A. Proof of Theorem 2

In this section, we provide a proof for Theorem 2:

For an MCEF representation given in Definition 3, the FIM $\mathbf{F}_{wz}(\boldsymbol{\lambda})$ given in (9) is positive-definite and invertible for all $\boldsymbol{\lambda} \in \Omega$. We prove this using a sequence of lemmas.

Lemma 1 *$\log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda})$ is twice differentiable with respect to $\boldsymbol{\lambda}$.*

Proof: From Definition 2, we see that the $\log q(\mathbf{z}, \mathbf{w})$ is differentiable when $A_w(\boldsymbol{\lambda}_w)$ and $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ are twice differentiable for each $\mathbf{w}$ sampled from $q(\mathbf{w})$. Since $q(\mathbf{w})$ is an EF, $A_w(\boldsymbol{\lambda}_w)$ is twice differentiable (Johansen, 1979). Similarly, since conditioned on $\mathbf{w}$, $q(\mathbf{z}|\mathbf{w})$ is also an EF, $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ is twice differentiable too. Therefore, the log of the joint distribution is twice differentiable. □

Lemma 2 *The FIM $\mathbf{F}_{wz}(\boldsymbol{\lambda})$ is block-diagonal with two blocks:*

$$\mathbf{F}_{wz}(\boldsymbol{\lambda}) = \begin{bmatrix} \mathbf{F}_z(\boldsymbol{\lambda}) & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_w(\boldsymbol{\lambda}_w) \end{bmatrix}, \tag{32}$$

where $\mathbf{F}_w(\boldsymbol{\lambda}_w)$ is the FIM of $q(\mathbf{w})$ and $\mathbf{F}_z(\boldsymbol{\lambda})$ is the expected of the FIM of $q(\mathbf{z}|\mathbf{w})$ where expectation is taken under $q(\mathbf{w})$ as shown below:

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

Proof: By Lemma 1, $\log q(\mathbf{w}, \mathbf{z}|\boldsymbol{\lambda})$ is twice differentiable, so the FIM is well defined. Below, we simplify the FIM to show that it has a block-diagonal structure. The first step below follows from the definition of the FIM. The second step is simply writing the FIM in a 2×2 block corresponding to $\boldsymbol{\lambda}_z$ and $\boldsymbol{\lambda}_w$. In the third step, we write the joint as the product of $q(\mathbf{z}|\mathbf{w})$ and $q(\mathbf{w})$. The fourth step is obtained since the two blocks are separable in $\boldsymbol{\lambda}_z$ and $\boldsymbol{\lambda}_w$. In the fifth step, we take the expectation inside which give us the desired result in the last step.

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

□

Lemma 3 *The first block of the FIM matrix $\mathbf{F}_z$ is equal to the derivative of the expectation parameter $\mathbf{m}_z(\boldsymbol{\lambda})$:*

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

Proof: We first show that the gradient of $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$ with respect to $\boldsymbol{\lambda}_z$ is equal to $\mathbb{E}_{q(z|w)}\left[\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})\right]$. By using the definition of $A_z(\boldsymbol{\lambda}_z, \mathbf{w})$, this is straightforward to show:

$$\begin{aligned}
\nabla_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w}) &= \nabla_{\lambda_z} \log \int h_z(\mathbf{z}, \mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle\right] d\mathbf{z} \\
&= \frac{\int \nabla_{\lambda_z} h_z(\mathbf{z}, \mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle\right] d\mathbf{z}}{\int h_z(\mathbf{z}, \mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle\right] d\mathbf{z}} \\
&= \frac{\int \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}) h_z(\mathbf{z}, \mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle\right] d\mathbf{z}}{\int h_z(\mathbf{z}, \mathbf{w}) \exp\left[\langle \boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w}), \boldsymbol{\lambda}_z \rangle\right] d\mathbf{z}} \qquad (33) \\
&= \mathbb{E}_{q(z|w)}\left[\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})\right] \qquad (34)
\end{aligned}$$

Using this, the expectation parameter $\mathbf{m}_z$ is simply the expected value of the gradient of the log-partition function.

$$\mathbf{m}_z = \mathbb{E}_{q(w)q(z|w)}\left[\boldsymbol{\phi}_z(\mathbf{z}, \mathbf{w})\right] = \mathbb{E}_{q(w)}\left[\nabla_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w})\right] \qquad (35)$$

Using this, it is easy to show the result by simply using the definition of the conditional EF, as shown below:

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

□

Lemma 4 *Let $\Omega_w \times \Omega_z$ be relatively relatively open. If the mapping $\mathbf{m}_w(\cdot) : \Omega_w \to \mathcal{M}_w$ is one-to-one, and, given every $\boldsymbol{\lambda}_w \in \Omega_w$, the conditional mapping $\mathbf{m}_z(\cdot, \boldsymbol{\lambda}_w) : \Omega_z \to \mathcal{M}_z$ is one-to-one, then $\mathbf{F}_{wz}(\boldsymbol{\lambda})$ is positive-definite in $\Omega_z \times \Omega_w$.*

Proof: When the mapping $\mathbf{m}_w$ is one-to-one, $q(\mathbf{w}|\boldsymbol{\lambda}_w)$ is a minimal EF, and given that Ω_w is relatively open, using the result discussed in Section 2, we conclude that the second block $\mathbf{F}_w(\boldsymbol{\lambda}_w)$ of $\mathbf{F}_{wz}(\boldsymbol{\lambda})$ given in (32) is positive definite and invertible for all Ω_z. Now we prove that the first block $\mathbf{F}_z(\boldsymbol{\lambda})$ is also positive definite.

The steps below establish the positive-*semi* definiteness first. The first step is simply the definition of the FIM, while the second step is obtained by using the fact that $\nabla \log f(\boldsymbol{\lambda}) = \nabla f(\boldsymbol{\lambda})/f(\boldsymbol{\lambda})$. The third step is obtained by using the chain-rule, and the fourth step simply uses the log-trick above to simply the second term. In the fifth step, we take the derivative out of the first term which cancels out $q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)$. The last step is straightfoward since the outer products are always nonnegative.

$$\begin{aligned}
\nabla^2_{\lambda_z} A_z(\boldsymbol{\lambda}_z, \mathbf{w}) &= -\mathbb{E}_{q(z|w)}\left[\nabla^2_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right], \\
&= -\mathbb{E}_{q(z|w)}\left[\nabla_{\lambda_z}\left(\frac{\nabla_{\lambda_z^T} q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)}{q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)}\right)\right], \\
&= -\mathbb{E}_{q(z|w)}\left[\frac{\nabla^2_{\lambda_z} q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)}{q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)} - \frac{\nabla_{\lambda_z} q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)}{q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)} \frac{\nabla_{\lambda_z^T} q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)}{q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)}\right] \\
&= \mathbb{E}_{q(z|w)}\left[-\frac{\nabla^2_{\lambda_z} q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)}{q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)}\right] + \mathbb{E}_{q(z|w)}\left[\nabla_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) \nabla_{\lambda_z^T} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right], \\
&= \int -\nabla^2_{\lambda_z} q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) d\mathbf{z} + \mathbb{E}_{q(z|w)}\left[\nabla_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) \nabla_{\lambda_z^T} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right], \\
&= \underbrace{-\nabla^2_{\lambda_z} \overbrace{\int q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) d\mathbf{z}}^{=1}}_{=0} + \mathbb{E}_{q(z|w)}\left[\nabla_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) \nabla_{\lambda_z^T} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right], \\
&= \mathbb{E}_{q(z|w)}\left[\nabla_{\lambda_z} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z) \nabla_{\lambda_z^T} \log q(\mathbf{z}|\mathbf{w}, \boldsymbol{\lambda}_z)\right] \succeq \mathbf{0}. \qquad (36)
\end{aligned}$$

Using Lemma 3 and (35), we see that FIM is positive semi-definite:

$$\mathbf{F}_z(\boldsymbol{\lambda}) = \nabla_{\lambda_z}\mathbf{m}_z^T = \nabla_{\lambda_z}\mathbb{E}_{q(w|\lambda_w)}\left[\nabla_{\lambda_z^T}A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right] = \mathbb{E}_{q(w)}\left[\nabla^2_{\lambda_z}A_z(\boldsymbol{\lambda}_z,\mathbf{w})\right] \succeq \mathbf{0}$$

Now, we prove the final claim that, for every $\boldsymbol{\lambda}_w \in \Omega_w$, if the conditional mapping $\mathbf{m}_z(\cdot,\boldsymbol{\lambda}_w)$ is one-to-one, then $\mathbf{F}_z(\boldsymbol{\lambda})$ is positive definite. We will prove this statement by contradiction. Suppose there exists $\boldsymbol{\lambda}$ such that $\mathbf{F}_z(\boldsymbol{\lambda})$ is positive semi-definite, since $\mathbf{F}_z(\boldsymbol{\lambda})$ is positive semi-definite, there exists a non-zero vector $\mathbf{a}$ such that $\mathbf{a}^T\mathbf{F}_z(\boldsymbol{\lambda})\mathbf{a} = 0$. Simplifying below, we show that this leads to a contradiction. The first and second step are obtained by simply plugging (36), while the third step is obtained by using the definition of $q(\mathbf{z}|\mathbf{w},\boldsymbol{\lambda}_z)$ and the fourth step is obtained by using (34). The last step is obtained by noting that the quantity is simply the variance of $\mathbf{a}^T\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w})$ conditioned on $\mathbf{w}$.

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

The expectation of a function positive quantity is equal to zero only when each function value is equal to zero, therefore for the above to be zeros, we need $\mathbf{a}^T\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w}) = 0$. However, as we show below, this is not possible since the representation $q(\mathbf{z}|\mathbf{w})$ is minimal conditioned on $\mathbf{w}$.

Since Ω_z is relatively open, there exists a small $\delta > 0$ to always be able to obtain a perturbed version $\boldsymbol{\lambda}'_z = \boldsymbol{\lambda}_z + \delta\mathbf{a}$, such that $\boldsymbol{\lambda}'_z \in \Omega_z$. Since the conditional mapping is one-to-one, $\mathbf{m}_z(\boldsymbol{\lambda}'_z,\boldsymbol{\lambda}_w) \neq \mathbf{m}_z(\boldsymbol{\lambda}_z,\boldsymbol{\lambda}_w)$. By using (35) and (33), when $\mathbf{a}^T\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w}) = 0$, we get a contradiction:

$$\begin{aligned}
\mathbf{m}_z(\boldsymbol{\lambda}'_z,\boldsymbol{\lambda}_w) &= \mathbb{E}_{q(w|\lambda_w)}\left[\nabla_{\lambda'_z}A_z(\boldsymbol{\lambda}'_z,\mathbf{w})\right] \\
&= \mathbb{E}_{q(w|\lambda_w)}\left[\frac{\int\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w})h_z(\mathbf{z},\mathbf{w})\exp\left[\langle\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w}),\boldsymbol{\lambda}'_z\rangle\right]d\mathbf{z}}{\int h_z(\mathbf{z},\mathbf{w})\exp\left[\langle\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w}),\boldsymbol{\lambda}'_z\rangle\right]d\mathbf{z}}\right] \qquad (37)\\
&= \mathbb{E}_{q(w|\lambda_w)}\left[\frac{\int\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w})h_z(\mathbf{z},\mathbf{w})\exp\left[\langle\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w}),\boldsymbol{\lambda}_z\rangle\right]d\mathbf{z}}{\int h_z(\mathbf{z},\mathbf{w})\exp\left[\langle\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w}),\boldsymbol{\lambda}_z\rangle\right]d\mathbf{z}}\right] \qquad (38)\\
&= \mathbf{m}_z(\boldsymbol{\lambda}_z,\boldsymbol{\lambda}_w)
\end{aligned}$$

where we can move from (37) to (38), since $\langle\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w}),\boldsymbol{\lambda}'_z\rangle = \langle\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w}),\boldsymbol{\lambda}_z + \delta\mathbf{a}\rangle = \langle\boldsymbol{\phi}_z(\mathbf{z},\mathbf{w}),\boldsymbol{\lambda}_z\rangle$, Due to the contradiction, $\mathbf{F}_z(\boldsymbol{\lambda})$ must be positive definite. This proves that both the blocks are positive definite and invertible.

□

Lemma 5 *The gradient with respect to $\boldsymbol{\lambda}$ can be expressed as the gradient with respect to $\mathbf{m}$:*

$$\nabla_\lambda\mathcal{L} = \left[\nabla_\lambda\mathbf{m}^T\right]\nabla_m\mathcal{L} = \left[\mathbf{F}_{wz}(\boldsymbol{\lambda})\right]\nabla_m\mathcal{L} \qquad (39)$$

Proof: Using Lemma 3, and chain rule, we can establish the results for $\boldsymbol{\lambda}_z$:

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

For $\boldsymbol{\lambda}_w$, this result holds trivially, which proves the statement. □

B. Finite Mixture of Gaussians

The finite mixture of EF distribution has the following conditional distribution $q(\mathbf{z}|w)$:

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

where we assume each component admits the same parametric form.

