \\
&= \mathbb{E}_{q(z|\lambda)}\begin{bmatrix} \partial^2_{\lambda^{(1)}} A(\boldsymbol{\lambda}) & 0 \\ 0 & \partial^2_{\lambda^{(2)}} A(\boldsymbol{\lambda}) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{4}\left(\lambda^{(1)}\right)^{-3/2}\lambda^{(2)} & 0 \\ 0 & \frac{1}{2}\left(\lambda^{(2)}\right)^{-2} \end{bmatrix}
\end{aligned}$$

□

It is easy to show that $\boldsymbol{\lambda}$ is a BCN parameterization since $\boldsymbol{\lambda}$ satisfies Assumption 1 to 3.

Due to the BCN parameterization, the Christoffel symbols can be readily computed as below.

$$\Gamma_{1,11} = \tfrac{1}{2}\partial^3_{\lambda^{(1)}} A(\boldsymbol{\lambda}) = -\frac{3}{16}\left(\lambda^{(1)}\right)^{-5/2}\lambda^{(2)}, \quad \Gamma_{2,22} = \tfrac{1}{2}\partial^3_{\lambda^{(2)}} A(\boldsymbol{\lambda}) = -\frac{1}{2}\left(\lambda^{(2)}\right)^{-3}$$
$$\Gamma^1{}_{11} = \frac{\Gamma_{1,11}}{F_{11}} = -\frac{3}{4\lambda^{(1)}}, \quad \Gamma^2{}_{22} = \frac{\Gamma_{2,22}}{F_{22}} = -\frac{1}{\lambda^{(2)}}$$

The proposed natural-gradient update with step-size t is

$$\lambda^{(1)} \leftarrow \lambda^{(1)} - t\hat{g}^{(1)} + \frac{t^2}{2}\left(\frac{3}{4\lambda^{(1)}}\right)\left(\hat{g}^{(1)}\right)^2$$
$$\lambda^{(2)} \leftarrow \lambda^{(2)} - t\hat{g}^{(2)} + \frac{t^2}{2}\left(\frac{1}{\lambda^{(2)}}\right)\left(\hat{g}^{(2)}\right)^2$$

Lemma 10 *The update above satisfies the underlying constraint.*

Proof: The proposed natural-gradient update with step-size t is given below.

$$\begin{aligned}
\lambda^{(1)} &\leftarrow \lambda^{(1)} - t\hat{g}^{(1)} + \frac{t^2}{2}\left(\frac{3}{4\lambda^{(1)}}\right)\left(\hat{g}^{(1)}\right)^2 \\
&= \frac{1}{4\lambda^{(1)}}\left[4\left(\lambda^{(1)}\right)^2 - 4t\hat{g}^{(1)}\lambda^{(1)} + \frac{3}{2}\left(t\hat{g}^{(1)}\right)^2\right] \\
&= \frac{1}{4\lambda^{(1)}}\left[\underbrace{\left(2\lambda^{(1)} - t\hat{g}^{(1)}\right)^2}_{\text{Term I}} + \underbrace{\frac{1}{2}\left(t\hat{g}^{(1)}\right)^2}_{\text{Term II}}\right] \\
\lambda^{(2)} &\leftarrow \lambda^{(2)} - t\hat{g}^{(2)} + \frac{t^2}{2}\left(\frac{1}{\lambda^{(2)}}\right)\left(\hat{g}^{(2)}\right)^2 \\
&= \frac{1}{2\lambda^{(2)}}\left[2\left(\lambda^{(2)}\right)^2 - 2t\hat{g}^{(2)}\lambda^{(2)} + \left(t\hat{g}^{(2)}\right)^2\right] \\
&= \frac{1}{2\lambda^{(2)}}\left[\left(\lambda^{(2)}\right)^2 + \left(\lambda^{(2)} - t\hat{g}^{(2)}\right)^2\right]
\end{aligned}$$

Note that Term I and Term II cannot be both zero at the same time when $\lambda^{(1)} > 0$. A similar argument can be made for the update about $\lambda^{(2)}$. Therefore, the proposed update satisfies the underlying constraint. □

Recall that $\hat{g}$ are the natural-gradients, which can be computed as shown below.

$$\hat{g}^{(1)} = \frac{4}{\lambda^{(2)}}\left(\lambda^{(1)}\right)^{3/2}\partial_{\lambda^{(1)}}\mathcal{L}, \;\; \hat{g}^{(2)} = 2\left(\lambda^{(2)}\right)^2\partial_{\lambda^{(2)}}\mathcal{L}$$

Using the chain rule, we know that

$$\partial_{\lambda^{(1)}}\mathcal{L} = \frac{1}{2\beta}\partial_\beta\mathcal{L}, \;\; \partial_{\lambda^{(2)}}\mathcal{L} = \partial_\alpha\mathcal{L}$$

$\partial_\alpha\mathcal{L}$ and $\partial_\beta\mathcal{L}$ can be computed by the implicit reparameterization trick (Salimans & Knowles, 2013; Figurnov et al., 2018) as $\partial_\eta\mathcal{L} \approx [\partial_\eta z]\,[\nabla_z b(z)]$, where $\eta = \{\alpha, \beta\}$, $z \sim q(z|\alpha,\beta)$ and $b(z) := \bar{\ell}(z) + \log q(z|\alpha,\beta)$

H.1. Implicit reparameterization gradient

Now, we discuss how to compute the gradients w.r.t. α and β using the implicit reparameterization trick. To use the implicit reparameterization trick, we have to compute the following term.

$$\begin{aligned}\partial_\eta z &= -\frac{\partial_\eta Q(z|\boldsymbol{\eta})}{q(z|\boldsymbol{\eta})} \\ &= -\frac{\partial_\eta\left[\Phi(\sqrt{\frac{\alpha}{z}}\,(z\beta-1)) + \exp(2\alpha\beta)\Phi(-\sqrt{\frac{\alpha}{z}}\,(z\beta+1))\right]}{\sqrt{\frac{1}{2\pi z^3}}\exp\left(-\frac{z\alpha\beta^2}{2} - \frac{\alpha}{2z} + \frac{\log\alpha}{2} + \alpha\beta\right)}\end{aligned}$$

where $\boldsymbol{\eta} = \{\alpha, \beta\}$, $Q(z|\boldsymbol{\eta})$ is the C.D.F. of the inverse Gaussian distribution, and $\Phi(x) = \int_{-\infty}^{x}\mathcal{N}(t|0,1)dt$ is the C.D.F. of the standard Gaussian distribution. We use the following fact to simplify the above expression.

$$\delta(z,\alpha,\beta) := \frac{\exp(2\alpha\beta)\Phi(-\sqrt{\frac{\alpha}{z}}\,(z\beta+1))}{\mathcal{N}(\sqrt{\frac{\alpha}{z}}\,(z\beta-1)\,|0,1)} = \frac{\Phi(-\sqrt{\frac{\alpha}{z}}\,(z\beta+1))}{\mathcal{N}(-\sqrt{\frac{\alpha}{z}}\,(z\beta+1)\,|0,1)}$$

where $\delta(z,\alpha,\beta)$ is known as the Mills ratio of Gaussian distribution. Using this fact, we can get the simplified expressions as follows.

$$\begin{aligned}\partial_\alpha z &= \frac{z}{\alpha} - 2\beta z^{3/2}\alpha^{-1/2}\delta(z,\alpha,\beta) \\ \partial_\beta z &= -2z^{3/2}\alpha^{1/2}\delta(z,\alpha,\beta)\end{aligned}$$

where we compute $\log(\delta(z,\alpha,\beta))$ for numerical stability since the logarithm of Gaussian cumulative distribution function can be computed by using existing libraries, such as the scipy.special.log_ndtr() function.

In fact, we have closed-form expressions of gradients of the entropy term as shown below.

$$\begin{aligned}\mathbb{E}_{q(z|\eta)}\left[-\log q(z|\boldsymbol{\eta})\right] &= \tfrac{1}{2}\left[-\log\alpha - 3\left(\log\beta + \exp(2\alpha\beta)E_1(2\alpha\beta)\right) + 1 + \log(2\pi)\right] \\ \partial_\alpha\mathbb{E}_{q(z|\eta)}\left[-\log q(z|\boldsymbol{\eta})\right] &= \frac{1}{\alpha} - 3\beta\exp(2\alpha\beta)E_1(2\alpha\beta) \\ \partial_\beta\mathbb{E}_{q(z|\eta)}\left[-\log q(z|\boldsymbol{\eta})\right] &= -3\alpha\exp(2\alpha\beta)E_1(2\alpha\beta)\end{aligned}$$

where $E_1(x) := \int_x^\infty \frac{e^{-t}}{t}dt$ is the exponential integral. It is not numerical stable to compute the product $\exp(x)E_1(x)$ when $x > 100$. In this case, we can use the asymptotic expansion (see Eq 3 at Tseng & Lee (1998)) for the exponential integral to approximate the product as shown below.

$$\exp(x)E_1(x) \approx \frac{1}{x}\left[1 + \sum_{n=1}^{N}\frac{(-1)^n n!}{x^n}\right] \quad \text{when } x > 100,$$

where N is an integer such as $N \le x < N+1$.

I. Mixture of Exponential Family Distributions

Let's consider the following mixture of exponential family distributions $q(\mathbf{z}) = \int q(\mathbf{z},\mathbf{w})d\mathbf{w}$. The joint distribution $q(\mathbf{z},\mathbf{w}|\boldsymbol{\lambda}) = q(\mathbf{w}|\boldsymbol{\lambda}_w)q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)$ is known as the conditional exponential family (CEF) defined by Lin et al. (2019a).

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

where $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w\}$.

We will use the joint Fisher information matrix(FIM) suggested by Lin et al. (2019a) as the metric $\mathbf{F}$ to derive our improved learning rule for mixture approximations.

I.1. The Joint Fisher Information Matrix and the Christoffel Symbol

Lin et al. (2019a) propose to use the FIM of the joint distribution $q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda})$, where they refer this FIM as the joint FIM. The joint FIM and the corresponding Christoffel symbol of the first kind are defined as follows.

$$\begin{aligned} F_{ab} &:= -\mathbb{E}_{q(w,z|\lambda)}\left[\partial_a \partial_b \log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda})\right] \\ \Gamma_{d,ab} &:= \tfrac{1}{2}\left[\partial_a F_{bd} + \partial_b F_{ad} - \partial_d F_{ab}\right] \end{aligned}$$

where we denote $\partial_a = \partial_{\lambda^a}$ for notation simplicity.

Like the exponential family cases as shown in Eq. (17), the Christoffel symbol of the first kind can be computed as

$$\begin{aligned} \Gamma_{d,ab} = \tfrac{1}{2}\Big[&\mathbb{E}_{q(w,z|\lambda)}\left[\partial_a \partial_b \log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda}) \partial_d \log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda})\right] - \mathbb{E}_{q(w,z|\lambda)}\left[\partial_b \partial_d \log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda}) \partial_a \log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda})\right] \\ &- \mathbb{E}_{q(w,z|\lambda)}\left[\partial_a \partial_d \log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda}) \partial_b \log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda})\right] - \mathbb{E}_{q(w,z|\lambda)}\left[\partial_a \partial_b \partial_d \log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda})\right]\Big] \end{aligned} \tag{26}$$

I.2. The BCN Parameterization

Now, we show that how to simplify the computation of the Christoffel symbol by extending the BCN parameterization for this kind of mixtures.

To this end, we first assume that $\boldsymbol{\lambda}$ can be partitioned with $(m+n)$ blocks to satisfy Assumption 1 in the main text.

$$\boldsymbol{\lambda} = \{\underbrace{\boldsymbol{\lambda}_z^{[1]}, \ldots, \boldsymbol{\lambda}_z^{[m]}}_{\boldsymbol{\lambda_z}}, \underbrace{\boldsymbol{\lambda}_w^{[m+1]}, \ldots, \boldsymbol{\lambda}_w^{[m+n]}}_{\boldsymbol{\lambda_w}}\}$$

Then, we extend the definition of BC parameterization to the conditional exponential family, which is similar to Assumption 2 in the main text and a concrete example of Definition 1 in Appendix B.1.

Assumption 2 [Block Coordinate Parameterization] : *A parameterization satisfied Assumption 1 is block coordinate (BC) if the joint FIM under this parameterization is block-diagonal according to the block structure of the parameterization.*

Lin et al. (2019a) show that, for any parameterization $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w\}$, the joint FIM has the following two blocks: $\mathbf{F}_z$ for block $\boldsymbol{\lambda}_z$ and $\mathbf{F}_w$ for block $\boldsymbol{\lambda}_w$ as given below.