From the above expression and using the EF form for the multinomial distribution, we can write the sufficient statistics, natural parameters, and expectation parameters as shown below, where $\mathbf{m}_c := \mathbb{E}_{q(z|w=c)}\left[\boldsymbol{\phi}_z(\mathbf{z})\right]$ is the expectation parameter of a component $q(\mathbf{z}|w=c)$.

$$\begin{bmatrix} \mathbb{I}_1(w) \\ \mathbb{I}_2(w) \\ \vdots \\ \mathbb{I}_{K-1}(w) \\ \mathbb{I}_1(w)\boldsymbol{\phi}_z(\mathbf{z}) \\ \mathbb{I}_2(w)\boldsymbol{\phi}_z(\mathbf{z}) \\ \vdots \\ \mathbb{I}_K(w)\boldsymbol{\phi}_z(\mathbf{z}) \end{bmatrix} \quad \begin{bmatrix} \log(\pi_1/\pi_K) \\ \log(\pi_2/\pi_K) \\ \vdots \\ \log(\pi_{K-1}/\pi_K) \\ \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \vdots \\ \boldsymbol{\lambda}_K \end{bmatrix} \quad \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_{K-1} \\ \pi_1\mathbf{m}_1 \\ \pi_2\mathbf{m}_2 \\ \vdots \\ \pi_K\mathbf{m}_K \end{bmatrix}$$

From the last two vectors, we can see that the mapping between $\boldsymbol{\lambda}$ and $\mathbf{m}$ is going to be one-to-one, when each EF $q(\mathbf{z}|\boldsymbol{\lambda}_c)$ is minimal (which makes sure that mapping $\boldsymbol{\lambda}_c$ and $\mathbf{m}_c$ is one-to-one), and all $\boldsymbol{\lambda}_c$ are distinct.

B.1. The Model and ELBO

We consider the following model: $p(\mathcal{D}, \mathbf{z}) = \prod_{n=1}^{N} p(\mathcal{D}_n|\mathbf{z})p(\mathbf{z})$. We approximate the posterior by using the finite mixture of EFs whose marginal is denoted by $q(\mathbf{z})$ as given in (15). The variational lower bound is given by the following:

$$\mathcal{L}(\boldsymbol{\lambda}) = \left[\sum_{n=1}^{N} \left[\log p(\mathcal{D}_n|\mathbf{z})\right] + \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \right]$$
$$= \mathbb{E}_{q(z)}\left[-h(\mathbf{z})\right], \text{ where } h(\mathbf{z}) := -\left[\log \frac{p(\mathbf{z})}{q(\mathbf{z})} + \sum_{n} \log p(\mathcal{D}_n|\mathbf{z}) \right].$$

Note that the lower bound is defined with the marginal $q(\mathbf{z})$ and the variable w is not part of the model but only the variational approximation $q(\mathbf{z}, w)$.

B.2. Finite Mixture of Gaussians Approximation

We now give details about the NGD update for finite mixture of Gaussians. Note that the NGD update for $\boldsymbol{\lambda}_z$ and $\boldsymbol{\lambda}_w$ can be computed separately since the FIM is block-diagonal. We first derive the NGD update for each component $q(\mathbf{z}|w=c)$, and then give an update for $\boldsymbol{\lambda}_w$.

As shown in Table 1, the natural and expectation parameters of the c'th component is given as follows:

$$\boldsymbol{\Lambda}_c := -\tfrac{1}{2}\boldsymbol{\Sigma}_c^{-1} \qquad \mathbf{M}_c := \pi_c(\boldsymbol{\mu}_c\boldsymbol{\mu}_c^T + \boldsymbol{\Sigma}_c)$$
$$\boldsymbol{\lambda}_c := \boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c \qquad \mathbf{m}_c := \pi_c\boldsymbol{\mu}_c$$

The expectation parameters $\mathbf{m}_c$ and $\mathbf{M}_c$ are functions of $\pi_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c$ and its gradient can be obtained in terms of the gradient

with respect to these quantities by using the chain rule. The final expressions are shown below:

$$\nabla_{m_c}\mathcal{L} = \frac{1}{\pi_c}\left(\nabla_{\mu_c}\mathcal{L} - 2\left[\nabla_{\Sigma_c}\mathcal{L}\right]\boldsymbol{\mu}_c\right)$$
$$\nabla_{M_c}\mathcal{L} = \frac{1}{\pi_c}\left(\nabla_{\Sigma_c}\mathcal{L}\right)$$

We can compute gradients with respect to $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ by using the gradient and Hessian of $h(\mathbf{z})$ at a sample $\mathbf{z}$ from $q(\mathbf{z}, w)$. This can be done by using the Bonnet's and Price's theorems (Bonnet, 1964; Price, 1958; Opper & Archambeau, 2009; Rezende et al., 2014). (Staines & Barber, 2012) and Lin et al. (2019) discuss the conditions of the target function $h(\mathbf{z})$ when it comes to applying these theorems. Firstly, we define $\delta_c = q(\mathbf{z}|w=c)/q(\mathbf{z}) := \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)/\sum_{c'=1}^{K}\pi_{c'}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{c'}, \boldsymbol{\Sigma}_{c'})$. Using these theorems, we obtain the following stochastic-gradient estimations for the mixture of Gaussian case:

$$\nabla_{\mu_c}\mathcal{L}(\boldsymbol{\lambda}) = -\mathbb{E}_{q(z)}\left[q(w=c|\mathbf{z})\nabla_z h(\mathbf{z})\right] \quad \approx -\pi_c\delta_c\nabla_z h(\mathbf{z})$$
$$\nabla_{\Sigma_c}\mathcal{L}(\boldsymbol{\lambda}) = -\mathbb{E}_{q(z)}\left[q(w=c|\mathbf{z})\nabla_z^2 h(\mathbf{z})\right] \quad \approx -\frac{\pi_c\delta_c}{2}\nabla_z^2 h(\mathbf{z}).$$

where $q(w=c|\mathbf{z}) = \pi_c\delta_c$ and $\mathbf{z}$ is sampled from $q(\mathbf{z})$.

We can then plug these gradient estimations in the natural-gradient update for $\boldsymbol{\Lambda}_c$:

$$-\tfrac{1}{2}\left[\boldsymbol{\Sigma}_c^{(\text{new})}\right]^{-1} \leftarrow -\tfrac{1}{2}\boldsymbol{\Sigma}_c^{-1} + \beta\nabla_{M_c}\mathcal{L} \qquad \Rightarrow \left[\boldsymbol{\Sigma}_c^{(\text{new})}\right]^{-1} \leftarrow \boldsymbol{\Sigma}_c^{-1} + \beta\delta_c\nabla_z^2 h(\mathbf{z})$$

Similar for λ_c:

$$\begin{aligned}
\left[\boldsymbol{\Sigma}_c^{(\text{new})}\right]^{-1}\boldsymbol{\mu}_c^{(\text{new})} &\leftarrow \boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c + \beta\nabla_{m_c}\mathcal{L} \\
&\leftarrow \boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c + \beta\frac{1}{\pi_c}\left(\nabla_{\mu_c}\mathcal{L} - 2\left[\nabla_{\Sigma_c}\mathcal{L}\right]\boldsymbol{\mu}_c\right) \\
&\leftarrow \left[\boldsymbol{\Sigma}_c^{-1} - 2\frac{\beta}{\pi_c}\left[\nabla_{\Sigma_c}\mathcal{L}\right]\right]\boldsymbol{\mu}_c + \beta\frac{1}{\pi_c}\left(\nabla_{\mu_c}\mathcal{L}\right) \\
&\leftarrow \left[\boldsymbol{\Sigma}_c^{-1} + \beta\delta_c\left[\nabla_z^2 h(\mathbf{z})\right]\right]\boldsymbol{\mu}_c + \beta\frac{1}{\pi_c}\left(\nabla_{\mu_c}\mathcal{L}\right) \\
&\leftarrow \left[\boldsymbol{\Sigma}_c^{(\text{new})}\right]^{-1}\boldsymbol{\mu}_c + \beta\frac{1}{\pi_c}\left(\nabla_{\mu_c}\mathcal{L}\right)
\end{aligned}$$

This gives the following update (by using the stochastic gradients):

$$\boldsymbol{\mu}_c^{(\text{new})} \leftarrow \boldsymbol{\mu}_c - \beta\delta_c\boldsymbol{\Sigma}_c^{(\text{new})}\nabla_z h(\mathbf{z})$$

B.3. Natural Gradients for $q(w)$

Now, we give the update for $q(w|\boldsymbol{\lambda}_w)$. Its natural parameter and expectation parameter are

$$\boldsymbol{\lambda}_w = \left\{\log\frac{\pi_c}{\pi_K}\right\}_{c=1}^{K-1} \qquad \mathbf{m}_w = \left\{\mathbb{E}_{q(w)}\left[\mathbb{I}_c(w)\right]\right\}_{c=1}^{K-1} = \left\{\pi_c\right\}_{c=1}^{K-1}$$

To derive the gradients, we note that only $q(\mathbf{z})$ depends on $\boldsymbol{\pi}$ since the model does not contain this as a parameter. Therefore, we need the gradient of the variational approximation which can be written as follows:

$$\nabla_{\pi_c}q(\mathbf{z}) = \nabla_{\pi_c}\sum_{k=1}^{K}\pi_k q(\mathbf{z}|\boldsymbol{\lambda}_k) = q(\mathbf{z}|\boldsymbol{\lambda}_c) - q(\mathbf{z}|\boldsymbol{\lambda}_K).$$

The second term appears because the last π_K depends on π_c.

For the convenience of our derivation, we will separate the lower bound into terms that depend on $q(\mathbf{z})$ and the rest of the terms, as shown below:

$$\begin{aligned}
\nabla_{\pi_c}\mathcal{L}(\boldsymbol{\lambda}) &= \nabla_{\pi_c}\mathbb{E}_{q(z)}\left[\sum_{n=1}^{N}\log p(\mathcal{D}_n|\mathbf{z}) + \log p(\mathbf{z}) - \log q(\mathbf{z})\right] \\
&= \int \underbrace{\nabla_{\pi_c} q(\mathbf{z})}_{q(\boldsymbol{z}|\boldsymbol{\lambda}_c)-q(\boldsymbol{z}|\boldsymbol{\lambda}_K)}\left[\sum_{n=1}^{N}\log p(\mathcal{D}_n|\mathbf{z}) + \log p(\mathbf{z}) - \log q(\mathbf{z})\right] d\mathbf{z} - \underbrace{\int q(\mathbf{z})\nabla_{\pi_c}\log q(\mathbf{z}) d\mathbf{z}}_{0} \\
&= \int \left[q(\mathbf{z}|\boldsymbol{\lambda}_c) - q(\mathbf{z}|\boldsymbol{\lambda}_K)\right]\left[\sum_{n=1}^{N}\left(\log p(\mathcal{D}_n|\mathbf{z}) + \log\frac{p(\mathbf{z})}{q(\mathbf{z})}\right)\right] d\mathbf{z} \\
&= \int q(\mathbf{z})\left[\frac{q(\mathbf{z}|\boldsymbol{\lambda}_c)}{q(\mathbf{z})} - \frac{q(\mathbf{z}|\boldsymbol{\lambda}_K)}{q(\mathbf{z})}\right]\left[\sum_{n=1}^{N}\left(\log p(\mathcal{D}_n|\mathbf{z}) + \log\frac{p(\mathbf{z})}{q(\mathbf{z})}\right)\right] d\mathbf{z} \\
&= \mathbb{E}_{q(z)}\Big[\,(\delta_c - \delta_K)\,[\underbrace{\sum_{n=1}^{N}\left(\log p(\mathcal{D}_n|\mathbf{z}) + \log\frac{p(\mathbf{z})}{q(\mathbf{z})}\right)}_{-h(\boldsymbol{z})}]\Big] \\
&\approx -(\delta_c - \delta_K)h(\mathbf{z})
\end{aligned}$$

where $\mathbf{z}$ is a sample from $q(\mathbf{z})$.

Using this, we can perform the following NGD update:

$$\log(\pi_c/\pi_K) \leftarrow \log(\pi_c/\pi_K) - \beta(\delta_c - \delta_K)h(\mathbf{z})$$

B.4. Extension to Finite Mixture of EFs

The algorithm presented in Section 4.1 can be extended to handle generic minimal EF components. We now present a general gradient estimator to do so. The update of π_c remains unaltered, so we do not discuss them here. We only discuss how to update natural parameters $\boldsymbol{\lambda}_c$ of $q(\mathbf{z}|\boldsymbol{\lambda}_c)$.

The natural parameter and sufficient statistics are $\boldsymbol{\lambda}_c$ and $\mathbb{I}_c(w)\boldsymbol{\phi}_z(\mathbf{z})$ respectively. We wish to perform the following update: $\boldsymbol{\lambda}_c \leftarrow \boldsymbol{\lambda}_c + \beta_z \nabla_{m_c}\mathcal{L}(\boldsymbol{\lambda})$. In general, we can compute the gradient $\nabla m_c\mathcal{L}(\boldsymbol{\lambda})$ by computing the FIM of each component as shown below:

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

Both of these gradients can be obtained given $\nabla_{\lambda_z}\mathbf{z}$, where $\mathbf{z}$ is a sample from $q(\mathbf{z})$ as shown below.

$$\begin{aligned}
\nabla_{\lambda_c}\mathbb{E}_{q(w,z)}\left[\mathbb{I}_c(w)\boldsymbol{\phi}_z(\mathbf{z})\right] &= \nabla_{\lambda_c}\int \pi_c q(\mathbf{z}|w=c)\boldsymbol{\phi}_z(\mathbf{z})d\mathbf{z} = \int \pi_c q(\mathbf{z}|w=c)\nabla_z\left[\boldsymbol{\phi}_z(\mathbf{z})\right]\left[\nabla_{\lambda_c}\mathbf{z}\right]d\mathbf{z} \\
\nabla_{\lambda_c}\mathcal{L}(\boldsymbol{\lambda}) &= \int \pi_c\nabla_{\lambda_c}q(\mathbf{z}|w=c)\left[-h(\mathbf{z})\right]d\mathbf{z} + \underbrace{\int q(\mathbf{z})\left[\nabla_{\lambda_c}\log q(\mathbf{z})\right]d\mathbf{z}}_{0} \\
&= \int \pi_c q(\mathbf{z}|w=c)\left[\nabla_z\left(-h(\mathbf{z})\right)\right]\left[\nabla_{\lambda_c}\mathbf{z}\right]d\mathbf{z}
\end{aligned}$$

If we assume that $q(z|w)$ is an univariate continuous exponential family distribution, we can use the implicitly re-parameterization trick (Salimans & Knowles, 2013; Figurnov et al., 2018) to get:

$$\nabla_{\lambda_c}z = -\frac{\nabla_{\lambda_c}Q_c(z|\boldsymbol{\lambda}_c)}{q(z|w=c)},$$

where $Q_c(\cdot|\boldsymbol{\lambda}_c)$ is the cumulative distribution function (CDF) of $q(z|w=c)$. Therefore, we now can compute the required gradient as below:

$$\begin{aligned}\nabla_{\lambda_c}\mathcal{L}(\boldsymbol{\lambda}) =& \int \pi_c q(z|w=c)\left[-\nabla_z h(z)\right]\left[-\frac{\nabla_{\lambda_c}Q_c(z|\boldsymbol{\lambda}_c)}{q(z|w=c)}\right]dz \\ =&\mathbb{E}_{q(z)}\left[\pi_c\frac{q(z|w=c)}{q(z)}\left[-\nabla_z h(z)\right]\left[-\frac{\nabla_{\lambda_c}Q_c(z|\boldsymbol{\lambda}_c)}{q(z|w=c)}\right]\right] \\ =&\mathbb{E}_{q(z)}\left[\frac{\pi_c\nabla_{\lambda_c}Q_c(z|\boldsymbol{\lambda}_c)}{q(z)}\left[\nabla_z h(z)\right]\right],\end{aligned}$$

and

$$\begin{aligned}\nabla_{\lambda_c}\mathbb{E}_{q(w,z)}\left[\mathbb{I}_c(w)\boldsymbol{\phi}_z(z)\right] =& \int \pi_c q(\mathbf{z}|w=c)\nabla_z\left[\boldsymbol{\phi}_z(z)\right]\left[-\frac{\nabla_{\lambda_c}Q_c(z|\boldsymbol{\lambda}_c)}{q(z|w=c)}\right]d\mathbf{z} \\ =&\mathbb{E}_{q(z)}\left[\frac{-\pi_c\nabla_{\lambda_c}Q_c(z|\boldsymbol{\lambda}_c)}{q(z)}\nabla_z\left[\boldsymbol{\phi}_z(z)\right]\right].\end{aligned}$$

This is not the most efficient way to compute NGDs, however, for the specific cases (e.g., Gaussian, exponential distribution, inverse Gaussian), we can get simplifications whenever gradient with respect to the expectation parameters are easy to compute.

The Birnbaum-Saunders distribution, which is a finite mixture of inverse Gaussians, is presented in Appendix C. This example is different from examples given in this section since we allow each mixing component takes a different and tied parametric form.

B.5. Result for the Toy Example

See Figure 6

C. Birnbaum-Saunders Distribution

Firstly, we denote the inverse Gaussian distribution by $\text{InvGauss}(z;\mu,v) = \left(\frac{v}{2\pi z^3}\right)^{1/2}\exp\left[-\frac{vz}{2\mu^2}-\frac{v}{2z}+\frac{v}{\mu}\right]$, where $z>0$, $v>0$, and $\mu>0$. We consider the following mixture distribution.

$$\begin{aligned}q(w) &= p^{\mathbb{I}(w=0)}(1-p)^{\mathbb{I}(w=1)} \\ q(z|w) &= \mathbb{I}(w=0)\text{InvGauss}(z;\mu,v)+\mathbb{I}(w=1)\frac{z\text{InvGauss}(z;\mu,v)}{\mu},\end{aligned} \tag{40}$$

where $\frac{z}{\mu}\text{InvGauss}(z;\mu,v)$ is a normalized distribution since it is the distribution of $z=y^{-1}$ where y is distributed by $\text{InvGauss}(y;\mu^{-1},v/\mu^2)$.

As we can observe from Eq (40), each mixing component has a distinct parametric form and variational parameters are shared between the components, which is different from examples discussed in Appendix B. According to Desmond (1986); Jørgensen et al. (1991); Balakrishnan & Kundu (2019), the marginal distribution is known as the Birnbaum-Saunders distribution (Birnbaum & Saunders, 1969) shown as below, where $p=\frac{1}{2}$.

$$\begin{aligned}q(z|v,\mu) &= \sum_w q(w)q(z|w) \\ &= \tfrac{1}{2}\left\{\left(\frac{v}{2\pi z^3}\right)^{1/2}\exp\left[-\frac{vz}{2\mu^2}-\frac{v}{2z}+\frac{v}{\mu}\right]\left(1+\frac{z}{\mu}\right)\right\} \\ &= \frac{\sqrt{v}}{2\sqrt{2\pi}}\left[\frac{1}{z^{3/2}}+\frac{1}{\mu z^{1/2}}\right]\exp\left[-\frac{vz}{2\mu^2}-\frac{v}{2z}+\frac{v}{\mu}\right] \\ &= \frac{\sqrt{v}}{2\mu\sqrt{2\pi\mu}}\left[\left(\frac{\mu}{z}\right)^{1/2}+\left(\frac{\mu}{z}\right)^{3/2}\right]\exp\left\{-\frac{v(\frac{z}{\mu}+\frac{\mu}{z}-2)}{2\mu}\right\}\end{aligned}$$

Lemma 6 *The joint distribution of the Birnbaum-Saunders distribution given in* (40) *can be written in a conditional EF form.*

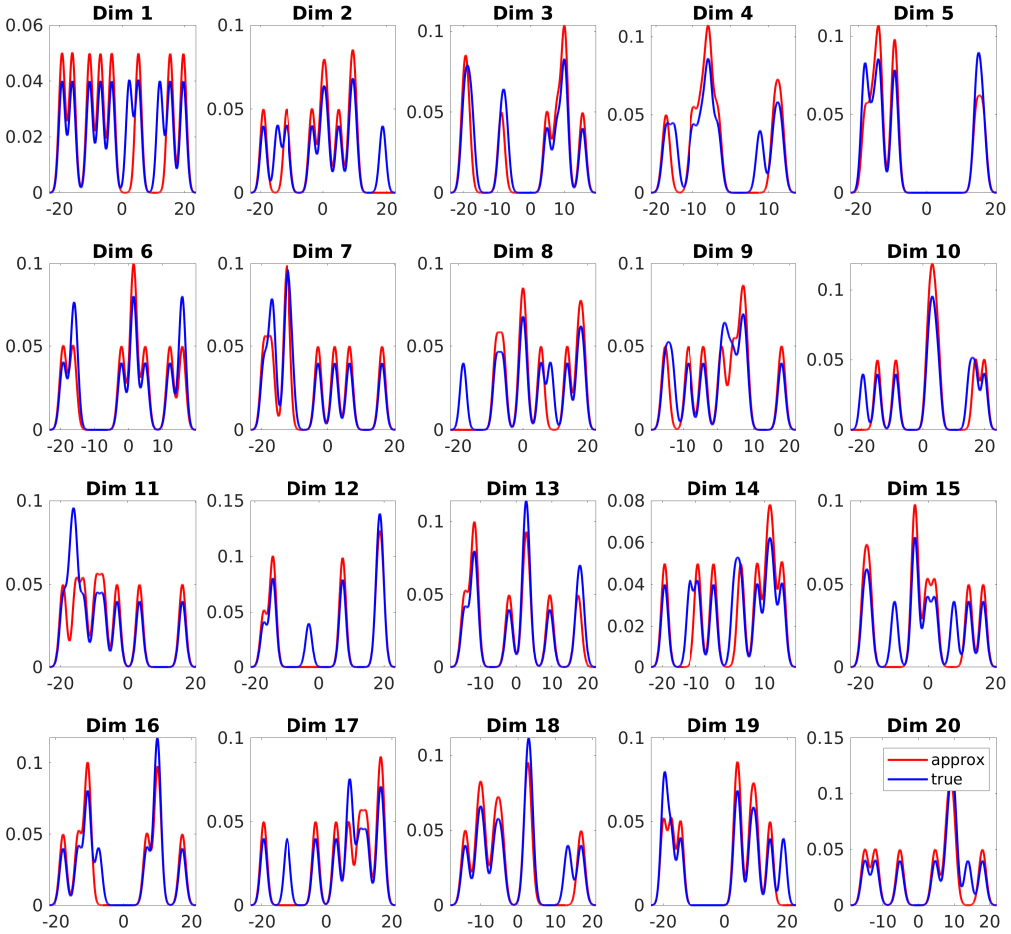

Figure 6. This is a complete version of the leftmost figure in Figure 2. The figure shows MOG approximation (with $K = 20$) to fit an MOG model with 10 components in a 20 dimensional problem.

Proof: It is obvious that $q(w)$ is Bernoulli distribution with $p = \frac{1}{2}$, which is an EF distribution. Now, we show that $q(z|w)$ is a conditional EF distribution as below.

$$
\begin{aligned}
q(z|w) &= \exp\left\{\mathbb{I}(w=0)\left[-\frac{vz}{2\mu^2} - \frac{v}{2z} + \frac{v}{\mu} + \tfrac{1}{2}\log\frac{v}{2\pi z^3}\right] + \mathbb{I}(w=1)\left[-\frac{vz}{2\mu^2} - \frac{v}{2z} + \frac{v}{\mu} + \tfrac{1}{2}\log\frac{v}{2\pi z} - \log\mu\right]\right\} \\
&= \exp\left\{-\frac{vz}{2\mu^2} - \frac{v}{2z} + \frac{v}{\mu} + \mathbb{I}(w=0)\left[\tfrac{1}{2}\log\frac{v}{2\pi z^3}\right] + \mathbb{I}(w=1)\left[\tfrac{1}{2}\log\frac{v}{2\pi z} - \log\mu\right]\right\} \\
&= \frac{1}{\sqrt{2\pi}} z^{-3\mathbb{I}(w=0)/2} z^{-\mathbb{I}(w=1)/2} \exp\left\{-\frac{vz}{2\mu^2} - \frac{v}{2z} + \frac{v}{\mu} + \tfrac{1}{2}\log(v) - \mathbb{I}(w=1)\log(\mu)\right\}
\end{aligned}
$$

The natural parameters and sufficient statistics are $\{-\frac{v}{2\mu^2}, -\frac{v}{2}\}$ and $\{z, \frac{1}{z}\}$ respectively. □

According to Balakrishnan & Kundu (2019), the expectation parameters are

$$m_1 = \mathbb{E}_{q(z)}\left[z\right] = \mu + \frac{\mu^2}{2v}$$
$$m_2 = \mathbb{E}_{q(z)}\left[z^{-1}\right] = \mu^{-1} + \frac{1}{2v}$$

The sufficient statistics, natural parameters, and expectation parameters are summarized below:

$$\begin{bmatrix} z \\ \frac{1}{z} \end{bmatrix} \quad \begin{bmatrix} -\frac{v}{2\mu^2} \\ -\frac{v}{2} \end{bmatrix} \quad \begin{bmatrix} \mu + \frac{\mu^2}{2v} \\ \mu^{-1} + \frac{1}{2v} \end{bmatrix}$$

Lemma 7 *The joint distribution given in* (40) *is a minimal conditional-EF.*

Proof: Since $\boldsymbol{\lambda}_w$ is known in this case, we only need to show there exists an one-to-one mapping between the natural parameter and the expectation parameter. Just by observing the parameters given above, we can see that there exists an one-to-one mapping between the natural parameter and $\{\mu, v\}$ since $\mu > 0$ and $v > 0$. Furthermore, we know that $m_1 m_2 > 1$ and $m_1 > 0$. We can show that there also exists an one-to-one mapping between $\{\mu, v\}$ and the expectation parameter by noticing that

$$\mu = \sqrt{m_1/m_2}$$
$$v = \frac{1}{2(m_2 - \sqrt{m_2/m_1})}$$

Since one-to-one mapping is transitive, we know that mapping between natural and expectation parameters is one-to-one. Hence proved. □

Note that we can use the implicitly re-parametrization trick to compute the gradient w.r.t. μ and v. Furthermore, the expectation parameters m_1 and m_2 are functions of μ, v and the gradients can be obtained in terms of the gradient with respect to μ and v by using the chain rule.

D. Studnet's t-distribution

Lemma 8 *The joint distribution $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\mathcal{IG}(w|a, a)$, where $\mathbf{z} \in \mathcal{R}^d$, is a curved exponential family distribution.*

Proof: The joint-distribution can be expressed as a four-parameter exponential form as shown below:

$$\begin{aligned}
\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\mathcal{IG}(w|a, a) &= \det\left(2\pi w\boldsymbol{\Sigma}\right)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\left(w\boldsymbol{\Sigma}\right)^{-1}(\mathbf{z}-\boldsymbol{\mu})\}\frac{a^a}{\Gamma(a)}w^{-a-1}\exp\{-\frac{a}{w}\} \\
&= (2\pi w)^{-d/2}\, w^{-1} \exp\{-\tfrac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\left(w\boldsymbol{\Sigma}\right)^{-1}(\mathbf{z}-\boldsymbol{\mu}) - \tfrac{1}{2}\log\det(\boldsymbol{\Sigma}) \\
&\qquad - \frac{a}{w} - a\log w - (\log\Gamma(a) - a\log a)\} \\
&= (2\pi w)^{-d/2}\, w^{-1} \exp\{\langle -\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}, w\mathbf{z}\mathbf{z}^T\rangle + \langle \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, w\mathbf{z}\rangle + \langle -\tfrac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, w^{-1}\rangle \\
&\qquad + \langle -a, w^{-1} + \log w\rangle - \left[\log\Gamma(a) - a\log a + \tfrac{1}{2}\log\det(\boldsymbol{\Sigma})\right]\} \\
&= (2\pi w)^{-d/2}\, w^{-1} \exp\{\langle \boldsymbol{\Lambda}_1, w\mathbf{z}\mathbf{z}^T\rangle + \langle \boldsymbol{\lambda}_2, w\mathbf{z}\rangle + \langle \lambda_3, w^{-1}\rangle \\
&\qquad + \langle \lambda_4, w^{-1} + \log w\rangle - \left[\log\Gamma(-\lambda_4) + \lambda_4\log(-\lambda_4) - \tfrac{1}{2}\log\det(-2\boldsymbol{\Lambda}_1)\right]\},
\end{aligned}$$

where the following are the natural parameters:

$$\boldsymbol{\Lambda}_1 := -\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}, \quad \boldsymbol{\lambda}_2 := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \quad \lambda_3 := -\tfrac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \quad \lambda_4 := -a$$

We can see that λ_3 is fully determined by $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\lambda}_2$, i.e.,

$$\lambda_3 = -\tfrac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = -\tfrac{1}{2}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)^T\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) = -\tfrac{1}{2}\left(\left(-2\boldsymbol{\Lambda}_1\right)^{-1}\boldsymbol{\lambda}_2\right)^T\boldsymbol{\lambda}_2$$

In a minimal representation we can specify all four parameters freely, but in this case we have less degree of freedom. Therefore, this is a curved EF representation.