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_z & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_w \end{bmatrix}$$

Assumption 2 implies that $\mathbf{F}_w$ and $\mathbf{F}_z$ are both block-diagonal according to the block structure of $\boldsymbol{\lambda}_w$ and $\boldsymbol{\lambda}_z$, respectively. The block diagonal structure is given below if $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_z^{[1]}, \ldots, \boldsymbol{\lambda}_z^{[m]}, \boldsymbol{\lambda}_w^{[m+1]}, \ldots, \boldsymbol{\lambda}_w^{[m+n]}\}$ is a BC parameterization.

$$\mathbf{F} = \begin{bmatrix} \overbrace{\begin{bmatrix} \mathbf{F}_z^{[1]} & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{F}_z^{[m]} \end{bmatrix}}^{\mathbf{F}_z} & \mathbf{0} \\ \mathbf{0} & \underbrace{\begin{bmatrix} \mathbf{F}_w^{[m+1]} & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{F}_w^{[m+n]} \end{bmatrix}}_{\mathbf{F}_w} \end{bmatrix}$$

Assumption 3 [Block Natural Parameterization for the Conditional Exponential-Family] : *For a conditional exponential-family distribution $q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda}) = q(\mathbf{w}|\boldsymbol{\lambda}_w)q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$, a parameterization $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w\}$ has the following properties.*

- *$\boldsymbol{\lambda}_w$ is a BCN parameterization of the exponential family distribution $q(\mathbf{w}|\boldsymbol{\lambda}_w)$ as defined in the main text.*
- *$\boldsymbol{\lambda}_z$ is a parameterization of $q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$, where there exist function ϕ_{z_i} and h_{z_i} for each block $\boldsymbol{\lambda}_z^{[i]}$ such that conditioning on $\mathbf{w}$, $q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$ can be re-expressed as a minimal conditional exponential family distribution (see Lin et al. (2019a) for the definition of the minimality) given that the rest of blocks $\boldsymbol{\lambda}_z^{[-i]}$ are known.*

$$q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) \equiv h_{z_i}(\mathbf{w}, \mathbf{z}, \boldsymbol{\lambda}_z^{[-i]}) \exp\left[\langle \boldsymbol{\phi}_{z_i}(\mathbf{w}, \mathbf{z}, \boldsymbol{\lambda}_z^{[-i]}), \boldsymbol{\lambda}_z^{[i]} \rangle - A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right]$$

We say $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w\}$ is a BCN parameterization for the mixture if it satisfies Assumption 1 to 3.

Many mixture approximations studied in Lin et al. (2019a) have a BCN parameterization. For concrete examples, see Appendix J and K.

I.3. Our Learning Rule for Mixture Approximations

Now, we are ready to discuss the learning rule for mixture approximations. Under a BC parameterization $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w\}$, our learning rule remains the same as shown below.

$$\lambda^{c_i} \leftarrow \lambda^{c_i} - t\hat{g}^{c_i} - \frac{t^2}{2}\Gamma^{c_i}_{a_i b_i}\hat{g}^{a_i}\hat{g}^{b_i}$$

where block i can be either a block of $\boldsymbol{\lambda}_w$ or $\boldsymbol{\lambda}_z$.

First, note that the sub-block matrix $\mathbf{F}_w$ of the joint FIM is indeed the FIM of $q(\mathbf{w}|\boldsymbol{\lambda}_w)$. Furthermore, $q(\mathbf{w}|\boldsymbol{\lambda}_w)$ is an exponential family distribution. If $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_z, \boldsymbol{\lambda}_w\}$ is a BCN parameterization, it is easy to see that the computation of the Christoffel symbol for $\boldsymbol{\lambda}_w$ is exactly the same as the exponential family cases as discussed in Appendix D.

Furthermore, we can simplify the Christoffel symbol for $\boldsymbol{\lambda}_z$ due to the following Theorem.

Theorem 4 *If $\boldsymbol{\lambda}$ is a BCN parameterization of a conditional exponential family (CEF) with the joint FIM, natural gradient and the Christoffel symbol of the first kind for block $\boldsymbol{\lambda}_z^{[i]}$ can be simplified as*

$$\hat{g}^{a_i} = \partial_{m_{z_{a_i}}}\mathcal{L} \;;\; \Gamma_{d_i, a_i b_i} = \tfrac{1}{2}\mathbb{E}_{q(w|\lambda_w)}\left[\partial_{\lambda_{z^{a_i}}}\partial_{\lambda_{z^{b_i}}}\partial_{\lambda_{z^{d_i}}} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right]$$

where $\lambda_z^{a_i}$ is the a-th element of $\boldsymbol{\lambda}_z^{[i]}$; $m_{z_{a_i}}$ denotes the a-th element of the block coordinate expectation parameter $\mathbf{m}_{z_{[i]}} = \mathbb{E}_{q(w,z|\lambda)}\left[\boldsymbol{\phi}_{z_i}(\mathbf{w}, \mathbf{z}, \boldsymbol{\lambda}_z^{[-i]})\right] = \mathbb{E}_{q(w|\lambda_w)}\left[\partial_{\lambda_z^{[i]}} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right]$.

I.4. Proof of Theorem 4

Proof: We assume $\boldsymbol{\lambda}_z = \{\boldsymbol{\lambda}_z^{[1]}, \cdots, \boldsymbol{\lambda}_z^{[m]}\}$ is partitioned with m blocks.

Since $\boldsymbol{\lambda}$ is a BCN parameterization, conditioning on $\mathbf{w}$ and given $\boldsymbol{\lambda}_z^{[-i]}$ and $\boldsymbol{\lambda}_w$ are known, we can re-express $q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$ as

$$q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) = h_{z_i}(\mathbf{z}, \mathbf{w}, \boldsymbol{\lambda}_z^{[-i]}) \exp\left[\langle \boldsymbol{\phi}_{z_i}(\mathbf{z}, \mathbf{w}, \boldsymbol{\lambda}_z^{[-i]}), \boldsymbol{\lambda}_z^{[i]} \rangle - A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right]$$

where $q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$ is also a one-parameter EF distribution conditioning on $\boldsymbol{\lambda}_z^{[-i]}$ and $\mathbf{w}$. Similarly, we have the following results.

$$\partial_{a_i}\partial_{b_i}\log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) = -\partial_{a_i}\partial_{b_i} A_z(\boldsymbol{\lambda}_z, \mathbf{w})$$
$$\mathbb{E}_{q(z|w,\lambda_z)}\left[\partial_{a_i}\log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right] = 0$$

where $\partial_{a_i} = \partial_{\lambda_z^{a_i}}$ is for notation simplicity. Using the above identities, we have

$$\begin{aligned}
\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{a_i}\partial_{b_i}\log q(\mathbf{z},\mathbf{w}|\boldsymbol{\lambda})\partial_{d_i}\log q(\mathbf{z},\mathbf{w}|\boldsymbol{\lambda})\right] &= \mathbb{E}_{q(z,w|\lambda)}\left[\partial_{a_i}\partial_{b_i}\log q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)\partial_{d_i}\log q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)\right] \\
&= \mathbb{E}_{q(w|\lambda_w)}\left[\mathbb{E}_{q(z|w,\lambda_z)}\left[\partial_{a_i}\partial_{b_i}\log q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)\partial_{d_i}\log q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)\right]\right] \\
&= -\mathbb{E}_{q(w|\lambda_w)}\Big[\partial_{a_i}\partial_{b_i}A_z(\boldsymbol{\lambda}_z,\mathbf{w})\underbrace{\mathbb{E}_{q(z|w,\lambda_z)}\left[\partial_{d_i}\log q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)\right]}_{0}\Big] \\
&= 0
\end{aligned}$$

Therefore, by Eq. (26), we can simplify the Christoffel symbol for $\boldsymbol{\lambda}_z^{[i]}$ as follows.

$$\begin{aligned}
\Gamma_{d_i,a_ib_i} &= -\tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{a_i}\partial_{b_i}\partial_{d_i}\log q(\mathbf{z},\mathbf{w}|\boldsymbol{\lambda})\right] \\
&= -\tfrac{1}{2}\mathbb{E}_{q(w|\lambda_w)}\left[\partial_{a_i}\partial_{b_i}\partial_{d_i}\log q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)\right] \\
&= \tfrac{1}{2}\mathbb{E}_{q(w|\lambda_w)}\left[\partial_{a_i}\partial_{b_i}\partial_{d_i}A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right]
\end{aligned}$$

where we use d_i to denote the d-th entry of block $\boldsymbol{\lambda}_z^{[i]}$.

Likewise, let $\mathbf{m}_{z_{[i]}} = \mathbb{E}_{q(z,w|\lambda)}\left[\boldsymbol{\phi}_{z_i}(\mathbf{z},\mathbf{w},\boldsymbol{\lambda}_z^{[-i]})\right]$ denote the block coordinate expectation parameter. We have

$$0 = \mathbb{E}_{q(w|\lambda_w)}\Big[\underbrace{\mathbb{E}_{q(z|w,\lambda_z)}\left[\partial_{a_i}\log q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)\right]}_{0}\Big] = m_{z_{a_i}} - \mathbb{E}_{q(w|\lambda_w)}\left[\partial_{a_i}A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right]$$

where $m_{z_{a_i}}$ denotes the a-th element of $\mathbf{m}_{z_{[i]}}$.

Therefore, we know that $m_{z_{a_i}} = \mathbb{E}_{q(w|\lambda_w)}\left[\partial_{a_i}A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right]$.

Recall that the sub-block of the joint FIM for $\boldsymbol{\lambda}_z^{[i]}$ denoted by $\mathbf{F}_z^{[i]}$ can be computed as

$$\begin{aligned}
F_{a_ib_i} &= -\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{b_i}\partial_{a_i}\log q(\mathbf{z},\mathbf{w}|\boldsymbol{\lambda})\right] \\
&= -\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{b_i}\partial_{a_i}\log q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)\right] \\
&= -\mathbb{E}_{q(z,w|\lambda)}\left[-\partial_{b_i}\partial_{a_i}A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right] \\
&= \mathbb{E}_{q(w|\lambda_w)}\left[\partial_{b_i}\partial_{a_i}A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right] \\
&= \partial_{b_i}\mathbb{E}_{q(w|\lambda_w)}\left[\partial_{a_i}A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right] \\
&= \partial_{b_i}m_{z_{a_i}}
\end{aligned}$$

where we use the fact that $\boldsymbol{\lambda}_w$ does not depend on $\lambda_z^{b_i} \in \boldsymbol{\lambda}_z$ and $\partial_{b_i} = \partial_{\lambda_z^{b_i}}$ to move from the fourth step to the fifth step.

Recall that when $\boldsymbol{\lambda}$ is a BC parameterization, the joint FIM $\mathbf{F}$ is block-diagonal as shown below.

$$\mathbf{F} = \begin{bmatrix} \overbrace{\begin{bmatrix} \mathbf{F}_z^{[1]} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{F}_z^{[m]} \end{bmatrix}}^{\mathbf{F}_z} & \mathbf{0} \\ \mathbf{0} & \underbrace{\begin{bmatrix} \mathbf{F}_w^{[m+1]} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{F}_w^{[m+n]} \end{bmatrix}}_{\mathbf{F}_w} \end{bmatrix}$$

If $\mathbf{F}_z^{[i]}$ is positive definite everywhere, we have

$$F^{a_ib_i} = \partial_{m_{z_{a_i}}}\lambda_z^{b_i}$$

The above assumption is true if given that $\boldsymbol{\lambda}_z^{[-i]}$ and $\boldsymbol{\lambda}_w$ are known, $q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda})$ is a one-parameter minimal CEF distribution (See Theorem 2 of Lin et al. (2019a)).

The above result implies that we can compute natural gradients as follows.

$$\hat{g}^{a_i} = F^{a_i b_i} g_{b_i} = \left[\partial_{m_{z_{a_i}}} \lambda_z^{b_i}\right]\left[\partial_{\lambda_z^{b_i}} \mathcal{L}\right] = \partial_{m_{z_{a_i}}} \mathcal{L}$$

where $g_{b_i} = \partial_{\lambda_z^{b_i}} \mathcal{L}$.

□

If we can interchange the differentiations and the integration, we can show, by Theorem 4, we have $\Gamma^{c_i}{}_{a_i b_i} = \frac{1}{2}\partial_{m_{z_{c_i}}} \partial_{\lambda_{z^{a_i}}} \partial_{\lambda_{z^{b_i}}} \mathbb{E}_{q(w|\lambda_w)}\left[A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right]$ since $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ is C^3-smooth w.r.t. $\boldsymbol{\lambda}_z^{[i]}$.