□

Instead of using the above 4 parameter form, we can write the distribution in the conditional EF form given in Definition 2.

Lemma 9 *The joint distribution of Studnet's t-distribution given in* (17) *can be written in a conditional EF form.*

Proof: We can rewrite the conditional $q(\mathbf{z}|w)$ in an EF-form as follows:

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

The sufficient statistics $\boldsymbol{\phi}_z(\mathbf{z}, w) = \left\{w^{-1}\mathbf{z}, w^{-1}\mathbf{z}\mathbf{z}^T\right\}$. The natural parameter is $\boldsymbol{\lambda}_z = \left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\right\}$. Since $q(w)$ is a inverse Gamma distribution, which is a EF distribution as shown below, the joint $q(\mathbf{z}, w)$ is a conditional EF. The EF form of the inverse gamma distribution is shown below:

$$q(w|a, a) = w^{-1}\exp(-(\log(w) + \frac{1}{w})a - (\log\Gamma(a) - a\log(a)))$$

We can read the sufficient statistics $\boldsymbol{\phi}_w(w) = -\log(w) - \frac{1}{w}$ and the natural parameter $\lambda_w = a$ from this form. □

Using the fact that $\mathbb{E}_{q(w|a)}\left[1/w\right] = a/a = 1$, and $\mathbb{E}_{q(w|a)}\left[\log w\right] = \log a - \psi(a)$, we can derive the expectation parameter shown below:

$$\begin{aligned} \mathbf{m} &:= \mathbb{E}_{q(w|a)q(z|\mu,w\Sigma)}\left[w^{-1}\mathbf{z}\right] = \boldsymbol{\mu}, \\ \mathbf{M} &:= \mathbb{E}_{q(w|a)q(z|\mu,w\Sigma)}\left[w^{-1}\mathbf{z}\mathbf{z}^T\right] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}, \\ m_a &:= \mathbb{E}_{q(w|a)}\left[-\frac{1}{w} - \log(w)\right] = -1 - \log a + \psi(a) \end{aligned}$$

The sufficient statistics, natural parameters, and expectation parameters are summarized below:

$$\begin{bmatrix} -1/w - \log w \\ \mathbf{z}/w \\ \mathbf{z}\mathbf{z}^T/w \end{bmatrix} \quad \begin{bmatrix} a \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{bmatrix} \quad \begin{bmatrix} -1 - \log a + \psi(a) \\ \boldsymbol{\mu} \\ \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma} \end{bmatrix}$$

The following lemma shows that the Student's t-distribution is an MCEF obtained by establishing one-to-one mapping between natural and expectation parameters.

Lemma 10 *The joint distribution of Student's t-distribution shown in* (17) *is a minimal conditional-EF.*

Proof: The proof is rather simple. First we note that the expectation parameters for $q(\mathbf{z}|w)$ do not depend on $\boldsymbol{\lambda}_w := a$. In fact, the mapping between the last two natural and expectation parameter is one-to-one since they correspond to a Gaussian distribution which has a minimal representation.

The only thing remaining is to show that the mapping between a and $m_w(a) := -1 - \log a + \psi(a)$ is one-to-one. Since $\nabla_a m_w(a)$ is the Fisher information of $q(w)$, we can show this when $\nabla_a m_w(a) > 0$. The gradient $\nabla_a m_w(a) = \nabla_a\psi(a) - 1/a$. According to Eq. 1.4 in Batir (2005), we have $\nabla_\alpha\psi(\alpha) - 1/a - 1/(2a^2) > 0$ when $a > 0$. Therefore, $\nabla_a\psi(a) - 1/a > 0$ which establishes that the Fisher information is positive, therefore the distribution is a minimal EF. This completes the proof. □

D.1. Derivation of the NGD Update

Let's consider $q(w) = \mathcal{IG}(w|a, a)$ and $q(\mathbf{z}|w) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})$. We denote the log-likelihood for the n'th data point by $f_n(\mathbf{z}) := -\log p(\mathcal{D}_n|\mathbf{z})$ with a Student's t-prior on $\mathbf{z}$ expressed as a scale mixture of Gaussians $p(\mathbf{z}, w) =$

$\mathcal{IG}(w|a_0, a_0)\mathcal{N}(\mathbf{z}|\mathbf{0}, w\mathbf{I})$. We use the lower bound defined in the joint-distribution $p(\mathcal{D}, \mathbf{z}, w)$:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda}) &= \mathbb{E}_{q(z,w)}\left[\log p(\mathcal{D}, \mathbf{z}, w) - \log q(\mathcal{D}, \mathbf{z}, w)\right] \\ &= \mathbb{E}_{q(z,w)}\left[\sum_{n=1}^{N} \underbrace{\log p(\mathcal{D}_n|\mathbf{z})}_{:=-f_n(\mathbf{z})} + \log \frac{\mathcal{N}(\mathbf{z}|\mathbf{0}, w\mathbf{I})}{\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})} + \log \frac{\mathcal{IG}(w|a_0, a_0)}{\mathcal{IG}(w|a, a)}\right],\end{aligned}$$

Our goal is to compute the gradient of this ELBO with respect to the expectation parameters.

Since the expectation parameters $\mathbf{m}_z$ only depend on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we can write the gradient with respect to them using the chain rule (similar to the finite mixture of Gaussians case):

$$\begin{aligned}\nabla_m \mathcal{L}(\boldsymbol{\lambda}) &= \nabla_\mu \mathcal{L}(\boldsymbol{\lambda}) - 2\nabla_\Sigma \mathcal{L}(\boldsymbol{\lambda})\boldsymbol{\mu} \\ \nabla_M \mathcal{L}(\boldsymbol{\lambda}) &= \nabla_\Sigma \mathcal{L}(\boldsymbol{\lambda})\end{aligned}$$

These gradients of the lower bound can be obtained as follows:

$$\begin{aligned}\nabla_\mu \mathcal{L}(\boldsymbol{\lambda}) &= -\sum_{n=1}^{N} \nabla_\mu \mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right] - \boldsymbol{\mu} \\ \nabla_\Sigma \mathcal{L}(\boldsymbol{\lambda}) &= -\sum_{n=1}^{N} \nabla_\Sigma \mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right] - \tfrac{1}{2}\mathbf{I} + \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\end{aligned}$$

Plugging these in the natural-gradient updates (14), we get the following updates (we have simplified these in the same way as explained in Appendix B.2; more details in Khan et al. (2018)):

$$\begin{aligned}\boldsymbol{\Sigma}^{-1} &\leftarrow (1-\beta)\boldsymbol{\Sigma}^{-1} + 2\beta \sum_{n=1}^{N} \nabla_\Sigma \mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right] + \beta\mathbf{I} \\ \boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} - \beta \sum_{n=1}^{N} \boldsymbol{\Sigma}\left(\nabla_\mu \mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right] + \boldsymbol{\mu}\right)\end{aligned}$$

To compute the gradients, the reparametrization trick (Kingma & Welling, 2013) can be used. However, we can do better by using the extended Bonnet's and Price's theorems for Student's t-distribution (Lin et al., 2019). Assuming that $f_n(\mathbf{z})$ satisfies the assumptions needed for these two theorems, we can use the following stochastic-gradient approximations for the graidents:

$$\begin{aligned}\nabla_\mu \mathbb{E}_{q(w)\mathcal{N}(z|\mu,w\Sigma)}\left[f(\mathbf{z})\right] &= \mathbb{E}_{q(z)}\left[\nabla_z f(\mathbf{z})\right] \quad \approx \nabla_z f(\mathbf{z}), \\ \nabla_\Sigma \mathbb{E}_{q(w)\mathcal{N}(z|\mu,w\Sigma)}\left[f(\mathbf{z})\right] &= \tfrac{1}{2}\mathbb{E}_{q(z)}\left[u(\mathbf{z})\nabla_z^2 f(\mathbf{z})\right] \quad \approx \tfrac{1}{2}u(\mathbf{z})\nabla_z^2 f(\mathbf{z}), \\ &= \tfrac{1}{2}\mathbb{E}_{q(w,z)}\left[w\nabla_z^2 f(\mathbf{z})\right] \quad \approx \tfrac{1}{2}w\nabla_z^2 f(\mathbf{z}),\end{aligned}$$

where $\mathbf{z} \in \mathcal{R}^d$ is generated from $q(\mathbf{z})$, w is generated from $q(w)$, and

$$u(\mathbf{z}) := \frac{a + \frac{1}{2}\left(\mathbf{z} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{z} - \boldsymbol{\mu}\right)}{(a + d/2 - 1)}.$$

Using these gradient, we get the following update:

$$\begin{aligned}\boldsymbol{\Sigma}^{-1} &\leftarrow (1-\beta)\boldsymbol{\Sigma}^{-1} + \beta\left[u(\mathbf{z})\nabla_z^2 f_n(\mathbf{z}) + \mathbf{I}/N\right], \\ \boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} - \beta\boldsymbol{\Sigma}\left[\nabla_z f_n(\mathbf{z}) + \boldsymbol{\mu}/N\right].\end{aligned}$$

Now we derive the NGD update for a. Recall that the natural parameter is a and the expectation parameter is $m_a = -1 - \log a + \psi(a)$. The gradient of the lower bound can be expressed as

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

which gives us the following update:

$$a \leftarrow (1-\beta)a + \beta\left(a_0 - \sum_{n=1}^{N} \nabla_{m_a} \mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right]\right) \tag{41}$$

While the gradient with respect to the expectation parameter does not admit a closed-form expression, we can compute the gradient using the re-parametrization trick. According to (39), the gradient $\nabla_{m_a}\mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right]$ can be computed as

$$\nabla_{m_a}\mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right] = \left(\nabla_a m_a\right)^{-1}\nabla_a \mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right] = \left(\nabla_a \mathbb{E}_{q(w)}\left[\phi_w(w)\right]\right)^{-1}\nabla_a \mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right]$$

Note that $\nabla_a \mathbb{E}_{q(w)}\left[\phi_w(w)\right] = \nabla_a(m_a)$ has a closed-form expression. However we have found that using stochastic approximation for both the numerator and denominator works better. Salimans & Knowles (2013) show that such approximation reduces the variance but introduce a bit bias. We use the reparameterization trick for both terms. Since $q(w)$ is (implicitly) re-parameterizable (Salimans & Knowles, 2013; Figurnov et al., 2018) , the gradient can be computed as

$$\nabla_a \mathbb{E}_{q(w)}\left[\phi_w(w)\right] = -\int \mathcal{IG}(w|a,a)\left(\nabla_w\left[w^{-1}+\log w\right]\right)\left(\nabla_a w\right)dw \quad \approx (w^{-2}-w^{-1})\nabla_a w$$

where w is generated from $\mathcal{IG}(w|a,a)$. Similarly,

$$\begin{aligned}\nabla_a \mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right] &= \int\int q(w)\left(\nabla_w q(\mathbf{z}|w)\right)\left(\nabla_a w\right) f_n(\mathbf{z})dwd\mathbf{z} \\ &= \int\int \mathcal{IG}(w|a,a)\left(\nabla_w \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\right)\left(\nabla_a w\right) f_n(\mathbf{z})dwd\mathbf{z}\end{aligned}$$

For stochastic approximation, we generate w from $q(w)$ and let $\hat{\boldsymbol{\Sigma}} = w\boldsymbol{\Sigma}$. The above expression can be approximated as below.

$$\begin{aligned}\nabla_a \mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right] &\approx \int \left(\nabla_w \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w\boldsymbol{\Sigma})\right)\left(\nabla_a w\right) f_n(\mathbf{z})d\mathbf{z} \\ &= \int \mathrm{Tr}\left(\boldsymbol{\Sigma}\nabla_{\hat{\Sigma}}\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}})\right)\left(\nabla_a w\right) f_n(\mathbf{z})d\mathbf{z} \\ &= \nabla_a w \mathrm{Tr}\left(\boldsymbol{\Sigma}\nabla_{\hat{\Sigma}}\mathbb{E}_{\mathcal{N}(z|\mu,\hat{\Sigma})}\left[f_n(\mathbf{z})\right]\right) \\ &= \frac{\nabla_a w}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}\mathbb{E}_{\mathcal{N}(z|\mu,\hat{\Sigma})}\left[\nabla_z^2 f_n(\mathbf{z})\right]\right),\end{aligned}$$

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

We then use stochastic approximation to get the desired update.

$$\nabla_a \mathbb{E}_{q(z,w)}\left[f_n(\mathbf{z})\right] \approx \frac{\nabla_a w}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}\nabla_z^2 f_n(\mathbf{z})\right)$$

where $\mathbf{z}$ is generated from $q(\mathbf{z})$ and w is generated from $q(w)$.

Anther example is the symmetric normal inverse-Gaussian distribution, which can be found at Appendix H.

E. Multivariate Skew-Gaussian Distribution

We consider the following variational distribution .

$$q(\mathbf{z}, w) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + |w|\boldsymbol{\alpha}, \boldsymbol{\Sigma})\mathcal{N}(w|0,1) \tag{42}$$

The marginal distribution is known as the multivariate skew Gaussian distribution (Azzalini, 2005) as shown in the following lemma.

Lemma 11 *The marginal distribution $q(\mathbf{z})$ is $2\Phi\left(\frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\sqrt{1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right)\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}+\boldsymbol{\alpha}\boldsymbol{\alpha}^T)$, where $\Phi(\cdot)$ is CDF of the standard univariate Gaussian distribution.*

Proof: The marginal distribution $\mathbf{z}$ is

$$\begin{aligned} q(\mathbf{z}) =& 2\int_0^{+\infty} \mathcal{N}(w|0,1)\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}+w\boldsymbol{\alpha},\boldsymbol{\Sigma})dw \\ =& 2\int_0^{+\infty} \mathcal{N}(w|0,1)\mathcal{N}(\mathbf{z}-w\boldsymbol{\alpha}|\boldsymbol{\mu},\boldsymbol{\Sigma})dw \end{aligned}$$

By grouping terms related to w together and completing a Gaussian form for w, we obtain the following expression

$$\begin{aligned} q(\mathbf{z}) =& 2\int_0^{+\infty} \mathcal{N}(w|\frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}, \left(1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right)^{-1})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}+\boldsymbol{\alpha}\boldsymbol{\alpha}^T)dw \\ =& 2\Phi\left(\frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\sqrt{1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right)\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}+\boldsymbol{\alpha}\boldsymbol{\alpha}^T) \end{aligned}$$

where we move from the first step to the second step using the fact that

$$\begin{aligned} \int_0^{+\infty}\mathcal{N}(w|u,\sigma^2)dw &= \int_{-u/\sigma}^{+\infty}\mathcal{N}(w'|0,1)dw' \\ &= 1-\Phi(-u/\sigma) \\ &= \Phi(u/\sigma). \end{aligned}$$

□

Lemma 12 *The joint distribution of skew-Gaussian distribution given at Eq.* (42) *can be written in a conditional EF form.*

Proof: We first rewrite $q(\mathbf{z}|w)$ in a EF-form as follows:

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

The sufficient statistics $\boldsymbol{\phi}_z(\mathbf{z},w) = \left\{\mathbf{z}, |w|\mathbf{z}, \mathbf{z}\mathbf{z}^T\right\}$ and the natural parameter $\boldsymbol{\lambda}_z = \left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\right\}$ can be read from the form. Since $q(w)$ is a univariate Gaussian distribution with known parameters, which is a EF distribution. Therefore, the joint distribution $q(\mathbf{z},w)$ is a conditional EF. □

Let $c=\sqrt{\frac{2}{\pi}}$. Using the fact that $\mathbb{E}_{q(w|0,1)}\left[|w|\right]=c$, we can derive the expectation parameter shown below:

$$\begin{aligned} \mathbf{m} &:= \mathbb{E}_{\mathcal{N}(w|0,1)\mathcal{N}(z|\mu+|w|\alpha,\Sigma)}\left[\mathbf{z}\right] = \boldsymbol{\mu}+c\boldsymbol{\alpha}, \\ \mathbf{m}_\alpha &:= \mathbb{E}_{\mathcal{N}(w|0,1)\mathcal{N}(z|\mu+|w|\alpha,\Sigma)}\left[|w|\mathbf{z}\right] = c\boldsymbol{\mu}+\boldsymbol{\alpha}, \\ \mathbf{M} &:= \mathbb{E}_{\mathcal{N}(w|0,1)\mathcal{N}(z|\mu+|w|\alpha,\Sigma)}\left[\mathbf{z}\mathbf{z}^T\right] = \boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\alpha}\boldsymbol{\alpha}^T+c\left(\boldsymbol{\mu}\boldsymbol{\alpha}^T+\boldsymbol{\alpha}\boldsymbol{\mu}^T\right)+\boldsymbol{\Sigma} \end{aligned}$$

The sufficient statistics, natural parameters, and expectation parameters are summarized below:

$$\begin{bmatrix} \mathbf{z} \\ |w|\mathbf{z} \\ \mathbf{z}\mathbf{z}^T \end{bmatrix} \quad \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{bmatrix} \quad \begin{bmatrix} \boldsymbol{\mu}+c\boldsymbol{\alpha} \\ c\boldsymbol{\mu}+\boldsymbol{\alpha} \\ \boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\alpha}\boldsymbol{\alpha}^T+c\left(\boldsymbol{\mu}\boldsymbol{\alpha}^T+\boldsymbol{\alpha}\boldsymbol{\mu}^T\right)+\boldsymbol{\Sigma} \end{bmatrix}$$

The following lemma shows that the skew-Gaussian distribution is indeed a minimal conditional-EF.

Lemma 13 *Multivariate skew Gaussians is a minimal conditional-EF.*

Proof: Since $\boldsymbol{\lambda}_w$ is known in this case, we only need to show there exists an one-to-one mapping between the natural parameter and the expectation parameter. Just by observing the parameters given above, we can see that there exists an one-to-one mapping between the natural parameter and $\{\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\}$. We can show that there also exists an one-to-one mapping between $\{\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\}$ and the expectation parameter by noticing that

$$\boldsymbol{\mu} = \frac{\mathbf{m} - c\mathbf{m}_\alpha}{1 - c^2} \tag{43}$$

$$\boldsymbol{\alpha} = \frac{\mathbf{m}_\alpha - c\mathbf{m}}{1 - c^2} \tag{44}$$

$$\boldsymbol{\Sigma} = \mathbf{M} - \frac{\mathbf{m}\mathbf{m}^T + \mathbf{m}_\alpha\mathbf{m}_\alpha^T - c\left(\mathbf{m}_\alpha\mathbf{m}^T + \mathbf{m}\mathbf{m}_\alpha^T\right)}{1 - c^2} \tag{45}$$

Since one-to-one mapping is transitive, we know that mapping between natural and expectation parameters is one-to-one. Hence proved. □

E.1. Derivation of the NGD Update

Let's consider the variational approximation using the skew-Gaussian distribution $q(\mathbf{z}|\boldsymbol{\lambda})$. We consider the following model with a Gaussian-prior $\mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})$ on $\mathbf{z}$, where the log-likelihood for the n'th data point is denoted by $p(\mathcal{D}_n|\mathbf{z})$.

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

We use the lower bound defined in the following distribution $p(\mathcal{D}, \mathbf{z})$:

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

where $q(\mathbf{z}) = 2\Phi\left(\frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\sqrt{1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right)\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T)$ and recall that $\Phi(\cdot)$ denotes the CDF of the standard univariate normal distribution.

Our goal is to compute the gradient of this ELBO with respect to the expectation parameters.

E.2. Natural Gradient for $q(\mathbf{z}|w)$

We do not need to compute these gradients with respect to the expectation parameters directly. The gradients can be computed in terms of $\{\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\}$.

Using the mapping at (43)-(45) and the chain rule, we can express the following gradients with respect to the expectation parameters in terms of the gradients with respect to $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma}$.

$$\begin{aligned}
\nabla_m\mathcal{L} &= \frac{1}{1-c^2}\nabla_\mu\mathcal{L} - \frac{c}{1-c^2}\nabla_\alpha\mathcal{L} - 2\left(\nabla_\Sigma\mathcal{L}\right)\boldsymbol{\mu} \\
\nabla_{m_\alpha}\mathcal{L} &= \frac{1}{1-c^2}\nabla_\alpha\mathcal{L} - \frac{c}{1-c^2}\nabla_\mu\mathcal{L} - 2\left(\nabla_\Sigma\mathcal{L}\right)\boldsymbol{\alpha} \\
\nabla_M\mathcal{L} &= \nabla_\Sigma\mathcal{L}
\end{aligned}$$

By plugging the gradients into the update in (14) and then re-expressing the update in terms of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Sigma}$ (we have simplified these in the same way as explained in Appendix B.2), we obtain the natural gradient update in terms of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma}^{-1} \leftarrow \boldsymbol{\Sigma}^{-1} - 2\beta\nabla_\Sigma\mathcal{L} \tag{46}$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \beta\boldsymbol{\Sigma}\left(\frac{1}{1-c^2}\nabla_\mu\mathcal{L} - \frac{c}{1-c^2}\nabla_\alpha\mathcal{L}\right) \tag{47}$$

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + \beta\boldsymbol{\Sigma}\left(\frac{1}{1-c^2}\nabla_\alpha\mathcal{L} - \frac{c}{1-c^2}\nabla_\mu\mathcal{L}\right) \tag{48}$$

Recall that the lower bound is

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

For the prior term, there is a closed-form expression for gradient computation.

$$\mathbb{E}_{q(z)}\left[\log \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})\right] = -\frac{d}{2}\log(2\pi) + \frac{d\delta}{2} - \frac{\delta}{2}\left(\boldsymbol{\alpha}^T\boldsymbol{\alpha} + 2c\boldsymbol{\mu}^T\boldsymbol{\alpha} + \mathrm{Tr}\left(\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^T\boldsymbol{\mu}\right)$$

It is easy to show that the gradients about the prior term are

$$\begin{aligned}
\mathbf{g}_{\mu}^{\text{prior}} &= -\delta\left(\boldsymbol{\mu} + c\boldsymbol{\alpha}\right) \\
\mathbf{g}_{\alpha}^{\text{prior}} &= -\delta\left(\boldsymbol{\alpha} + c\boldsymbol{\mu}\right) \\
\mathbf{g}_{\Sigma}^{\text{prior}} &= -\frac{\delta}{2}\mathbf{I}
\end{aligned}$$

For the entropy term, by Contreras-Reyes & Arellano-Valle (2012); Arellano-Valle et al. (2013), it can be expressed as follows.

$$\mathbb{E}_{q(z)}\left[-\log q(\mathbf{z}|\boldsymbol{\lambda})\right] = \frac{d}{2}(\log(2\pi)+1) + \tfrac{1}{2}\log\left|\boldsymbol{\Sigma} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T\right| - 2\mathbb{E}_{\mathcal{N}(z_3|0,\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})}\Big[\Phi\left(z_2\right)\log\left(\Phi\left(z_2\right)\right)\Big] - \log(2) \tag{49}$$

We can use the re-parametrization trick to compute the gradients about the entropy term. However, we can found out the exact gradients usually works better. Using the expression in (49), the gradients of the entropy term are given as follows:

$$\mathbf{g}_{\mu}^{\text{entropy}} = 0 \tag{50}$$

$$\mathbf{g}_{\alpha}^{\text{entropy}} = \left(\boldsymbol{\Sigma} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T\right)^{-1}\boldsymbol{\alpha} - \mathbb{E}_{\mathcal{N}\left(z_3|0,\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}/\left(1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right)\right)}\left[\frac{\log\left(\Phi\left(z_3\right)\right)}{\sqrt{2\pi\left(1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right)}}\frac{2z_3\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right] \tag{51}$$

$$\mathbf{g}_{\Sigma}^{\text{entropy}} = \tfrac{1}{2}\left(\boldsymbol{\Sigma} + \boldsymbol{\alpha}\boldsymbol{\alpha}^T\right)^{-1} + \mathbb{E}_{\mathcal{N}\left(z_3|0,\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}/\left(1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right)\right)}\left[\frac{\log\left(\Phi\left(z_3\right)\right)}{\sqrt{2\pi\left(1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right)}}\frac{z_3\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right] \tag{52}$$

where the expectations involve 1d integrations, which can be computed by Gauss-Hermite quadrature.

The remaining thing is to compute the gradients about $\mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right]$. To compute the gradients, the reparametrization trick can be used. However, we can do better by using the extended Bonnet's and Price's theorems for skew-Gaussian distribution (Lin et al., 2019). Assuming that $f_n(\mathbf{z})$ satisfies the assumptions needed for these two theorems, we obtain the following gradient expression:

$$\begin{aligned}
\mathbf{g}_1^n &:= \nabla_{\mu}\mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right] = \mathbb{E}_{q(z)}\left[\nabla_z f_n(\mathbf{z})\right] \approx \nabla_z f_n(\mathbf{z}) \\
\mathbf{g}_2^n &:= \nabla_{\alpha}\mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right] \\
&= \mathbb{E}_{q(z)}\left[u(\mathbf{z})\nabla_z f_n(\mathbf{z})\right] + v\mathbb{E}_{\mathcal{N}(\hat{z}|\mu,\Sigma)}\left[\nabla_{\hat{z}} f_n(\hat{\mathbf{z}})\right] \approx u(\mathbf{z})\nabla_z f_n(\mathbf{z}) + v\nabla_{\hat{z}} f_n(\hat{\mathbf{z}}) \\
&= \mathbb{E}_{q(w,z)}\left[|w|\nabla_z f_n(\mathbf{z})\right] \approx |w|\nabla_z f_n(\mathbf{z}) \\
\mathbf{g}_3^n &:= 2\nabla_{\Sigma}\mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right] = \mathbb{E}_{q(z)}\left[\nabla_z^2 f_n(\mathbf{z})\right] \approx \nabla_z^2 f_n(\mathbf{z})
\end{aligned}$$

where $v = \frac{c}{(1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})}$, $u(\mathbf{z}) := \frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}$, and $w \sim \mathcal{N}(w|0,1)$, $\hat{\mathbf{z}} \sim \mathcal{N}(\hat{\mathbf{z}}|\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\mathbf{z} = \hat{\mathbf{z}} + |w|\boldsymbol{\alpha}$.