J. Example: Finite Mixture of Gaussians Approximation

We consider a K-mixture of Gaussians under this parameterization $\boldsymbol{\lambda} = \{\{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K, \boldsymbol{\lambda}_w\}$

$$q(\mathbf{z}|\boldsymbol{\pi}, \{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K) = \sum_{c=1}^K \pi_c \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_c, \mathbf{S}_c)$$

where π_c is the mixing weight so that $\sum_{c=1}^K \pi_c = 1$, $\mathbf{S}_c = \boldsymbol{\Sigma}_c^{-1}$, $\boldsymbol{\lambda}_w = \{\log(\pi_c/\pi_K)\}_{c=1}^{K-1}$ and $\pi_K = 1 - \sum_{c=1}^{K-1} \pi_c$. The constraints are $\boldsymbol{\lambda}_w \in \mathbb{R}^{K-1}$, $\boldsymbol{\mu}_c \in \mathbb{R}^d$, and $\mathbf{S}_c \in \mathbb{S}_{++}^{d\times d}$.

Under this parameterization, the joint distribution can be expressed as below.

$$\begin{aligned}
q(\mathbf{z}, w|\boldsymbol{\lambda}) &= q(w|\boldsymbol{\lambda}_w) q(\mathbf{z}|w, \{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K) \\
q(w|\boldsymbol{\lambda}_w) &= \exp(\sum_{c=1}^{K-1} \mathbb{I}(w=c)\lambda_{w_c} - A_w(\boldsymbol{\lambda}_w)) \\
q(\mathbf{z}|w, \{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K) &= \exp\Big(\sum_{c=1}^K \mathbb{I}(w=c)\left[-\tfrac{1}{2}\mathbf{z}^T\mathbf{S}_c\mathbf{z} + \mathbf{z}^T\mathbf{S}_c\boldsymbol{\mu}_c\right] - A_z(\{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K, w)\Big)
\end{aligned}$$

where $B(\boldsymbol{\mu}_c, \mathbf{S}_c) = \frac{1}{2}\left[\boldsymbol{\mu}_c^T\mathbf{S}_c\boldsymbol{\mu}_c - \log|\mathbf{S}_c/(2\pi)|\right]$, $A_z(\{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K, w) = \sum_{c=1}^K \mathbb{I}(w=c)B(\boldsymbol{\mu}_c, \mathbf{S}_c)$, $\lambda_{w_c} = \log(\frac{\pi_c}{\pi_K})$, $A_w(\boldsymbol{\lambda}_w) = \log(1 + \sum_{c=1}^{K-1}\exp(\lambda_{w_c}))$.

Lemma 11 *The joint FIM is block diagonal under this parameterization.*

$$\mathbf{F} = \begin{bmatrix}
\begin{bmatrix}\mathbf{F}_{\mu_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_{S_1}\end{bmatrix} & \cdots & \mathbf{0} & \mathbf{0} \\
\vdots & \ddots & \vdots & \vdots \\
\mathbf{0} & \cdots & \begin{bmatrix}\mathbf{F}_{\mu_K} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_{S_K}\end{bmatrix} & \mathbf{0} \\
\mathbf{0} & \cdots & \mathbf{0} & \mathbf{F}_w
\end{bmatrix}$$

Therefore, this parameterization is a BC parameterization.

Proof: We will prove this lemma by showing that all cross terms are zeros.

Case 1: First, we will show that cross terms (shown in red) between $\boldsymbol{\lambda}_w$ and $\boldsymbol{\lambda}_z := \{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K$ are zeros.

Let's denote λ_w^i be an element of $\boldsymbol{\lambda}_w$ and λ_z^j be an element of $\boldsymbol{\lambda}_z$. By the definition, each cross term in this case is defined as belows.

$$-\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\lambda_w^i}\partial_{\lambda_z^j}\log q(\mathbf{z}, w|\boldsymbol{\lambda})\right] = -\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\lambda_w^i}\partial_{\lambda_z^j}\big(\log q(w|\boldsymbol{\lambda}_w) + \log q(\mathbf{z}|w, \boldsymbol{\lambda}_z)\big)\right] = 0$$

Case 2: Next, we will show that cross terms between (shown in blue) any two Gaussian components are zeros.

Let's denote λ_a^i be an element of $\{\boldsymbol{\mu}_a, \mathbf{S}_a\}$ and λ_b^j be an element of $\{\boldsymbol{\mu}_b, \mathbf{S}_b\}$, where $a \neq b$.

By the definition, each cross term in this case is defined as belows.

$$
\begin{aligned}
&- \mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\lambda_a^i}\partial_{\lambda_b^j}\log q(\mathbf{z}, w|\boldsymbol{\lambda})\right] \\
=&- \mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\lambda_a^i}\partial_{\lambda_b^j}\left(\log q(\mathbf{z}|w, \{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K)\right)\right] \\
=&- \mathbb{E}_{q(z,w|\lambda)}\Big[\mathbb{I}(w=b)\partial_{\lambda_a^i}\underbrace{\left(\partial_{\lambda_b^j}\left[-\tfrac{1}{2}\mathbf{z}^T\mathbf{S}_b\mathbf{z} + \mathbf{z}^T\mathbf{S}_b\boldsymbol{\mu}_b - B(\boldsymbol{\mu}_b, \mathbf{S}_b)\right]\right)}_{u(\boldsymbol{z}, \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)}\Big] = 0
\end{aligned}
$$

It is obvious that the above expression is 0 since $\partial_{\lambda_a^i} u(\mathbf{z}, \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) = 0$ when $a \neq b$.

Case 3: Finally, we will show that for each component a, cross terms (shown in green) between $\boldsymbol{\mu}_a$ and $\mathbf{S}_a$ are zeros.

Let's denote μ_a^i be the i-th element of $\boldsymbol{\mu}_a$ and S_a^{jk} be the element of $\mathbf{S}_a$ at position (j,k). Furthermore, $\mathbf{e}_i$ denotes an one-hot vector where all entries are zeros except the i-th entry with value 1, and $\mathbf{I}_{jk}$ denotes an one-hot matrix where all entries are zeros except the entry at position (j,k) with value 1. By the definition, the cross term is defined as belows.

$$
\begin{aligned}
&- \mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\mu_a^i}\partial_{S_a^{jk}}\log q(\mathbf{z}, w|\boldsymbol{\lambda})\right] \\
=&- \mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\mu_a^i}\partial_{S_a^{jk}}\left(\log q(\mathbf{z}|w, \{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K)\right)\right] \\
=&- \mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\mu_a^i}\partial_{S_a^{jk}}\Big(\mathbb{I}(w=a)\left[-\tfrac{1}{2}\mathbf{z}^T\mathbf{S}_a\mathbf{z} + \mathbf{z}^T\mathbf{S}_a\boldsymbol{\mu}_a - B(\boldsymbol{\mu}_a, \mathbf{S}_a)\right]\Big)\right] \\
=&- \mathbb{E}_{q(z,w|\lambda)}\left[\mathbb{I}(w=a)\left[\mathbf{e}_i^T\mathbf{I}_{jk}\mathbf{z} - \mathbf{e}_i^T\mathbf{I}_{jk}\boldsymbol{\mu}_a\right]\right] \\
=&- \mathbb{E}_{q(z,w|\lambda)}\left[\mathbb{I}(w=a)\mathbf{e}_i^T\mathbf{I}_{jk}\mathbf{z}\right] + \mathbb{E}_{q(z,w|\lambda)}\left[\mathbb{I}(w=a)\mathbf{e}_i^T\mathbf{I}_{jk}\boldsymbol{\mu}_a\right] \\
=&- \pi_a\mathbf{e}_i^T\mathbf{I}_{jk}\boldsymbol{\mu}_a + \pi_a\mathbf{e}_i^T\mathbf{I}_{jk}\boldsymbol{\mu}_a = 0
\end{aligned}
$$

where we use the following fact in the last step.

$$
\begin{aligned}
\mathbb{E}_{q(z,w|\lambda)}\left[\mathbb{I}(w=a)\mathbf{z}\right] &= \pi_a\boldsymbol{\mu}_a \\
\mathbb{E}_{q(z,w|\lambda)}\left[\mathbb{I}(w=a)\right] &= \pi_a
\end{aligned}
$$

□

Lemma 12 *The parameterization $\boldsymbol{\lambda} = \{\{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K, \boldsymbol{\lambda}_w\}$ is a BCN parameterization.*

Proof: Clearly, this parameterization satisfies Assumption 1 described in the main text. By Lemma 11, we know that this parameterization is a BC parameterization. Now, we will show that this parameterization also satisfies Assumption 3 in Appendix I.2.

First note that $\boldsymbol{\lambda}_w$ has only one block and it is the natural parameterization of exponential family distribution $q(w|\boldsymbol{\lambda}_w)$, which implies that $\boldsymbol{\lambda}_w$ is a BCN parameterization for $q(w|\boldsymbol{\lambda}_w)$.

Note that given the rest blocks are known and conditioning on w, $q(\mathbf{z}|w, \boldsymbol{\lambda}_z)$ can be re-expressed as follows in terms of block $\boldsymbol{\mu}_k$.

$$
\begin{aligned}
&q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) = \exp\Big(\sum_{c=1}^K \mathbb{I}(w=c)\left[-\tfrac{1}{2}\mathbf{z}^T\mathbf{S}_c\mathbf{z} + \mathbf{z}^T\mathbf{S}_c\boldsymbol{\mu}_c\right] - A_z(\{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K, w)\Big) \\
=&\underbrace{\exp\Big(\sum_{c\neq k}\left[\mathbb{I}(w=c)\left[-\tfrac{1}{2}\mathbf{z}^T\mathbf{S}_c\mathbf{z} + \mathbf{z}^T\mathbf{S}_c\boldsymbol{\mu}_c\right]\right] + \mathbb{I}(w=k)\left[-\tfrac{1}{2}\mathbf{z}^T\mathbf{S}_k\mathbf{z}\right]\Big)}_{h_{z_{k_1}}(w, \mathbf{z}, \lambda_z^{[-k_1]})} \exp\Big(\langle \underbrace{\mathbb{I}(w=k)\mathbf{S}_k\mathbf{z}}_{\phi_{z_{k_1}}(w, \mathbf{z}, \lambda_z^{[-k_1]})}, \underbrace{\boldsymbol{\mu}_k}_{\lambda_z^{k_1}}\rangle - A_z(\{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K, w)\Big)
\end{aligned}
$$

Similarly, for block $\mathbf{S}_k$, $q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$ can be re-expressed as follows

$$q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$$
$$= \underbrace{\exp\left(\sum_{c\neq k}\left[\mathbb{I}(w=c)\left[-\tfrac{1}{2}\mathbf{z}^T\mathbf{S}_c\mathbf{z} + \mathbf{z}^T\mathbf{S}_c\boldsymbol{\mu}_c\right]\right]\right)}_{h_{z_{k_2}}(w,\mathbf{z},\lambda_z^{[-k_2]})} \exp\Big(\langle \underbrace{\mathbb{I}(w=k)\left[-\tfrac{1}{2}\mathbf{z}\mathbf{z}^T + \boldsymbol{\mu}_k\mathbf{z}^T\right]}_{\phi_{z_{k_2}}(w,\mathbf{z},\lambda_z^{[-k_2]})}, \underbrace{\mathbf{S}_k}_{\lambda_z^{k_2}}\rangle - A_z(\{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K, w)\Big)$$

Since this parameterization satisfies Assumption 1 to 3, this parameterization is a BCN parameterization. □

We denote the Christoffel symbols of the first kind and the second kind for $\boldsymbol{\mu}_k$ as $\Gamma_{a_{k_1},b_{k_1}c_{k_1}}$ and $\Gamma^{a_{k_1}}{}_{b_{k_1}c_{k_1}}$ respectively.

Lemma 13 *For each component k, all entries of $\Gamma^{a_{k_1}}{}_{b_{k_1}c_{k_1}}$ for $\boldsymbol{\mu}_k$ are zeros.*

Proof: The proof is very similar to the proof of Lemma 3. We will prove this by showing that all entries of $\Gamma_{a_{k_1},b_{k_1}c_{k_1}}$ are zeros. For notation simplicity, we use $\Gamma_{a,bc}$ to denote $\Gamma_{a_{k_1},b_{k_1}c_{k_1}}$. Let μ_k^a denote the a-th element of $\boldsymbol{\mu}_k$.

The following expression holds for any valid a, b, and c.

$$\begin{aligned}
\Gamma_{a,bc} &= \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\mu_k^b}\partial_{\mu_k^c}\partial_{\mu_k^a}A_z(\{\boldsymbol{\mu}_j, \mathbf{S}_j\}_{j=1}^K, w)\right] \\
&= \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\mathbb{I}(w=k)\partial_{\mu_k^b}\partial_{\mu_k^c}\partial_{\mu_k^a}B(\boldsymbol{\mu}_k, \mathbf{S}_k)\right] \\
&= \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\mathbb{I}(w=k)\partial_{\mu_k^b}\partial_{\mu_k^c}\left(\mathbf{e}_a^T\mathbf{S}_k\boldsymbol{\mu}_k\right)\right] \\
&= \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\Big[\mathbb{I}(w=k)\underbrace{\partial_{\mu_k^b}\left(\mathbf{e}_a^T\mathbf{S}_k\mathbf{e}_c\right)}_{0}\Big] = 0
\end{aligned}$$

where in the last step we use the fact that $\mathbf{S}_k$, $\mathbf{e}_a$, and $\mathbf{e}_c$ do not depend on $\boldsymbol{\mu}_k$.