Putting together, we can express the gradients in the following form, which will be used for deriving the extended variational

Adam update.

$$\nabla_\mu \mathcal{L}(\boldsymbol{\lambda}) = -\sum_{n=1}^{N} \left(\mathbf{g}_1^n \underbrace{-\mathbf{g}_\mu^{\text{entropy}}/N}_{0} \right) \underbrace{-\delta(\boldsymbol{\mu} + c\boldsymbol{\alpha})}_{\mathbf{g}_\mu^{\text{prior}}}$$

$$= -\sum_{n=1}^{N} \underbrace{\mathbf{g}_1^n}_{:=\mathbf{g}_\mu^n} - \delta(\boldsymbol{\mu} + c\boldsymbol{\alpha}) \tag{53}$$

$$\nabla_\alpha \mathcal{L}(\boldsymbol{\lambda}) = -\sum_{n=1}^{N} \left(\underbrace{\mathbf{g}_2^n - \mathbf{g}_\alpha^{\text{entropy}}/N}_{:=\mathbf{g}_\alpha^n} \right) \underbrace{-\delta(\boldsymbol{\alpha} + c\boldsymbol{\mu})}_{\mathbf{g}_\alpha^{\text{prior}}} \tag{54}$$

$$\nabla_\Sigma \mathcal{L}(\boldsymbol{\lambda}) = -\tfrac{1}{2}\sum_{n=1}^{N} \left(\mathbf{g}_3^n - 2\mathbf{g}_\Sigma^{\text{entropy}}/N\right) \underbrace{-\frac{\delta}{2}\mathbf{I}}_{\mathbf{g}_\Sigma^{\text{prior}}} + \tfrac{1}{2}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1})$$

$$= -\tfrac{1}{2}\sum_{n=1}^{N} \left(\underbrace{\mathbf{g}_3^n - 2\mathbf{g}_\Sigma^{\text{entropy}}/N + \boldsymbol{\Sigma}^{-1}/N}_{:=\mathbf{g}_S^n} \right) - \frac{\delta}{2}\mathbf{I} + \tfrac{1}{2}\boldsymbol{\Sigma}^{-1} \tag{55}$$

For stochastic approximation, we can sub-sampling a data point n and use MC samples to approximate $\mathbf{g}_1^n$, $\mathbf{g}_2^n$, and $\mathbf{g}_3^n$. Plugging these stochastic gradients into (46)-(48), we obtain the NGD update:

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1} &\leftarrow (1-\beta)\boldsymbol{\Sigma}^{-1} + \beta(\delta\mathbf{I} + N\mathbf{g}_S^n) \\
\boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} - \beta\boldsymbol{\Sigma}\left(\frac{N}{1-c^2}(\mathbf{g}_\mu^n - c\mathbf{g}_\alpha^n) + \delta\boldsymbol{\mu}\right) \\
\boldsymbol{\alpha} &\leftarrow \boldsymbol{\alpha} - \beta\boldsymbol{\Sigma}\left(\frac{N}{1-c^2}(\mathbf{g}_\alpha^n - c\mathbf{g}_\mu^n) + \delta\boldsymbol{\alpha}\right)
\end{aligned}$$

F. Multivariate Exponentially Modified Gaussian Distribution

We consider the following mixture distribution.

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

The marginal distribution is a multivariate extension of the exponentially modified Gaussian distribution (Grushka, 1972) and the Gaussian minus exponential distribution (Carr & Madan, 2009) due to the following lemma.

Lemma 14 *The marginal distribution $q(\mathbf{z})$ ($\boldsymbol{\alpha} \neq \mathbf{0}$) is*

$$q(\mathbf{z}) = \frac{\sqrt{2\pi}\det\left(2\pi\boldsymbol{\Sigma}\right)^{-\frac{1}{2}}}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\Phi\left(\frac{\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} - 1}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right)\exp\left\{\tfrac{1}{2}\left[\frac{\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} - 1\right)^2}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}} - \left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\right]\right\},$$

where $\Phi(\cdot)$ is the CDF of the standard univariate Gaussian distribution.

In the univariate case, the marginal distribution becomes the exponentially modified Gaussian distribution when $\alpha > 0$ and the Gaussian minus exponential distribution when $\alpha < 0$.

Proof: The marginal distribution $\mathbf{z}$ is

$$q(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}) = \int_0^{+\infty} \text{Exp}(w|0,1)\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}+w\boldsymbol{\alpha},\boldsymbol{\Sigma})dw$$

By grouping terms related to w together and completing a Gaussian form for w, we obtain the following expression.

$$= \frac{\det\left(2\pi\boldsymbol{\Sigma}\right)^{-\frac{1}{2}}}{\sqrt{\frac{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{2\pi}}} \int_0^{+\infty} \mathcal{N}(w|\frac{\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}, \frac{1}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}) \exp\left\{\frac{1}{2}\left[\frac{\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1\right)^2}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}} - \left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\right]\right\} dw \tag{56}$$

$$= \frac{\sqrt{2\pi}\det\left(2\pi\boldsymbol{\Sigma}\right)^{-\frac{1}{2}}}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}} \Phi\left(\frac{\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right) \exp\left\{\frac{1}{2}\left[\frac{\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1\right)^2}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}} - \left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\right]\right\} \tag{57}$$

where we move from Eq. (56) to Eq. (57) using the fact that $\int_0^{+\infty}\mathcal{N}(w|u,\sigma^2)dw = \Phi(u/\sigma)$. □

Similar to the skew-Gaussian case, in this example, the sufficient statistics $\boldsymbol{\phi}_z(\mathbf{z},w) = \left\{\mathbf{z}, w\mathbf{z}, \mathbf{z}\mathbf{z}^T\right\}$ and the natural parameter $\boldsymbol{\lambda}_z = \left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\right\}$ can be read from $q(\mathbf{z}|w) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}+w\boldsymbol{\alpha},\boldsymbol{\Sigma})$. The joint distribution $q(\mathbf{z},w)$ is a conditional EF since $q(w)$ is an exponential distribution with known parameters, which is an EF distribution. Likewise, we can show that the joint distribution is also a minimal conditional EF. We can derive the expectation parameter shown below:

$$\begin{aligned}
\mathbf{m} &:= \mathbb{E}_{\text{Exp}(w|1)\mathcal{N}(z|\mu+w\alpha,\Sigma)}\left[\mathbf{z}\right] = \boldsymbol{\mu}+\boldsymbol{\alpha},\\
\mathbf{m}_\alpha &:= \mathbb{E}_{\text{Exp}(w|1)\mathcal{N}(z|\mu+w\alpha,\Sigma)}\left[w\mathbf{z}\right] = \boldsymbol{\mu}+2\boldsymbol{\alpha},\\
\mathbf{M} &:= \mathbb{E}_{\text{Exp}(w|1)\mathcal{N}(z|\mu+w\alpha,\Sigma)}\left[\mathbf{z}\mathbf{z}^T\right] = \boldsymbol{\mu}\boldsymbol{\mu}^T + 2\boldsymbol{\alpha}\boldsymbol{\alpha}^T + \left(\boldsymbol{\mu}\boldsymbol{\alpha}^T + \boldsymbol{\alpha}\boldsymbol{\mu}^T\right) + \boldsymbol{\Sigma}
\end{aligned}$$

The sufficient statistics, natural parameters, and expectation parameters are summarized below:

$$\begin{bmatrix} \mathbf{z} \\ w\mathbf{z} \\ \mathbf{z}\mathbf{z}^T \end{bmatrix} \quad \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{bmatrix} \quad \begin{bmatrix} \boldsymbol{\mu}+\boldsymbol{\alpha} \\ \boldsymbol{\mu}+2\boldsymbol{\alpha} \\ \boldsymbol{\mu}\boldsymbol{\mu}^T + 2\boldsymbol{\alpha}\boldsymbol{\alpha}^T + \left(\boldsymbol{\mu}\boldsymbol{\alpha}^T + \boldsymbol{\alpha}\boldsymbol{\mu}^T\right) + \boldsymbol{\Sigma} \end{bmatrix}$$

F.1. Derivation of the NGD Update

As shown in Appendix E.1, we consider the variational approximation using the exponentially modified Gaussian distribution $q(\mathbf{z}|\boldsymbol{\lambda})$. We consider the same model as Appendix E.1 with a Gaussian-prior $\mathcal{N}(\mathbf{z}|\mathbf{0},\delta^{-1}\mathbf{I})$ on $\mathbf{z}$. The lower bound is defined as below:

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

where $q(\mathbf{z}) = \frac{\sqrt{2\pi}\det\left(2\pi\boldsymbol{\Sigma}\right)^{-\frac{1}{2}}}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}} \Phi\left(\frac{\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right) \exp\left\{\frac{1}{2}\left[\frac{\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1\right)^2}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}} - \left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)\right]\right\}$ and recall that $\Phi(\cdot)$ denotes the CDF of the standard univariate normal distribution.

Our goal is to compute the gradient of this ELBO with respect to the expectation parameters.

F.2. Natural Gradient for $q(\mathbf{z}|w)$

We do not need to compute these gradients with respect to the expectation parameters directly. The gradients can be computed in terms of $\{\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma}\}$.

Similarly, by the chain rule, we can express the following gradients with respect to the expectation parameters in terms of the gradients with respect to $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma}$.

$$\begin{aligned}\nabla_m \mathcal{L} &= 2\nabla_\mu \mathcal{L} - \nabla_\alpha \mathcal{L} - 2\left(\nabla_\Sigma \mathcal{L}\right)\boldsymbol{\mu} \\ \nabla_{m_\alpha} \mathcal{L} &= \nabla_\alpha \mathcal{L} - \nabla_\mu \mathcal{L} - 2\left(\nabla_\Sigma \mathcal{L}\right)\boldsymbol{\alpha} \\ \nabla_M \mathcal{L} &= \nabla_\Sigma \mathcal{L}\end{aligned}$$

By plugging the gradients into the update in (14) and then re-expressing the update in terms of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Sigma}$, we obtain the natural gradient update in terms of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma}^{-1} \leftarrow \boldsymbol{\Sigma}^{-1} - 2\beta\nabla_\Sigma \mathcal{L} \tag{58}$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \beta\boldsymbol{\Sigma}\left(2\nabla_\mu \mathcal{L} - \nabla_\alpha \mathcal{L}\right) \tag{59}$$

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + \beta\boldsymbol{\Sigma}\left(\nabla_\alpha \mathcal{L} - \nabla_\mu \mathcal{L}\right) \tag{60}$$

Recall that the lower bound is

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

For the prior term, there is a closed-form expression for gradient computation.

$$\mathbb{E}_{q(z)}\left[\log \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})\right] = -\frac{d}{2}\log(2\pi) + \frac{d\delta}{2} - \frac{\delta}{2}\left(2\boldsymbol{\alpha}^T\boldsymbol{\alpha} + 2\boldsymbol{\mu}^T\boldsymbol{\alpha} + \mathrm{Tr}\left(\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^T\boldsymbol{\mu}\right)$$

It is easy to show that the gradients about the prior term are

$$\begin{aligned}\mathbf{g}_\mu^{\text{prior}} &= -\delta\left(\boldsymbol{\mu} + \boldsymbol{\alpha}\right) \\ \mathbf{g}_\alpha^{\text{prior}} &= -\delta\left(2\boldsymbol{\alpha} + \boldsymbol{\mu}\right) \\ \mathbf{g}_\Sigma^{\text{prior}} &= -\frac{\delta}{2}\mathbf{I}\end{aligned}$$

For the entropy term, it can be expressed as follows.

$$\mathbb{E}_{q(z)}\left[-\log q(\mathbf{z})\right] = \tfrac{1}{2}\left\{\log\det\left(2\pi\boldsymbol{\Sigma}\right) + \log\left(\frac{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{2\pi}\right) - \frac{1}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}} + (d+1)\right\} - \mathbb{E}_{\text{Exp}(w|1)\mathcal{N}(z_2|w\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}},1)}\left[\log\phi\left(z_2 - \frac{1}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right)\right] \tag{61}$$

Using the expression in (61), the gradients of the entropy term are given as follows:

$$\mathbf{g}_\mu^{\text{entropy}} = 0 \tag{62}$$

$$\mathbf{g}_\alpha^{\text{entropy}} = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}} + \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\left(\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right)^2} - \mathbb{E}_{\text{Exp}(w|1)\mathcal{N}(z_2|0,1)}\left[\frac{\exp\left(-\frac{t^2}{2} - \log\phi(t)\right)}{\sqrt{2\pi}}\left(w + \frac{1}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right)\left(\frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right)\right] \tag{63}$$

$$\mathbf{g}_\Sigma^{\text{entropy}} = \frac{\boldsymbol{\Sigma}^{-1}}{2} - \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}}{2\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}} - \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}}{2\left(\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right)^2} + \mathbb{E}_{\text{Exp}(w|1)\mathcal{N}(z_2|0,1)}\left[\frac{\exp\left(-\frac{t^2}{2} - \log\phi(t)\right)}{\sqrt{2\pi}}\left(w + \frac{1}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}\right)\left(\frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}}{2\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}\right)\right] \tag{64}$$

where $t = z_2 + \frac{w\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}}$ and the expectations involve 2d integrations, which can be computed by Gauss-Hermite quadrature and Gauss-Laguerre quadrature.