□

Similarly, we denote the Christoffel symbols of the second kind for $\mathrm{vec}(\mathbf{S}_k)$ as $\Gamma^{a_{k_2}}{}_{b_{k_2}c_{k_2}}$.

Lemma 14 *For each component k, the additional term for $\mathbf{S}_k$ is $-\hat{\boldsymbol{g}}_k^{[2]}\mathbf{S}_k^{-1}\hat{\boldsymbol{g}}_k^{[2]}$*

Proof: Recall that, in the Gaussian case $\mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\mathbf{S}})$, the additional term for $\bar{\mathbf{S}}$ is $\mathrm{Mat}(\bar{\Gamma}^{a_2}{}_{b_2c_2}\hat{g}^{b_2}\hat{g}^{c_2}) = \hat{\boldsymbol{g}}^{[2]}\bar{\mathbf{S}}^{-1}\hat{\boldsymbol{g}}^{[2]}$, where $\bar{\Gamma}^{a_2}{}_{b_2c_2}$ denotes the Christoffel symbols of the second kind for $\mathrm{vec}(\bar{\mathbf{S}})$.

To prove the statement, we will show that the Christoffel symbols of the second kind for $\mathrm{vec}(\mathbf{S}_k)$ is exactly the same as the Gaussian case, when $\bar{\mathbf{S}} = \mathbf{S}_k$. In other words, when $\bar{\mathbf{S}} = \mathbf{S}_k$, we will show $\Gamma^{a_{k_2}}{}_{b_{k_2}c_{k_2}} = \bar{\Gamma}^{a_2}{}_{b_2c_2}$.

We denote the Christoffel symbols of the second kind for $\mathrm{vec}(\mathbf{S}_k)$ using $\Gamma^{a_{k_2}}{}_{b_{k_2}c_{k_2}}$. By definition, the Christoffel symbols of the second kind for $\mathrm{vec}(\mathbf{S}_k)$ is defined as follows since $\boldsymbol{\lambda}$ is a BC parameterization.

$$\Gamma^{a_{k_2}}{}_{b_{k_2}c_{k_2}} = F^{a_{k_2}d_{k_2}}\Gamma_{d_{k_2},b_{k_2}c_{k_2}}$$

We will first show that $\Gamma_{d_{k_2},b_{k_2}c_{k_2}} = \pi_k\bar{\Gamma}_{d_2,b_2c_2}$.

In the Gaussian case, by definition, we have

$$\bar{\Gamma}_{d_2,b_2c_2} = \tfrac{1}{2}\mathbb{E}_{q(z|\bar{\lambda})}\left[\partial_{\bar{S}^b}\partial_{\bar{S}^c}\partial_{\bar{S}^d}A(\bar{\boldsymbol{\mu}}, \bar{\mathbf{S}})\right] = -\frac{1}{4}\partial_{\bar{S}^b}\partial_{\bar{S}^c}\partial_{\bar{S}^d}\left(\log\left|\bar{\mathbf{S}}\right|\right)$$

where $A(\bar{\boldsymbol{\mu}}, \bar{\mathbf{S}}) = \frac{1}{2}\left[\bar{\boldsymbol{\mu}}^T\bar{\mathbf{S}}\bar{\boldsymbol{\mu}} - \log\left|\bar{\mathbf{S}}/(2\pi)\right|\right]$ is the log partition function of the Gaussian distribution and $\bar{S}^d$ denotes the d-th element of $\mathrm{vec}(\bar{\mathbf{S}})$ in the Gaussian case.

Therefore, we have the following result in the MOG case when $\mathbf{S}_k = \bar{\mathbf{S}}$.

$$\begin{aligned}
\Gamma_{d_{k_2}, b_{k_2} c_{k_2}} &= \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{S_k^b}\partial_{S_k^c}\partial_{S_k^d} A_z(\{\boldsymbol{\mu}_j, \mathbf{S}_j\}_{j=1}^K, w)\right] \\
&= \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\mathbb{I}(w=k)\partial_{S_k^b}\partial_{S_k^c}\partial_{S_k^d} B(\boldsymbol{\mu}_k, \mathbf{S}_k)\right] \\
&= \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\mathbb{I}(w=k)\partial_{S_k^b}\partial_{S_k^c}\partial_{S_k^d}\left(-\tfrac{1}{2}\log|\mathbf{S}_k/(2\pi)|\right)\right] \\
&= -\frac{\pi_k}{4}\partial_{S_k^b}\partial_{S_k^c}\partial_{S_k^d}\left(\log|\mathbf{S}_k|\right) \\
&= \pi_k \bar{\Gamma}_{d_2, b_2 c_2}
\end{aligned}$$

where S_k^a denotes the a-th element of $\mathrm{vec}(\mathbf{S}_k)$ and $\mathbb{E}_{q(z,w|\lambda)}\left[\mathbb{I}(w=k)\right] = \pi_k$.

Let $F_{a_{k_2} d_{k_2}}$ denote the element at position (a, d) of the sub-block matrix of the joint FIM for block $\mathrm{vec}(\mathbf{S}_k)$ in the MOG case. Similarly, when $\mathbf{S}_k = \bar{\mathbf{S}}$, we can show that $F_{a_{k_2} d_{k_2}} = \pi_k \bar{F}_{a_2 d_2}$, where $\bar{F}_{a_2 d_2}$ denotes the element at position (a, d) of the sub-block matrix of the FIM for block $\mathrm{vec}(\bar{\mathbf{S}})$ in the Gaussian case.

Therefore, $F^{a_{k_2} d_{k_2}} = \pi_k^{-1} \bar{F}^{a_2 d_2}$ when $\bar{\mathbf{S}} = \mathbf{S}_k$.

Finally, when $\bar{\mathbf{S}} = \mathbf{S}_k$, we obtain the desired result since

$$\Gamma^{a_{k_2}}_{\ b_{k_2} c_{k_2}} = F^{a_{k_2} d_{k_2}} \Gamma_{d_{k_2}, b_{k_2} c_{k_2}} = \left(\pi_k^{-1} \bar{F}^{a_2 d_2}\right)\left(\pi_k \bar{\Gamma}_{d_2, b_2 c_2}\right) = \bar{F}^{a_2 d_2} \bar{\Gamma}_{d_2, b_2 c_2} = \bar{\Gamma}^{a_2}_{\ b_2 c_2}$$

where $\bar{\Gamma}^{a_2}_{\ b_2 c_2}$ denotes the Christoffel symbols of the second kind for $\mathrm{vec}(\bar{\mathbf{S}})$ in the Gaussian case.

□

J.1. Natural Gradients

Recall that $\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(z|\lambda)}\left[\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z}) + \log q(\mathbf{z}|\boldsymbol{\lambda})\right]$, where $q(\mathbf{z}|\boldsymbol{\lambda}) = \int q(\mathbf{z}, w|\boldsymbol{\lambda}) dw$.

Lin et al. (2019a) propose to use the importance sampling technique so that the number of Monte Carlo gradient evaluations is independent of the number of mixing components K.

Note that $\boldsymbol{\lambda}_w$ is the natural parameter of exponential family distribution $q(w|\boldsymbol{\lambda}_w)$, we can obtain the natural gradient by computing the gradient w.r.t. the mean parameter as shown by Lin et al. (2019a).

$$\hat{g}_w = \partial_\pi \mathcal{L}.$$

where $\pi_c := \mathbb{E}_{q(w)}\left[\mathbb{I}(w=c)\right]$, $\partial_{\pi_c}\mathcal{L}$ denotes the c-th element of $\partial_\pi \mathcal{L}$, and the gradient $\partial_{\pi_c}\mathcal{L}$ can be computed as below as suggested by Lin et al. (2019a).

$$\partial_{\pi_c}\mathcal{L} = \mathbb{E}_{q(z)}[(\delta_c - \delta_K) b(\mathbf{z})]$$

where $b(\mathbf{z}) := \ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z}) + \log q(\mathbf{z}|\lambda)$, and $\delta_c := \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_c, \mathbf{S}_c) / \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k, \mathbf{S}_k)$.

Recall that $\boldsymbol{\lambda}_w$ is unconstrained in this case, there is no need to compute the addition term for $\boldsymbol{\lambda}_w$.

Now, we discuss how to compute the natural gradients $\{\hat{\boldsymbol{g}}_c^{[1]}, \hat{\boldsymbol{g}}_c^{[2]}\}_{c=1}^K$. Since $\{\boldsymbol{\mu}_c, \mathbf{S}_c\}_{c=1}^K$ are BCN parameters, we can obtain the natural gradients by computing gradients w.r.t. its BC expectation parameter due to Theorem 4.

Given the rest of blocks are known, the BC expectation parameter for block $\boldsymbol{\mu}_k$ is

$$\mathbf{m}_{k_1} = \mathbb{E}_{q(w,z)}\left[\mathbb{I}(w=k)\left(\mathbf{S}_k \mathbf{z}\right)\right] = \pi_k \mathbf{S}_k \boldsymbol{\mu}_k$$

In this case, we know that $\partial_{\mu_k}\mathcal{L} = \pi_k \mathbf{S}_k \partial_{m_{k_1}}\mathcal{L}$. Therefore, the natural gradient w.r.t. $\boldsymbol{\mu}_k$ is $\hat{\boldsymbol{g}}_k^{[1]} = \partial_{m_{k_1}}\mathcal{L} = \pi_k^{-1}\mathbf{S}_k^{-1}\partial_{\mu_k}\mathcal{L} = \pi_k^{-1}\boldsymbol{\Sigma}_k \partial_{\mu_k}\mathcal{L}$, where the gradient $\partial_{\mu_k}\mathcal{L}$ can be computed as belows as suggested by Lin et al. (2019a).

$$\partial_{\mu_k}\mathcal{L} = \mathbb{E}_{q(z)}[\pi_k \delta_k \nabla_z b(\mathbf{z})]$$

Likewise, given the rest of blocks are known, the BC expectation parameter for block $\mathbf{S}_k$ is

$$\mathbf{m}_{k_2} = \mathbb{E}_{q(w,z)}\left[\mathbb{I}(w=k)\left(-\tfrac{1}{2}\mathbf{z}\mathbf{z}^T + \boldsymbol{\mu}_k\mathbf{z}^T\right)\right] = \frac{\pi_k}{2}\left(\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T - \mathbf{S}_k^{-1}\right)$$

Therefore, the natural gradient w.r.t. $\mathbf{S}_k$ is $\hat{\boldsymbol{g}}_k^{[2]} = \partial_{m_{k_2}}\mathcal{L} = -\frac{2}{\pi_k}\partial_{S_k^{-1}} f = -\frac{2}{\pi_k}\partial_{\Sigma_k} f$, where where the gradient $\partial_{\Sigma_k} f$ can be computed as belows as suggested by Lin et al. (2019a).

$$\partial_{\Sigma_k}\mathcal{L} = \tfrac{1}{2}\mathbb{E}_{q(z)}\left[\pi_k\delta_k\nabla_z^2 b(\mathbf{z})\right]$$

Alternatively, we can use the re-parametrization trick to compute the gradient as below.

$$\partial_{\Sigma_k}\mathcal{L} = \tfrac{1}{2}\mathbb{E}_{q(z)}\left[\pi_k\delta_k\mathbf{S}_k(\mathbf{z}-\boldsymbol{\mu}_k)\nabla_z^T b(\mathbf{z})\right]$$

By Lemma 13 and 14, the proposed update induced by our rule is

$$\begin{aligned}
\log(\pi_c/\pi_K) &\leftarrow \log(\pi_c/\pi_K) - t\mathbb{E}_{q(z)}[(\delta_c - \delta_K)b(\mathbf{z})] \\
\boldsymbol{\mu}_c &\leftarrow \boldsymbol{\mu}_c - t\mathbf{S}_c^{-1}\mathbb{E}_{q(z)}[\delta_c\nabla_z b(\mathbf{z})] \\
\mathbf{S}_c &\leftarrow \mathbf{S}_c - t\hat{\mathbf{G}}_c + \frac{t^2}{2}\hat{\mathbf{G}}_c\left(\mathbf{S}_c\right)^{-1}\hat{\mathbf{G}}_c
\end{aligned} \tag{27}$$

where we do not compute the additional term for $\boldsymbol{\lambda}_w$ since $\boldsymbol{\lambda}_w$ is unconstrained, $\delta_c := \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_c, \mathbf{S}_c)/\sum_{k=1}^K \pi_k\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k, \mathbf{S}_k)$, $b(\mathbf{z}) := \ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z}) + \log q(\mathbf{z}|\boldsymbol{\lambda})$ and $\hat{\mathbf{G}}_c$ can be computed as below.