The remaining thing is to compute the gradients about $\mathbb{E}_{q(z)}[f_n(\mathbf{z})]$. To compute the gradients, the reparametrization trick can be used. However, we can use the extended Bonnet's and Price's theorems (Lin et al., 2019). Assuming that $f_n(\mathbf{z})$ satisfies the assumptions needed for these two theorems, we obtain the following gradient expression:

$$\begin{aligned}
\mathbf{g}_1^n &:= \nabla_\mu \mathbb{E}_{q(z)}[f_n(\mathbf{z})] = \mathbb{E}_{q(z)}[\nabla_z f_n(\mathbf{z})] \approx \nabla_z f_n(\mathbf{z}) \\
\mathbf{g}_2^n &:= \nabla_\alpha \mathbb{E}_{q(z)}[f_n(\mathbf{z})] \\
&= \mathbb{E}_{q(z)}[u(\mathbf{z})\nabla_z f_n(\mathbf{z})] + v\mathbb{E}_{\mathcal{N}(\hat{z}|\mu,\Sigma)}[\nabla_{\hat{z}} f_n(\hat{\mathbf{z}})] \approx u(\mathbf{z})\nabla_z f_n(\mathbf{z}) + v\nabla_{\hat{z}} f_n(\hat{\mathbf{z}}) \\
&= \mathbb{E}_{q(w,z)}[w\nabla_z f_n(\mathbf{z})] \approx w\nabla_z f_n(\mathbf{z}) \\
\mathbf{g}_3^n &:= 2\nabla_\Sigma \mathbb{E}_{q(z)}[f_n(\mathbf{z})] = \mathbb{E}_{q(z)}\left[\nabla_z^2 f_n(\mathbf{z})\right] \approx \nabla_z^2 f_n(\mathbf{z})
\end{aligned}$$

where $v = \frac{1}{(\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})}$, $u(\mathbf{z}) := \frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}-1}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}$, and $w \sim \text{Exp}(w|1)$, $\hat{\mathbf{z}} \sim \mathcal{N}(\hat{\mathbf{z}}|\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\mathbf{z} = \hat{\mathbf{z}} + w\boldsymbol{\alpha}$.

Putting together, we can express the gradients in the following form, which will be used for deriving the extended variational Adam update.

$$\begin{aligned}
\nabla_\mu \mathcal{L}(\boldsymbol{\lambda}) &= -\sum_{n=1}^N \left(\mathbf{g}_1^n \underbrace{-\mathbf{g}_\mu^{\text{entropy}}/N}_{0}\right) \underbrace{-\delta(\boldsymbol{\mu}+\boldsymbol{\alpha})}_{\mathbf{g}_\mu^{\text{prior}}} \\
&= -\sum_{n=1}^N \underbrace{\mathbf{g}_1^n}_{:=\mathbf{g}_\mu^n} - \delta(\boldsymbol{\mu}+\boldsymbol{\alpha}) \qquad (65) \\
\nabla_\alpha \mathcal{L}(\boldsymbol{\lambda}) &= -\sum_{n=1}^N \left(\underbrace{\mathbf{g}_2^n - \mathbf{g}_\alpha^{\text{entropy}}/N}_{:=\mathbf{g}_\alpha^n}\right) \underbrace{-\delta(2\boldsymbol{\alpha}+\boldsymbol{\mu})}_{\mathbf{g}_\alpha^{\text{prior}}} \qquad (66) \\
\nabla_\Sigma \mathcal{L}(\boldsymbol{\lambda}) &= -\tfrac{1}{2}\sum_{n=1}^N \left(\mathbf{g}_3^n - 2\mathbf{g}_\Sigma^{\text{entropy}}/N\right) \underbrace{-\frac{\delta}{2}\mathbf{I}}_{\mathbf{g}_\Sigma^{\text{prior}}} + \tfrac{1}{2}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}) \\
&= -\tfrac{1}{2}\sum_{n=1}^N \left(\underbrace{\mathbf{g}_3^n - 2\mathbf{g}_\Sigma^{\text{entropy}}/N + \boldsymbol{\Sigma}^{-1}/N}_{:=\mathbf{g}_S^n}\right) - \frac{\delta}{2}\mathbf{I} + \tfrac{1}{2}\boldsymbol{\Sigma}^{-1} \qquad (67)
\end{aligned}$$

Similarly, for stochastic approximation, we can sub-sampling a data point n and use MC samples to approximate $\mathbf{g}_1^n$, $\mathbf{g}_2^n$, and $\mathbf{g}_3^n$. Plugging these stochastic gradients into (58)-(60), we obtain the NGD update:

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1} &\leftarrow (1-\beta)\boldsymbol{\Sigma}^{-1} + \beta(\delta\mathbf{I} + N\mathbf{g}_S^n) \\
\boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} - \beta\boldsymbol{\Sigma}(N(2\mathbf{g}_\mu^n - \mathbf{g}_\alpha^n) + \delta\boldsymbol{\mu}) \\
\boldsymbol{\alpha} &\leftarrow \boldsymbol{\alpha} - \beta\boldsymbol{\Sigma}(N(\mathbf{g}_\alpha^n - \mathbf{g}_\mu^n) + \delta\boldsymbol{\alpha})
\end{aligned}$$

G. Multivariate Normal Inverse-Gaussian Distribution

We consider the following mixture distribution (Barndorff-Nielsen, 1997), which is a Gaussian variance-mean mixture distribution. For simplicity, we assume λ is known.

$$q(w,\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}+w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})\text{InvGauss}(w|1,\lambda)$$

where $\text{InvGauss}(w|1,\lambda) = \left(\frac{\lambda}{2\pi w^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda}{2}\left(w + w^{-1}\right) + \lambda\right\}$ denotes the inverse Gaussian distribution.

Lemma 15 *The marginal distribution is*

$$q(\mathbf{z}) = \frac{\lambda^{\frac{1}{2}}}{(2\pi)^{\frac{d+1}{2}}}\det\left(\boldsymbol{\Sigma}\right)^{-1/2}\exp\left[\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda\right]\frac{2\mathcal{K}_{\frac{d+1}{2}}\left(\sqrt{\left(\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda\right)\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)+\lambda\right)}\right)}{\left(\sqrt{\frac{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda}{\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)+\lambda}}\right)^{\frac{-d-1}{2}}},$$

where $\mathcal{K}_v(x)$ denotes the modified Bessel function of the second kind.

Proof: By definition, we can compute the marginal distribution as follows.

$$q(\mathbf{z}) = \int_0^{+\infty} q(\mathbf{z}|w)q(w)dw$$

$$= \int_0^{+\infty}\det\left(2\pi w\boldsymbol{\Sigma}\right)^{-1/2}\exp\left\{-\tfrac{1}{2}\left[\left(\mathbf{z}-\boldsymbol{\mu}-w\boldsymbol{\alpha}\right)^T\left(w\boldsymbol{\Sigma}\right)^{-1}\left(\mathbf{z}-\boldsymbol{\mu}-w\boldsymbol{\alpha}\right)\right]\right\}\left(\frac{\lambda}{2\pi w^3}\right)^{1/2}\exp\left[-\frac{\lambda}{2}\left(w+\frac{1}{w}\right)+\lambda\right]dw$$

By grouping all terms related to w together, we have

$$=\frac{\lambda^{\frac{1}{2}}}{(2\pi)^{\frac{d+1}{2}}}\det\left(\boldsymbol{\Sigma}\right)^{-1/2}\exp\left[\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda\right]\int_0^{+\infty}w^{-\frac{d+3}{2}}\exp\left\{-\tfrac{1}{2}\left[w\left(\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda\right)+\frac{\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)+\lambda}{w}\right]\right\}dw$$

By completing a generalized inverse Gaussian form, we have

$$=\frac{\lambda^{\frac{1}{2}}}{(2\pi)^{\frac{d+1}{2}}}\det\left(\boldsymbol{\Sigma}\right)^{-1/2}\exp\left[\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda\right]\frac{2\mathcal{K}_{\frac{-d-1}{2}}\left(\sqrt{\left(\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda\right)\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)+\lambda\right)}\right)}{\left(\sqrt{\frac{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda}{\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)+\lambda}}\right)^{\frac{-d-1}{2}}}$$

We obtain the last step by the fact that $\mathcal{K}_v(x) = \mathcal{K}_{-v}(x)$

$$=\frac{\lambda^{\frac{1}{2}}}{(2\pi)^{\frac{d+1}{2}}}\det\left(\boldsymbol{\Sigma}\right)^{-1/2}\exp\left[\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda\right]\frac{2\mathcal{K}_{\frac{d+1}{2}}\left(\sqrt{\left(\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda\right)\left(\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)+\lambda\right)}\right)}{\left(\sqrt{\frac{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda}{\left(\mathbf{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)+\lambda}}\right)^{\frac{-d-1}{2}}}$$

□

Similarly, the sufficient statistics $\boldsymbol{\phi}_z(\mathbf{z},w) = \left\{\mathbf{z}/w, \mathbf{z}, \mathbf{z}\mathbf{z}^T/w\right\}$ and the natural parameter $\boldsymbol{\lambda}_z = \left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\right\}$ can be read from $q(\mathbf{z}|w) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}+w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})$. The joint distribution $q(\mathbf{z},w)$ is a conditional EF because $q(w)$ is an inverse Gaussian distribution with known parameters, which is a EF distribution. Likewise, we can show that the joint distribution is also a minimal conditional EF.

We can derive the expectation parameter shown below:

$$\begin{aligned}
\mathbf{m} &:= \mathbb{E}_{\text{InvGauss}(w|1,\lambda)\mathcal{N}(z|\mu+w\alpha,w\Sigma)}\left[w^{-1}\mathbf{z}\right] = \left(1+\lambda^{-1}\right)\boldsymbol{\mu}+\boldsymbol{\alpha},\\
\mathbf{m}_\alpha &:= \mathbb{E}_{\text{InvGauss}(w|1,\lambda)\mathcal{N}(z|\mu+w\alpha,w\Sigma)}\left[\mathbf{z}\right] = \boldsymbol{\mu}+\boldsymbol{\alpha},\\
\mathbf{M} &:= \mathbb{E}_{\text{InvGauss}(w|1,\lambda)\mathcal{N}(z|\mu+w\alpha,w\Sigma)}\left[w^{-1}\mathbf{z}\mathbf{z}^T\right] = \left(1+\lambda^{-1}\right)\boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\alpha}\boldsymbol{\alpha}^T+\boldsymbol{\mu}\boldsymbol{\alpha}^T+\boldsymbol{\alpha}\boldsymbol{\mu}^T+\boldsymbol{\Sigma}
\end{aligned}$$

The sufficient statistics, natural parameters, and expectation parameters are summarized below:

$$\begin{bmatrix}\mathbf{z}/w\\ \mathbf{z}\\ \mathbf{z}\mathbf{z}^T/w\end{bmatrix}\quad\begin{bmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\end{bmatrix}\quad\begin{bmatrix}\left(1+\lambda^{-1}\right)\boldsymbol{\mu}+\boldsymbol{\alpha}\\ \boldsymbol{\mu}+\boldsymbol{\alpha}\\ \left(1+\lambda^{-1}\right)\boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\alpha}\boldsymbol{\alpha}^T+\boldsymbol{\mu}\boldsymbol{\alpha}^T+\boldsymbol{\alpha}\boldsymbol{\mu}^T+\boldsymbol{\Sigma}\end{bmatrix}$$

G.1. Derivation of the NGD Update

We consider the variational approximation using the normal inverse Gaussian distribution $q(\mathbf{z}|\boldsymbol{\lambda})$. We consider the same model as Appendix B.1 with a Gaussian-prior $\mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})$ on $\mathbf{z}$. The lower bound is defined as below:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda}) &= \mathbb{E}_{q(z|\lambda)}\left[\sum_{n=1}^{N} \log p(\mathcal{D}_n|\mathbf{z}) + \log \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I}) - \log q(\mathbf{z}|\boldsymbol{\lambda})\right] \\ &= \mathbb{E}_{q(z)}\left[-h(\mathbf{z})\right], \text{ where } h(\mathbf{z}) := -\left[\log \frac{\mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})}{q(\mathbf{z})} + \sum_n \log p(\mathcal{D}_n|\mathbf{z})\right].\end{aligned}$$

Our goal is to compute the gradient of this ELBO with respect to the expectation parameters.

G.2. Natural Gradient for $q(\mathbf{z}|w)$

Likewise, we do not need to compute these gradients with respect to the expectation parameters directly. The gradients can be computed in terms of $\{\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\}$. Similarly, by the chain rule, we can express the following gradients with respect to the expectation parameters in terms of the gradients with respect to $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma}$.

$$\begin{aligned}\nabla_m \mathcal{L} &= \lambda \nabla_\mu \mathcal{L} - \lambda \nabla_\alpha \mathcal{L} - 2\left(\nabla_\Sigma \mathcal{L}\right)\boldsymbol{\mu} \\ \nabla_{m_\alpha} \mathcal{L} &= (1+\lambda)\nabla_\alpha \mathcal{L} - \lambda \nabla_\mu \mathcal{L} - 2\left(\nabla_\Sigma \mathcal{L}\right)\boldsymbol{\alpha} \\ \nabla_M \mathcal{L} &= \nabla_\Sigma \mathcal{L}\end{aligned}$$

By plugging the gradients into the update in (14) and then re-expressing the update in terms of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Sigma}$, we obtain the natural gradient update in terms of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma}^{-1} \leftarrow \boldsymbol{\Sigma}^{-1} - 2\beta \nabla_\Sigma \mathcal{L} \tag{68}$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \beta \boldsymbol{\Sigma}\left(\lambda \nabla_\mu \mathcal{L} - \lambda \nabla_\alpha \mathcal{L}\right) \tag{69}$$

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + \beta \boldsymbol{\Sigma}\left((1+\lambda)\nabla_\alpha \mathcal{L} - \lambda \nabla_\mu \mathcal{L}\right) \tag{70}$$

We can compute gradients with respect to $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma}$ by the extended Bonnet's and Price's theorem (Lin et al., 2019). In Lin et al. (2019), they discuss the conditions of the target function $h(\mathbf{z})$ when it comes to applying these theorems.

$$\begin{aligned}\nabla_\mu \mathcal{L}(\boldsymbol{\lambda}) &= -\mathbb{E}_{q(z)}\left[\nabla_z h(\mathbf{z})\right] \quad \approx -\nabla_z h(\mathbf{z}) \\ \nabla_\alpha \mathcal{L}(\boldsymbol{\lambda}) &= -\mathbb{E}_{q(w,z)}\left[w \nabla_z h_n(\mathbf{z})\right] \approx -w \nabla_z h_n(\mathbf{z}) \\ &= -\mathbb{E}_{q(z)}\left[u(\mathbf{z}) \nabla_z h_n(\mathbf{z})\right] \approx -u(\mathbf{z}) \nabla_z h_n(\mathbf{z}) \\ \nabla_\Sigma \mathcal{L}(\boldsymbol{\lambda}) &= -\tfrac{1}{2}\mathbb{E}_{q(z)}\left[w \nabla_z^2 h(\mathbf{z})\right] \quad \approx -\frac{w}{2}\nabla_z^2 h(\mathbf{z}). \\ &= -\tfrac{1}{2}\mathbb{E}_{q(z)}\left[u(\mathbf{z}) \nabla_z^2 h(\mathbf{z})\right] \quad \approx -\frac{u(\mathbf{z})}{2}\nabla_z^2 h(\mathbf{z}).\end{aligned}$$

where $u(\mathbf{z}) := \sqrt{\frac{(\mathbf{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})+\lambda}{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda}} \frac{\mathcal{K}_{\frac{d-1}{2}}\left(\sqrt{\left(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda\right)\left((\mathbf{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})+\lambda\right)}\right)}{\mathcal{K}_{\frac{d+1}{2}}\left(\sqrt{\left(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}+\lambda\right)\left((\mathbf{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})+\lambda\right)}\right)}$ and $w \sim \text{InvGauss}(w|1, \lambda)$, $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})$.