Note that $b(\mathbf{z})$ can be the logarithm of an unnormalized target function as such $b(\mathbf{z}) = \bar{\ell}(\mathbf{z}) + \text{Constant} + \log q(\mathbf{z}|\boldsymbol{\lambda})$. Recall that $\ell(\mathcal{D}, \mathbf{z}) - \log p(\mathbf{z}) = \bar{\ell}(\mathbf{z}) + \text{Constant}$. Lin et al. (2019a) suggest using the Hessian trick to compute $\hat{\mathbf{G}}_c$ as shown in (29). We can also use the re-parameterization trick to compute $\hat{\mathbf{G}}_c$ as shown in (28).

$$\begin{aligned}
\hat{\mathbf{G}}_c &= -\mathbb{E}_{q(z)}\left[\delta_c\mathbf{S}_c(\mathbf{z}-\boldsymbol{\mu}_c)\nabla_z^T b(\mathbf{z})\right] = -\mathbb{E}_{q(z)}\left[\delta_c\mathbf{S}_c(\mathbf{z}-\boldsymbol{\mu}_c)\nabla_z^T\bar{\ell}(\mathbf{z})\right] - \mathbb{E}_{q(z)}\left[\delta_c\nabla_z^2\log q(\mathbf{z}|\boldsymbol{\lambda})\right] && (28)\\
&= -\mathbb{E}_{q(z)}\left[\delta_c\nabla_z^2 b(\mathbf{z})\right] = -\mathbb{E}_{q(z)}\left[\delta_c\nabla_z^2\bar{\ell}(\mathbf{z})\right] - \mathbb{E}_{q(z)}\left[\delta_c\nabla_z^2\log q(\mathbf{z}|\boldsymbol{\lambda})\right]. && (29)
\end{aligned}$$

We use the MC approximation to compute $\hat{\mathbf{G}}_c$ as below.

$$\begin{aligned}
\hat{\mathbf{G}}_c &\approx -\delta_c\left(\frac{\bar{\mathbf{S}}_c + \bar{\mathbf{S}}_c^T}{2} + \nabla_z^2\log q(\mathbf{z}|\boldsymbol{\lambda})\right) && \text{referred to as "-rep"}\\
\hat{\mathbf{G}}_c &\approx -\delta_c\left(\nabla_z^2\bar{\ell}(\mathbf{z}) + \nabla_z^2\log q(\mathbf{z}|\boldsymbol{\lambda})\right) && \text{referred to as "-hess"}
\end{aligned}$$

where $\mathbf{z}\sim q(\mathbf{z}|\boldsymbol{\lambda})$, $\bar{\mathbf{S}}_c := \mathbf{S}_c(\mathbf{z}-\boldsymbol{\mu}_c)\nabla_z^T\bar{\ell}(\mathbf{z})$ and $\nabla_z^2\log q(\mathbf{z}|\boldsymbol{\lambda})$ can be manually coded or computed by Auto-Diff.

Recall that when $q(\mathbf{z}|\boldsymbol{\lambda})$ is Gaussian, $-\mathbb{E}_{q(z)}\left[\nabla_z^2\log q(\mathbf{z}|\boldsymbol{\lambda})\right] = \boldsymbol{\Sigma}^{-1}$, which is positive definite. VOGN is proposed to approximate $\mathbb{E}_{q(z)}\left[\nabla_z^2\bar{\ell}(\mathbf{z})\right]$ by a positive definite matrix when $q(\mathbf{z}|\boldsymbol{\lambda})$ is Gaussian. In MOG cases, $-\mathbb{E}_{q(z)}\left[\nabla_z^2\log q(\mathbf{z}|\boldsymbol{\lambda})\right]$ is no longer a positive definite matrix. VOGN does not guarantee that the update for $\mathbf{S}_c$ stays in the constraint set. Furthermore, directly approximating $-\hat{\mathbf{G}}_c$ by naively extending the idea of VOGN does not give a good posterior approximation. Unlike VOGN, our update satisfies the constraint without the loss of the approximation accuracy for both Gaussian and MOG cases.

K. Example: Skew Gaussian Approximation

We consider the skew Gaussian approximation proposed by Lin et al. (2019a). The joint distribution is given below.

$$\begin{aligned}
q(\mathbf{z}, w|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= q(\mathbf{z}|w, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\mathcal{N}(w|0,1)\\
q(\mathbf{z}|w, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + |w|\boldsymbol{\alpha}, \boldsymbol{\Sigma})\\
&= \exp(\{\text{Tr}\left(-\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{z}\mathbf{z}^T\right) + |w|\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\mathbf{z} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{z} - \tfrac{1}{2}((\boldsymbol{\mu}+|w|\boldsymbol{\alpha})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}+|w|\boldsymbol{\alpha}) + \log|2\pi\boldsymbol{\Sigma}|\})
\end{aligned}$$

We consider the parameterization $\boldsymbol{\lambda} = \{\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{bmatrix}, \mathbf{S}\}$, where $\mathbf{S} = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\lambda}^{[1]} = \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{bmatrix}$, and $\boldsymbol{\lambda}^{[2]} = \mathbf{S}$. The open-set constraint is $\boldsymbol{\lambda} \in \mathbb{R}^{2d} \times \mathbb{S}_{++}^{d\times d}$. Under this parameterization, the distribution $q(\mathbf{z}|w)$ can be re-expressed as below.

$$q(\mathbf{z}|w, \boldsymbol{\lambda}) = \exp\left\{\text{Tr}\left(-\tfrac{1}{2}\mathbf{S}\mathbf{z}\mathbf{z}^T\right) + \mathbf{z}^T\mathbf{S}\left(\mathbf{Q}(w)\right)^T\boldsymbol{\lambda}^{[1]} - A_z(\boldsymbol{\lambda}, w)\right\}$$

where $\mathbf{Q}(w) := \begin{bmatrix}\mathbf{I}_d\\ |w|\,\mathbf{I}_d\end{bmatrix}$ is a $2d$-by-d matrix and $A_z(\boldsymbol{\lambda}, w) = \frac{1}{2}\left[\begin{bmatrix}\boldsymbol{\mu}^T & \boldsymbol{\alpha}^T\end{bmatrix}\mathbf{Q}(w)\mathbf{S}\left(\mathbf{Q}(w)\right)^T\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{bmatrix} - \log|\mathbf{S}/(2\pi)|\right]$.

Lemma 15 *The joint FIM is block diagonal with two blocks under this parameterization.*

$$\mathbf{F} = \begin{bmatrix}\mathbf{F}^{[1]} & \mathbf{0}\\ \mathbf{0} & \mathbf{F}^{[2]}\end{bmatrix}$$

Therefore, this parameterization is a BC parameterization.

Proof: We will prove this lemma by showing that all cross terms shown in red are zeros.

Let's denote λ^{a_1} be the a-th element of $\boldsymbol{\lambda}^{[1]}$ and S^{bc} be the element of $\mathbf{S}$ at position (b, c). Furthermore, $\mathbf{e}_a$ denotes an one-hot vector where all entries are zeros except the a-th entry with value 1, and $\mathbf{I}_{bc}$ denotes an one-hot matrix where all entries are zeros except the entry at position (b, c) with value 1.

By definition, the cross term is defined as belows.

$$\begin{aligned}
&-\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\lambda^{a_1}}\partial_{S^{bc}}\log q(\mathbf{z}, w|\boldsymbol{\lambda})\right]\\
=&-\mathbb{E}_{q(z,w|\lambda)}\left[\mathbf{z}^T\mathbf{I}_{bc}\left(\mathbf{Q}(w)\right)^T\mathbf{e}_a - \left(\boldsymbol{\lambda}^{[1]}\right)^T\mathbf{Q}(w)\mathbf{I}_{bc}\left(\mathbf{Q}(w)\right)^T\mathbf{e}_a\right]\\
=&-\mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w,\lambda)}\left[\mathbf{z}^T\mathbf{I}_{bc}\left(\mathbf{Q}(w)\right)^T\mathbf{e}_a - \left(\boldsymbol{\lambda}^{[1]}\right)^T\mathbf{Q}(w)\mathbf{I}_{bc}\left(\mathbf{Q}(w)\right)^T\mathbf{e}_a\right]\right]\\
=&-\mathbb{E}_{q(w)}\left[\mathbb{E}_{q(z|w,\lambda)}\left[\mathbf{z}^T\mathbf{I}_{bc}\left(\mathbf{Q}(w)\right)^T\mathbf{e}_a\right] - \left(\boldsymbol{\lambda}^{[1]}\right)^T\mathbf{Q}(w)\mathbf{I}_{bc}\left(\mathbf{Q}(w)\right)^T\mathbf{e}_a\right]\\
=&-\mathbb{E}_{q(w)}\left[\left(\boldsymbol{\lambda}^{[1]}\right)^T\mathbf{Q}(w)\mathbf{I}_{bc}\left(\mathbf{Q}(w)\right)^T\mathbf{e}_a - \left(\boldsymbol{\lambda}^{[1]}\right)^T\mathbf{Q}(w)\mathbf{I}_{bc}\left(\mathbf{Q}(w)\right)^T\mathbf{e}_a\right] = 0
\end{aligned}$$

where we use the following expression in the last step.

$$\mathbb{E}_{q(z|w,\lambda)}\left[\mathbf{z}\right] = |w|\,\boldsymbol{\alpha} + \boldsymbol{\mu} = \left(\mathbf{Q}(w)\right)^T\boldsymbol{\lambda}^{[1]}$$

□

Note that another parameterization $\{\boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{S}\}$ is not a BC parameterization since the joint FIM is not block-diagonal under this parameterization.

Lemma 16 *Parameterization $\boldsymbol{\lambda}$ is a BCN parameterization.*

Proof: Clearly, this parameterization satisfies Assumption 1 described in the main text. By Lemma 15, we know that this parameterization is a BC parameterization. Now, we will show that this parameterization also satisfies Assumption 3 in Appendix I.2.

Note that given the rest blocks are known and conditioning on w, $q(\mathbf{z}|w, \boldsymbol{\lambda})$ can be re-expressed as follows in terms of block $\boldsymbol{\lambda}^{[1]}$.

$$\begin{aligned}
q(\mathbf{z}|w, \boldsymbol{\lambda}) &= \exp\left\{\text{Tr}\left(-\tfrac{1}{2}\mathbf{S}\mathbf{z}\mathbf{z}^T\right) + \mathbf{z}^T\mathbf{S}\left(\mathbf{Q}(w)\right)^T\boldsymbol{\lambda}^{[1]} - A_z(\boldsymbol{\lambda}, w)\right\}\\
&= \underbrace{\exp\left\{\text{Tr}\left(-\tfrac{1}{2}\mathbf{S}\mathbf{z}\mathbf{z}^T\right)\right\}}_{h_1(w,\mathbf{z},\boldsymbol{\lambda}^{[-1]})}\exp\left[\langle\ \underbrace{\mathbf{Q}(w)\mathbf{S}\mathbf{z}}_{\phi_1(w,\mathbf{z},\boldsymbol{\lambda}^{[-1]})}\ , \boldsymbol{\lambda}^{[1]}\rangle - A_z(\boldsymbol{\lambda}, w)\right]
\end{aligned}$$

Similarly, for block $\mathbf{S}$, $q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda})$ can be re-expressed as follows

$$q(\mathbf{z}|w, \boldsymbol{\lambda}) = \underbrace{1}_{h_2(w,\mathbf{z},\lambda^{[-2]})} \exp\Big[\underbrace{\langle -\tfrac{1}{2}\mathbf{z}\mathbf{z}^T + \mathbf{z}\left(\boldsymbol{\lambda}^{[1]}\right)^T \mathbf{Q}(w), \mathbf{S}\rangle}_{\phi_2(w,\mathbf{z},\lambda^{[-2]})} - A_z(\boldsymbol{\lambda}, w)\Big]$$

Since this parameterization satisfies Assumption 1 to 3, this parameterization is a BCN parameterization. □

We denote the Christoffel symbols of the first kind and the second kind for $\boldsymbol{\lambda}^{[1]}$ as Γ_{a_1,b_1c_1} and $\Gamma^{a_1}{}_{b_1c_1}$ respectively.

Lemma 17 *All entries of* $\Gamma^{a_1}{}_{b_1c_1}$ *for* $\boldsymbol{\lambda}^{[1]}$ *are zeros.*

Proof: We will prove this by showing that all entries of $\Gamma^{a_1}_{b_1c_1}$ are zeros. Let λ^{a_1} denote the a-th element of $\boldsymbol{\lambda}^{[1]}$.

The following expression holds for any valid a, b, and c.

$$\begin{aligned}
\Gamma_{a_1,b_1c_1} &= \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\lambda^{b_1}}\partial_{\lambda^{c_1}}\partial_{\lambda^{a_1}} A_z(\boldsymbol{\lambda}, w)\right] \\
&= \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\lambda^{b_1}}\partial_{\lambda^{c_1}}\left((\mathbf{e}_a)^T \mathbf{Q}(w)\mathbf{S}\left(\mathbf{Q}(w)\right)^T \boldsymbol{\lambda}^{[1]}\right)\right] \\
&= \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{\lambda^{b_1}}\left(\mathbf{e}_a^T\mathbf{Q}(w)\mathbf{S}\left(\mathbf{Q}(w)\right)^T \mathbf{e}_c\right)\right] = 0
\end{aligned}$$

where in the last step we use the fact that $\mathbf{S}$, $\mathbf{Q}(w)$, $\mathbf{e}_a$, and $\mathbf{e}_c$ do not depend on $\boldsymbol{\lambda}^{[1]}$.

□

We denote the Christoffel symbols of the second kind for $\mathrm{vec}(\mathbf{S})$ as $\Gamma^{a_2}{}_{b_2c_2}$.

Lemma 18 *The additional term for* $\mathbf{S}$ *is* $-\hat{\boldsymbol{g}}^{[2]}\mathbf{S}^{-1}\hat{\boldsymbol{g}}^{[2]}$

Proof: Recall that, in the Gaussian case $\mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\mathbf{S}})$, the additional term for $\bar{\mathbf{S}}$ is $\mathrm{Mat}(\bar{\Gamma}^{a_2}_{b_2c_2}\hat{g}^{b_2}\hat{g}^{c_2}) = \hat{\boldsymbol{g}}^{[2]}\bar{\mathbf{S}}^{-1}\hat{\boldsymbol{g}}^{[2]}$, where $\bar{\Gamma}^{a_2}_{b_2c_2}$ denotes the Christoffel symbols of the second kind for $\mathrm{vec}(\bar{\mathbf{S}})$.

To prove the statement, we will show that the Christoffel symbols of the second kind for $\mathrm{vec}(\mathbf{S})$ is exactly the same as the Gaussian case, when $\bar{\mathbf{S}} = \mathbf{S}$.

We denote the Christoffel symbols of the second kind for $\mathrm{vec}(\mathbf{S})$ as $\Gamma^{a_2}_{b_2c_2}$. By definition, the Christoffel symbols of the second kind for $\mathrm{vec}(\mathbf{S})$ is defined as follows.

$$\Gamma^{a_2}_{b_2c_2} = F^{a_2d_2}\Gamma_{d_2,b_2c_2}$$

We will show that $\Gamma_{a_2,b_2c_2} = \bar{\Gamma}_{a_2,b_2c_2}$.

In the Gaussian case, we have

$$\bar{\Gamma}_{d_2,b_2c_2} = -\frac{1}{4}\partial_{\bar{S}^b}\partial_{\bar{S}^c}\partial_{\bar{S}^d}\left(\log\left|\bar{\mathbf{S}}\right|\right)$$

where $A(\bar{\boldsymbol{\mu}}, \bar{\mathbf{S}}) = \frac{1}{2}\left[\bar{\boldsymbol{\mu}}^T\bar{\mathbf{S}}\bar{\boldsymbol{\mu}} - \log\left|\bar{\mathbf{S}}/(2\pi)\right|\right]$ is the log partition function of the Gaussian distribution and $\bar{S}^a$ is the a-th element of $\mathrm{vec}(\bar{\mathbf{S}})$ in the Gaussian case.

Therefore, we have the following result when $\bar{\mathbf{S}} = \mathbf{S}$.

$$\Gamma_{d_2,b_2c_2} = \tfrac{1}{2}\mathbb{E}_{q(z,w|\lambda)}\left[\partial_{S^b}\partial_{S^c}\partial_{S^d} A_z(\boldsymbol{\lambda}, w)\right] = -\frac{1}{4}\partial_{S^b}\partial_{S^c}\partial_{S^d}\log|\mathbf{S}| = \bar{\Gamma}_{d_2,b_2c_2}$$

where S^a denotes the a-th element of $\mathrm{vec}(\mathbf{S})$.

Let $F_{a_2d_2}$ denote the element at position (a, d) of the sub-block matrix of the joint FIM for $\mathrm{vec}(\mathbf{S})$. Similarly, we can show that $F_{a_2d_2} = \bar{F}_{a_2d_2}$, where $\bar{F}_{a_2d_2}$ denotes the element at position (a, d) of the FIM for $\mathrm{vec}(\bar{\mathbf{S}})$ in the Gaussian case. Therefore, $F^{a_2d_2} = \bar{F}^{a_2d_2}$.

Finally, when $\bar{\mathbf{S}} = \mathbf{S}$, we obtain the desired result since

$$\Gamma^{a_2}_{b_2c_2} = F^{a_2d_2}\Gamma_{d_2,b_2c_2} = \bar{F}^{a_2d_2}\bar{\Gamma}_{d_2,b_2c_2} = \bar{\Gamma}^{a_2}_{b_2c_2}$$

where $\bar{\Gamma}^{a_2}_{b_2c_2}$ denotes the Christoffel symbols of the second kind for $\mathrm{vec}(\bar{\mathbf{S}})$ in the Gaussian case. □

Using these lemmas, the proposed update induced by our rule is

$$\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{bmatrix} \leftarrow \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{bmatrix} - t\hat{\boldsymbol{g}}^{[1]}$$

$$\mathbf{S} \leftarrow \mathbf{S} - t\hat{\boldsymbol{g}}^{[2]} + \frac{t^2}{2}\hat{\boldsymbol{g}}^{[2]}\mathbf{S}^{-1}\hat{\boldsymbol{g}}^{[2]}$$

where $\hat{\boldsymbol{g}}^{[1]}$ and $\hat{\boldsymbol{g}}^{[2]}$ are natural gradients.

Similarly, it can be shown that the above update satisfies the underlying constraints.

K.1. Natural Gradients

Now, we discuss how to compute the natural gradients. Since the parameterization is a BCN parameterization, gradients w.r.t. BC expectation parameters are natural gradients for BCN parameters due to Theorem 4.

Recall that $\boldsymbol{\lambda}^{[1]} = \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{bmatrix}$. Let $\mathbf{m}_{[1]} = \begin{bmatrix}\mathbf{m}_\mu\\ \mathbf{m}_\alpha\end{bmatrix}$ denote the BC expectation parameter for $\boldsymbol{\lambda}^{[1]}$. Given $\mathbf{S}$ is known, the BC expectation parameter is

$$\begin{aligned}\begin{bmatrix}\mathbf{m}_\mu\\ \mathbf{m}_\alpha\end{bmatrix} &= \mathbb{E}_{q(w,z)}\left[\mathbf{Q}(w)\mathbf{S}\mathbf{z}\right]\\ &= \mathbb{E}_{q(w)}\left[\mathbf{Q}(w)\mathbf{S}\left(\mathbf{Q}(w)\right)^T\boldsymbol{\lambda}^{[1]}\right]\\ &= \mathbb{E}_{q(w)}\left[\begin{bmatrix}\mathbf{S} & |w|\,\mathbf{S}\\ |w|\,\mathbf{S} & w^2\mathbf{S}\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{bmatrix}\right]\\ &= \begin{bmatrix}\mathbf{S} & c\mathbf{S}\\ c\mathbf{S} & \mathbf{S}\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{S}\boldsymbol{\mu} + c\mathbf{S}\boldsymbol{\alpha}\\ c\mathbf{S}\boldsymbol{\mu} + \mathbf{S}\boldsymbol{\alpha}\end{bmatrix}\end{aligned}$$

where $c = \mathbb{E}_{q(w)}\left[|w|\right] = \sqrt{\frac{2}{\pi}}$.

Since $\mathbf{S} = \boldsymbol{\Sigma}^{-1}$, we have the following expressions.

$$\boldsymbol{\mu} = \frac{1}{1-c^2}\boldsymbol{\Sigma}\left(\mathbf{m}_\mu - c\mathbf{m}_\alpha\right), \quad \boldsymbol{\alpha} = \frac{1}{1-c^2}\boldsymbol{\Sigma}\left(\mathbf{m}_\alpha - c\mathbf{m}_\mu\right)$$

By the chain rule, we have

$$\partial_{m_\mu}\mathcal{L} = \boldsymbol{\Sigma}\left(\frac{1}{1-c^2}\partial_\mu\mathcal{L} - \frac{c}{1-c^2}\partial_\alpha\mathcal{L}\right), \quad \partial_{m_\alpha}\mathcal{L} = \boldsymbol{\Sigma}\left(\frac{1}{1-c^2}\partial_\alpha\mathcal{L} - \frac{c}{1-c^2}\partial_\mu\mathcal{L}\right)$$

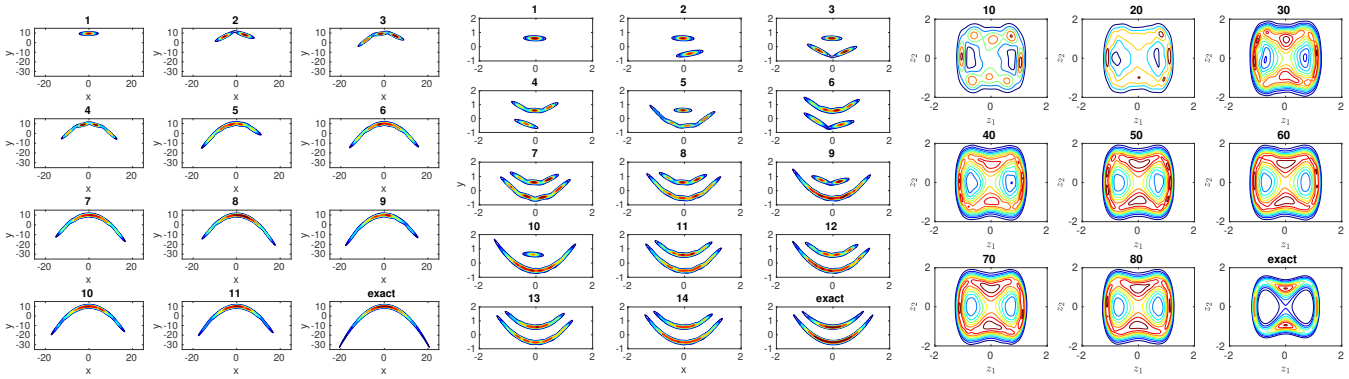
Therefore, the natural gradient w.r.t. $\boldsymbol{\lambda}^{[1]} = \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{bmatrix}$ is $\hat{\boldsymbol{g}}^{[1]} = \begin{bmatrix}\partial_{m_\mu}\mathcal{L}\\ \partial_{m_\alpha}\mathcal{L}\end{bmatrix}$ where the gradient $\partial_\mu\mathcal{L}$ and $\partial_\alpha\mathcal{L}$ can be computed as suggested by Lin et al. (2019a).

Likewise, the BC expectation parameter for block $\mathbf{S}$ is

$$\mathbf{m}_{[2]} = \mathbb{E}_{q(w,z)}\left[-\tfrac{1}{2}\mathbf{z}\mathbf{z}^T + \mathbf{z}\left(\boldsymbol{\lambda}^{[1]}\right)^T\mathbf{Q}(w)\right] = -\tfrac{1}{2}\mathbf{S}^{-1} + \mathbb{E}_{q(w)}\left[\tfrac{1}{2}\left(\mathbf{Q}(w)\right)^T\boldsymbol{\lambda}^{[1]}\left(\boldsymbol{\lambda}^{[1]}\right)^T\mathbf{Q}(w)\right]$$

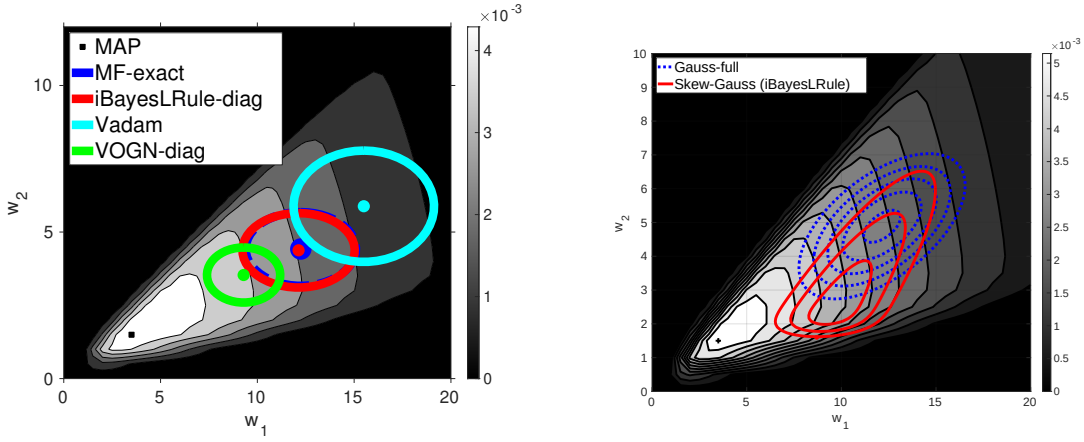
Since $\boldsymbol{\lambda}^{[1]}$ is known, $\mathbb{E}_{q(w)}\left[\tfrac{1}{2}\left(\mathbf{Q}(w)\right)^T\boldsymbol{\lambda}^{[1]}\left(\boldsymbol{\lambda}^{[1]}\right)^T\mathbf{Q}(w)\right]$ does not depend on $\mathbf{S}$. Therefore, the natural gradient w.r.t. $\mathbf{S}$ is $\hat{\boldsymbol{g}}^{[2]} = \partial_{m_{[2]}}\mathcal{L} = -2\partial_{S^{-1}}\mathcal{L} = -2\partial_\Sigma\mathcal{L}$, where we compute the gradient $\partial_\Sigma\mathcal{L}$ as suggested by Lin et al. (2019a).

L. More Results



(a) (b) (c)

Figure 5. The leftmost figure is MOG approximations for the banana distribution mentioned at Section 6.1, where the number indicates the number of components used in the approximations. The middle figure is a complete version of MOG approximations for the double banana distribution (the rightmost plot in Figure 2), where the number indicates the number of components used in the approximations. The rightmost figure is MOG approximations for the posterior $p(\mathbf{z}|y=1)$ of a BNN with a Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$ and a NN likelihood $p(y|\mathbf{z}) = \mathcal{N}(y|3z_1^2(z_1^2-1)+z_2^2, 0.5^2)$, where the number indicates the number of components used in the approximations.



(a) (b)

Figure 6. The leftmost plot is mean-field Gaussian approximations for the toy Bayesian logistic regression example considered at Section 6.1, where Vadam is proposed by Khan et al. (2018). The rightmost plot is a skew-Gaussian approximation with full covariance structure for the same example.

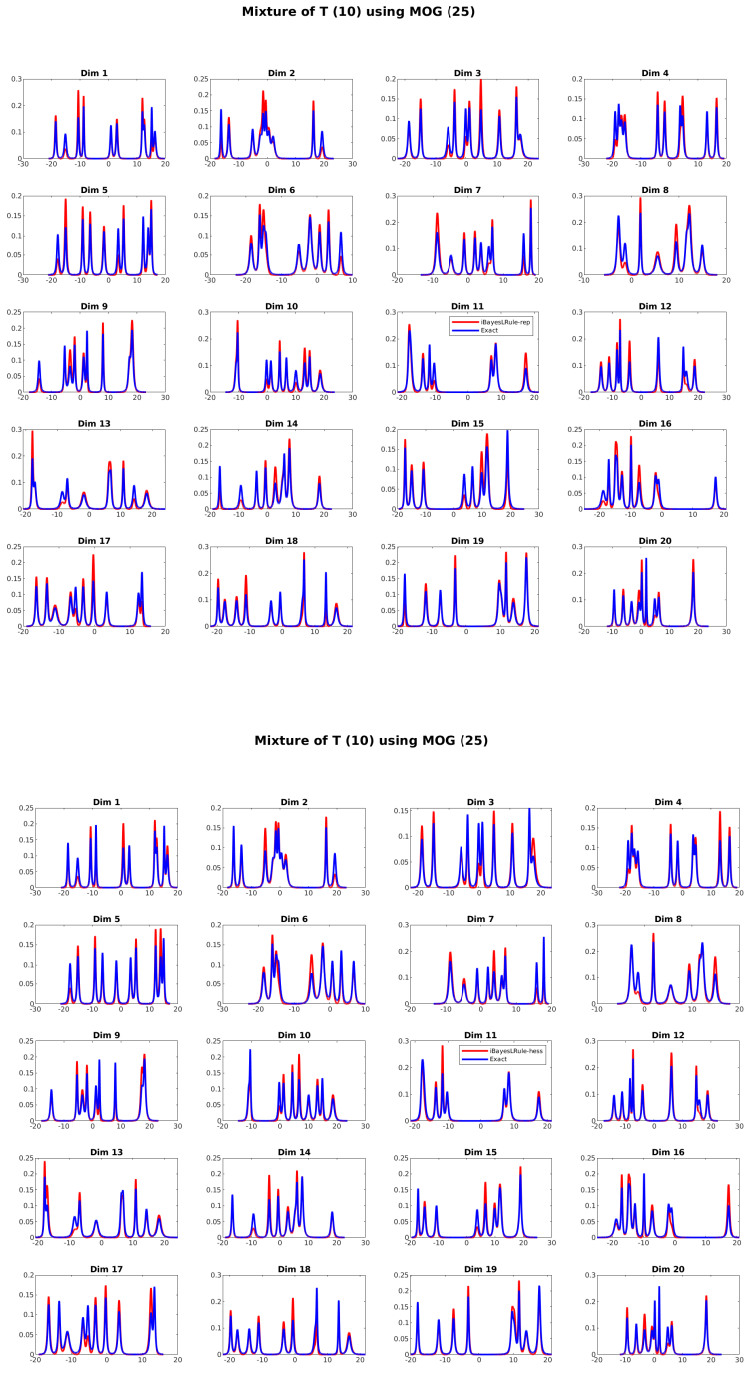


Figure 7. This is a complete version of the leftmost figure in Figure 2. The figure shows MOG approximation (with $K = 25$) to fit an MOG model with 10 components in a 20 dimensional problem.

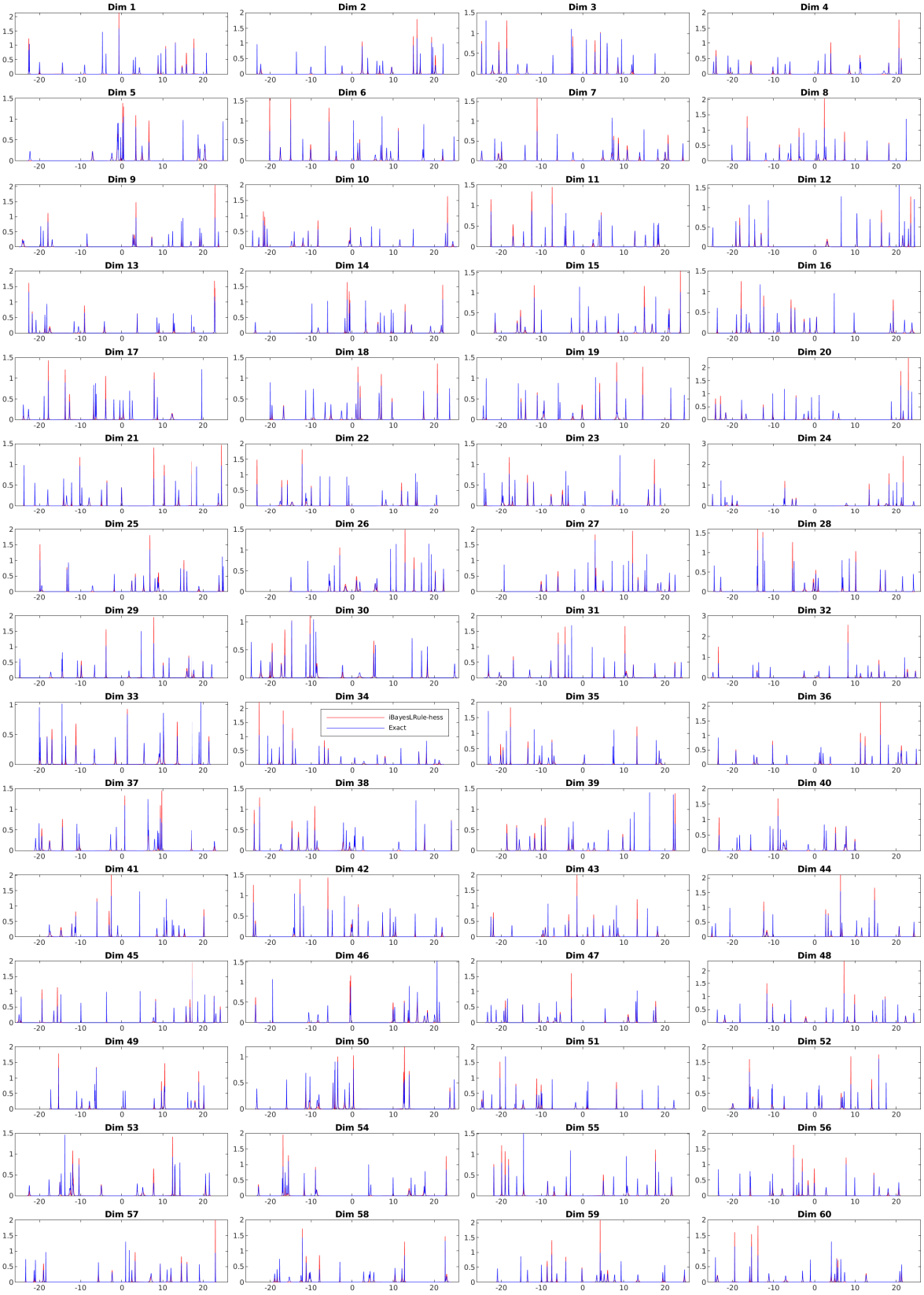


Figure 8. This is the first 60 marginal distributions obtained from a MOG approximation with $K = 60$ for a 300-dimensional mixture of Student's T distributions with 20 components. We describe the problem at Section 6.1, where the approximation is obtained by our method at the 50,000-th iteration.

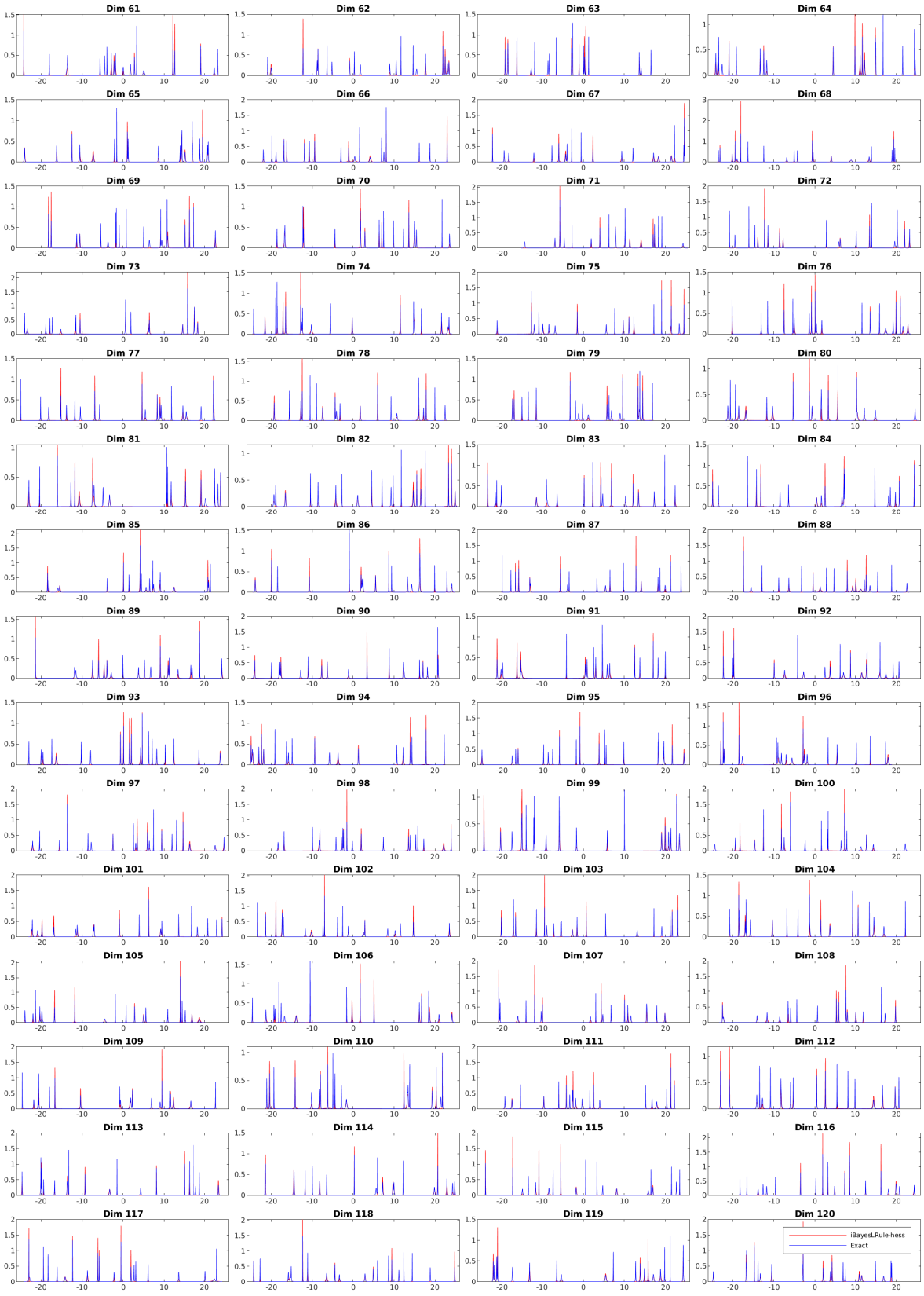


Figure 9. This is the second 60 marginal distributions obtained from a MOG approximation with $K = 60$ for a 300-dimensional mixture of Student's T distributions with 20 components. We describe the problem at Section 6.1, where the approximation is obtained by our method at the 50,000-th iteration.

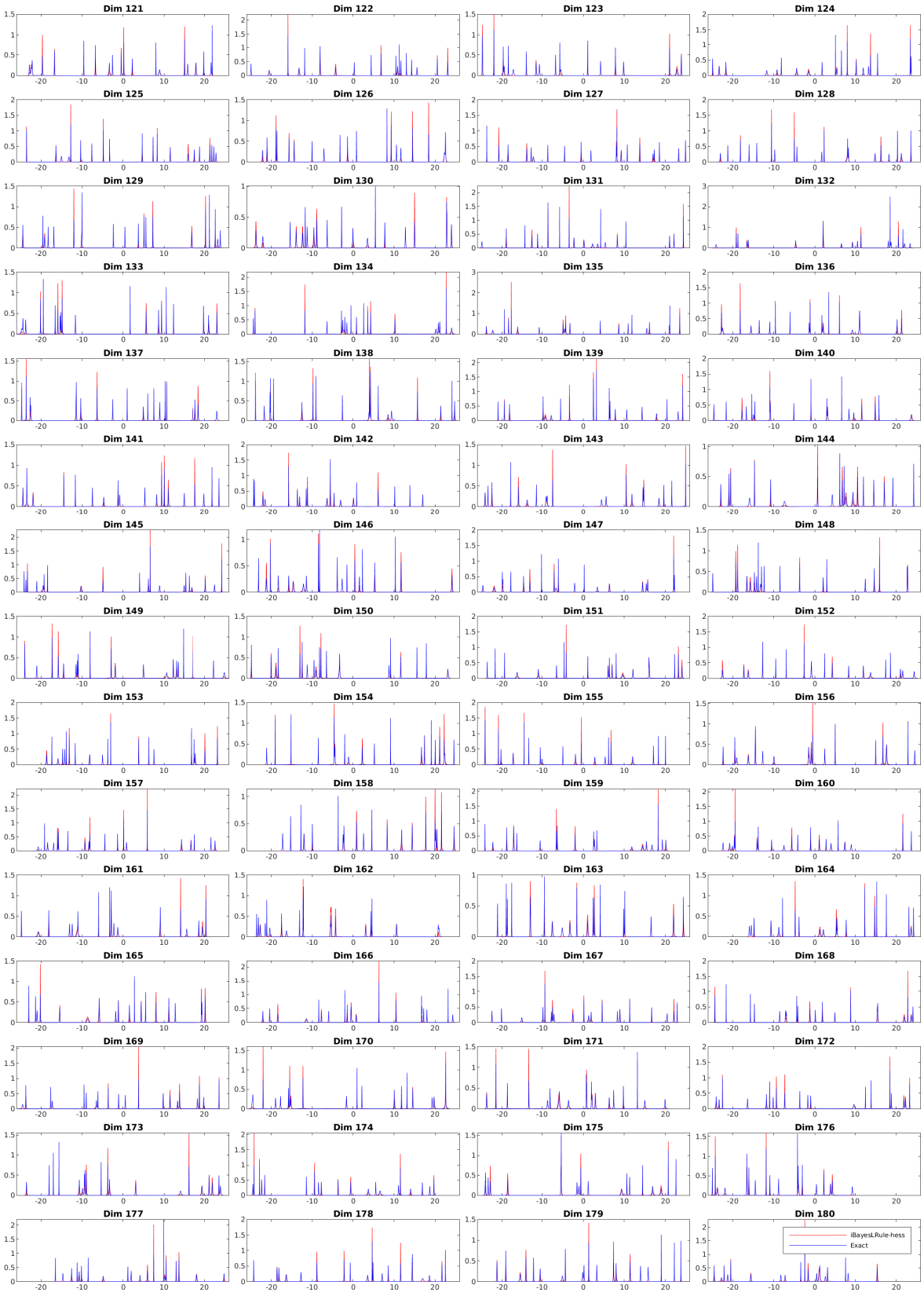


Figure 10. This is the third 60 marginal distributions obtained from a MOG approximation with $K = 60$ for a 300-dimensional mixture of Student's T distributions with 20 components. We describe the problem at Section 6.1, where the approximation is obtained by our method at the 50,000-th iteration.

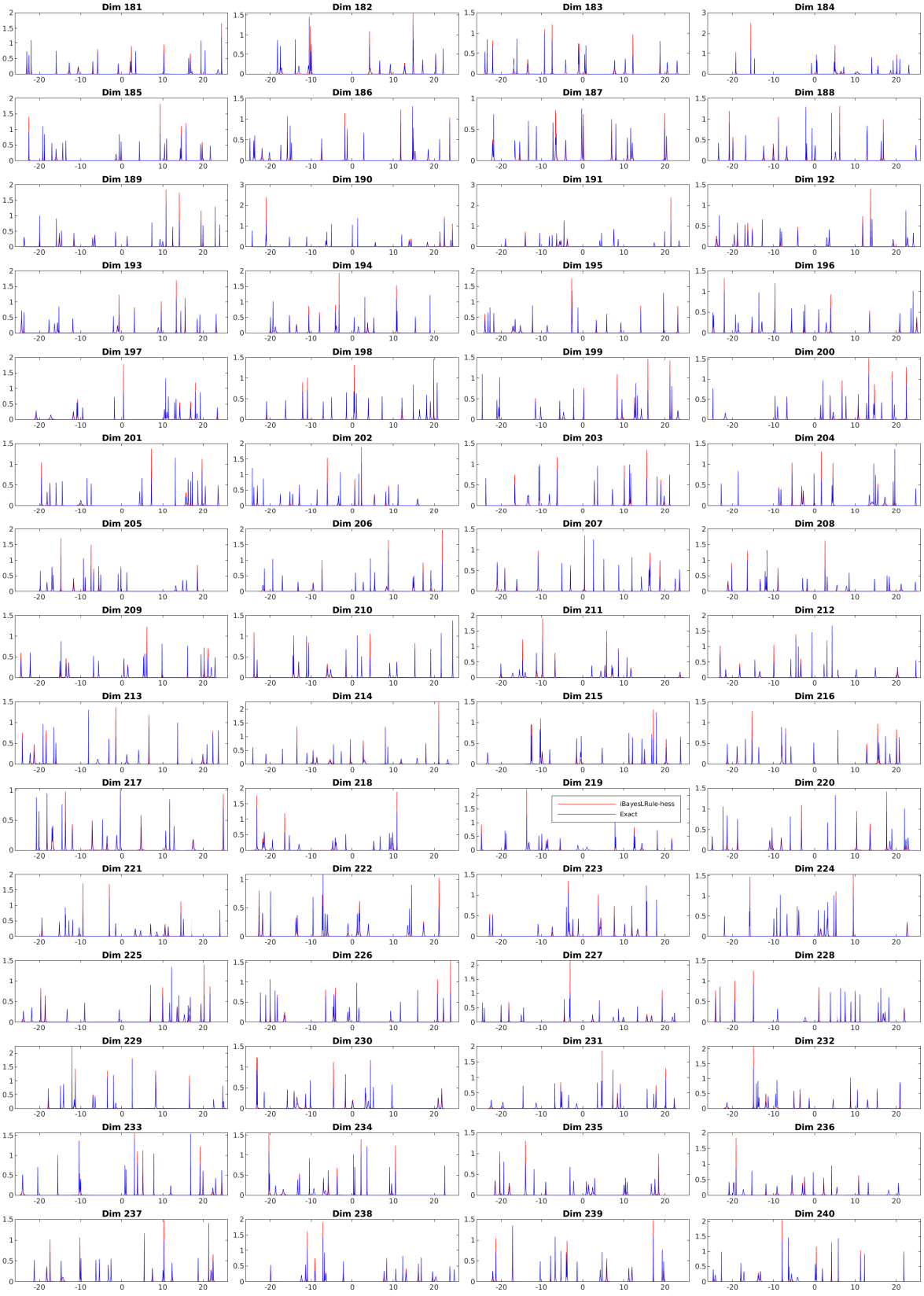


Figure 11. This is the fourth 60 marginal distributions obtained from a MOG approximation with $K = 60$ for a 300-dimensional mixture of Student's T distributions with 20 components. We describe the problem at Section 6.1, where the approximation is obtained by our method at the 50,000-th iteration.

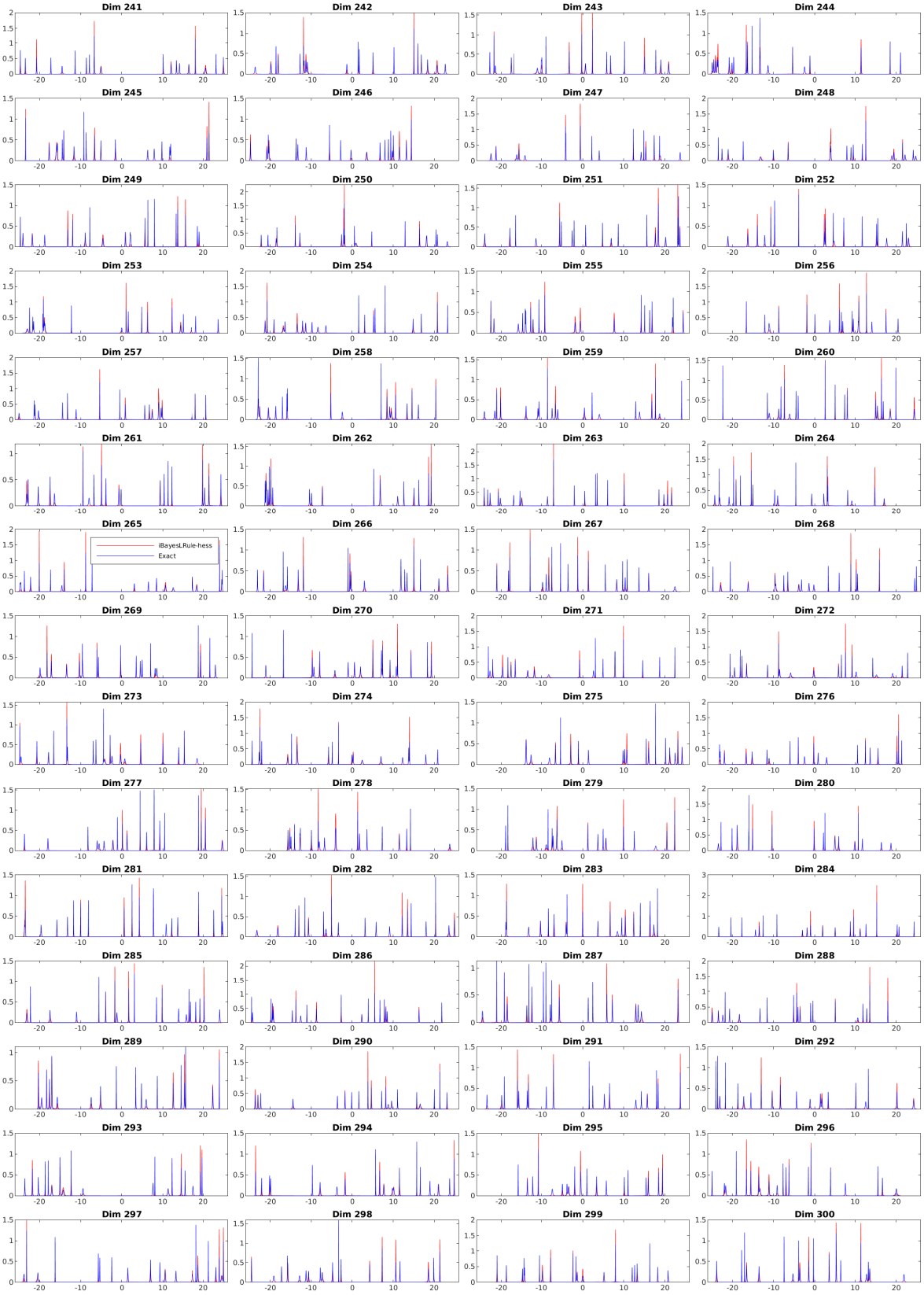


Figure 12. This is the last 60 marginal distributions obtained from a MOG approximation with $K = 60$ for a 300-dimensional mixture of Student's T distributions with 20 components. We describe the problem at Section 6.1, where the approximation is obtained by our method at the 50,000-th iteration.