Directly calculating the ratio between the modified Bessel functions of the second kind is expensive and numerically unstable when v is large. However, the ratio between consecutive order has a tight and algebraic bound (Ruiz-Antolín & Segura, 2016) as given below. When $v \in \mathcal{R}$ and $v \geq \frac{1}{2}$, the bound of the ratio is

$$D_{2v-1}(v, x) \leq \frac{\mathcal{K}_{v-1}(x)}{\mathcal{K}_v(x)} \leq D_0(v, x)$$

where function $D_\alpha(v,x)$ is defined as below.

$$D_\alpha(v,x) := \frac{x}{\psi_\alpha(v,x) + \sqrt{(\psi_\alpha(v,x))^2 + x^2}}$$
$$\psi_\alpha(v,x) := (v-\tfrac{1}{2}) - \frac{\tau_\alpha(v)}{2\sqrt{(\tau_\alpha(v))^2 + x^2}},\ \ \tau_\alpha(v) := v - \frac{\alpha+1}{2}$$

For natural number $v \in \mathbb{N}$, a tighter bound (see Eq (3.10) at Yang & Chu (2017)) with higher computation cost can be used, where we make use of the following relationship due to Eq (72) and (73).

$$\frac{\mathcal{K}_{v-1}(x)}{\mathcal{K}_v(x)} = \frac{\mathcal{K}_{v+1}(x)}{\mathcal{K}_v(x)} - \frac{2v}{x}$$

To compute the ratio, we propose to use the following approximation when $v \geq \frac{1}{2}$. A similar approach to approximate the ratio between two modified Bessel functions of the first kind is used in Oh et al. (2019) and Kumar & Tsvetkov (2018).

$$\frac{\mathcal{K}_{v-1}(x)}{\mathcal{K}_v(x)} \approx \frac{D_{2v-1}(v,x) + D_0(v,x)}{2} \tag{71}$$

Now, we discuss how to compute $\nabla_x \log \mathcal{K}_v(x)$ and $\nabla_x^2 \log \mathcal{K}_v(x)$. The first term appears when we compute $\nabla_z h(\mathbf{z})$. Similarly, the second term appears when we compute $\nabla_z^2 h(\mathbf{z})$.

First, we make use of the recurrence forms of the modified Bessel function of the second kind for $v \in \mathcal{R}$ (see page 20 at Culham (2004)).

$$\nabla_x \mathcal{K}_v(x) = -\mathcal{K}_{v-1}(x) - \frac{v}{x}\mathcal{K}_v(x) \tag{72}$$
$$\nabla_x \mathcal{K}_v(x) = \frac{v}{x}\mathcal{K}_v(x) - \mathcal{K}_{v+1}(x) \tag{73}$$

Using these recurrence forms, we have

$$\frac{\nabla_x \mathcal{K}_v(x)}{\mathcal{K}_v(x)} = -\frac{\mathcal{K}_{v-1}(x)}{\mathcal{K}_v(x)} - \frac{v}{x} \tag{74}$$

Furthermore, we have the following result due to the the recurrence forms.

$$\begin{aligned}
\nabla_x^2 \mathcal{K}_v(x) &= \nabla_x \left[\overbrace{-\mathcal{K}_{v-1}(x) - \frac{v}{x}\mathcal{K}_v(x)}^{\nabla_x \mathcal{K}_v(x)} \right] \\
&= -\nabla_x \mathcal{K}_{v-1}(x) - \frac{v}{x}\nabla_x \mathcal{K}_v(x) + \frac{v}{x^2}\mathcal{K}_v(x) \\
&= -\left(\underbrace{\frac{v-1}{x}\mathcal{K}_{v-1}(x) - \mathcal{K}_v(x)}_{\nabla_x \mathcal{K}_{v-1}(x)} \right) - \frac{v}{x}\nabla_x \mathcal{K}_v(x) + \frac{v}{x^2}\mathcal{K}_v(x) \\
&= -\frac{v-1}{x}\mathcal{K}_{v-1}(x) + \frac{v+x^2}{x^2}\mathcal{K}_v(x) - \frac{v}{x}\nabla_x \mathcal{K}_v(x) \\
&= -\frac{v-1}{x}\mathcal{K}_{v-1}(x) + \frac{v+x^2}{x^2}\mathcal{K}_v(x) - \frac{v}{x}\left(\underbrace{-\mathcal{K}_{v-1}(x) - \frac{v}{x}\mathcal{K}_v(x)}_{\nabla_x \mathcal{K}_v(x)} \right) \\
&= \frac{1}{x}\mathcal{K}_{v-1}(x) + \frac{v+x^2+v^2}{x^2}\mathcal{K}_v(x)
\end{aligned}$$

which implies that

$$\frac{\nabla_x^2 \mathcal{K}_v(x)}{\mathcal{K}_v} = \frac{1}{x}\frac{\mathcal{K}_{v-1}(x)}{\mathcal{K}_v(x)} + \frac{v + x^2 + v^2}{x^2} \tag{75}$$

Using Eq (74) and (75), we have

$$\begin{aligned}
\nabla_x \log \mathcal{K}_v(x) &= \frac{\nabla_x \mathcal{K}_v(x)}{\mathcal{K}_v(x)} \\
&= -\frac{\mathcal{K}_{v-1}(x)}{\mathcal{K}_v(x)} - \frac{v}{x} \\
\nabla_x^2 \log \mathcal{K}_v(x) &= \nabla_x \left[\frac{\nabla_x \mathcal{K}_v(x)}{\mathcal{K}_v(x)}\right] \\
&= \frac{\nabla_x^2 \mathcal{K}_v(x)}{\mathcal{K}_v(x)} - \left(\frac{\nabla_x \mathcal{K}_v(x)}{\mathcal{K}_v(x)}\right)^2 \\
&= \frac{1}{x}\frac{\mathcal{K}_{v-1}(x)}{\mathcal{K}_v(x)} + \frac{v + x^2 + v^2}{x^2} - \left(-\frac{\mathcal{K}_{v-1}(x)}{\mathcal{K}_v(x)} - \frac{v}{x}\right)^2 \\
&= \frac{1 - 2v}{x}\left(\frac{\mathcal{K}_{v-1}(x)}{\mathcal{K}_v(x)}\right) - \left(\frac{\mathcal{K}_{v-1}(x)}{\mathcal{K}_v(x)}\right)^2 + \frac{v}{x^2} + 1
\end{aligned}$$

where the ratio can be approximated by Eq (71).

H. Multivariate Symmetric Normal Inverse-Gaussian Distribution

The symmetric normal inverse-Gaussian distribution is a scale mixture distribution. A difference between the distribution at Appendix G is shown in red. Such difference allows this distribution to have heavy tails.

$$q(w, \mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w^{-1}\boldsymbol{\Sigma})\text{InvGauss}(w|1, \lambda)$$

where we assume λ is fixed for simplicity and $\text{InvGauss}(w|1,\lambda) = \left(\frac{\lambda}{2\pi w^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda}{2}\left(w + w^{-1}\right) + \lambda\right\}$.

Similarly, we can show that the marginal distribution is

$$q(\mathbf{z}) = \frac{\lambda^{\frac{1}{2}}}{(2\pi)^{\frac{d+1}{2}}} \det(\boldsymbol{\Sigma})^{-1/2} \exp(\lambda) \frac{2\mathcal{K}_{\frac{d-1}{2}}\left(\sqrt{\lambda\left((\mathbf{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) + \lambda\right)}\right)}{\left(\sqrt{\frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})+\lambda}{\lambda}}\right)^{\frac{d-1}{2}}}$$

Furthermore, the mixture distribution is a minimal conditional EF distribution. The sufficient statistics, natural parameters, and expectation parameters are summarized below:

$$\begin{bmatrix} w\mathbf{z} \\ w\mathbf{z}\mathbf{z}^T \end{bmatrix} \quad \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{bmatrix} \quad \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma} \end{bmatrix}$$

Likewise, we can derive the natural-gradient update in terms of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as shown below.

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1} &\leftarrow \boldsymbol{\Sigma}^{-1} - 2\beta\nabla_\Sigma \mathcal{L} \\
\boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} + \beta\boldsymbol{\Sigma}\nabla_\mu \mathcal{L}
\end{aligned}$$

Now, we discuss the gradient computation of the following lower bound.

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

For the prior term, we have

$$\mathbb{E}_{q(z)}\left[\log \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})\right] = -\frac{d}{2}\log(2\pi) + \frac{d\delta}{2} - \frac{\delta}{2}\left(\boldsymbol{\mu}^T\boldsymbol{\mu} + \left(1+\lambda^{-1}\right)\mathrm{Tr}(\boldsymbol{\Sigma})\right) \tag{76}$$

Due to Eq (76), the closed-form gradients are given below.

$$\mathbf{g}_{\mu}^{\text{prior}} = -\delta\boldsymbol{\mu}$$
$$\mathbf{g}_{\Sigma}^{\text{prior}} = -\frac{\delta\left(1+\lambda^{-1}\right)}{2}\mathbf{I}$$

In this example, we can re-express the entropy term as below.

$$\mathbb{E}_{q(z)}\left[-\log q(\mathbf{z})\right] = \frac{(d+1)\log(2\pi)}{2} + \frac{\log\det(\boldsymbol{\Sigma})}{2} - \lambda - \frac{d+1}{4}\log(\lambda) - \mathbb{E}_{\text{InvGauss}(w|1,\lambda)\chi^2(z_1|d)}\left[\log\left(\frac{2\mathcal{K}_{\frac{d-1}{2}}\left(\sqrt{\lambda\left(w^{-1}z_1+\lambda\right)}\right)}{\left(w^{-1}z_1+\lambda\right)^{\frac{d-1}{4}}}\right)\right] \tag{77}$$

where $\chi^2(z_1|d)$ denotes the chi-squared distribution with d degrees of freedom. The expectation can be computed by the inverse Gaussian quadrature (Choi et al., 2018) and the generalized Gauss-Laguerre quadrature.

By Eq (77), the closed-form gradients of the entropy term are shown below.

$$\mathbf{g}_{\mu}^{\text{entropy}} = 0$$
$$\mathbf{g}_{\Sigma}^{\text{entropy}} = \frac{\boldsymbol{\Sigma}^{-1}}{2}$$

The remaining step is to compute the gradients about $\mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right]$. To compute the gradients, we use the extended the Bonnet's and Price's theorems (Lin et al., 2019). Assuming that $f_n(\mathbf{z})$ satisfies the assumptions needed for these two theorems, we obtain the following gradient expression:

$$\begin{aligned}
\mathbf{g}_1^n &:= \nabla_{\mu}\mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right] = \mathbb{E}_{q(z)}\left[\nabla_z f_n(\mathbf{z})\right] \approx \nabla_z f_n(\mathbf{z}) \\
\mathbf{g}_2^n &:= 2\nabla_{\Sigma}\mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right] \\
&= \mathbb{E}_{q(z)}\left[u(\mathbf{z})\nabla_z^2 f_n(\mathbf{z})\right] \approx u(\mathbf{z})\nabla_z^2 f_n(\mathbf{z}) \\
&= \mathbb{E}_{q(w,z)}\left[w\nabla_z^2 f_n(\mathbf{z})\right] \approx w^{-1}\nabla_z^2 f_n(\mathbf{z})
\end{aligned}$$

where $u(\mathbf{z}) := \sqrt{\frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})+\lambda}{\lambda}}\frac{\mathcal{K}_{\frac{d-3}{2}}\left(\sqrt{\lambda\left((\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})+\lambda\right)}\right)}{\mathcal{K}_{\frac{d-1}{2}}\left(\sqrt{\lambda\left((\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})+\lambda\right)}\right)}$, where the ratio between the Bessel functions can be approximated by Eq (71) when $d \geq 2$, and $w \sim \text{InvGauss}(w|1,\lambda)$, $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, w^{-1}\boldsymbol{\Sigma})$.

Putting together, we can express the gradients in the following form:

$$\nabla_{\mu}\mathcal{L}(\boldsymbol{\lambda}) = -\sum_{n=1}^{N}\underbrace{\mathbf{g}_1^n}_{:=\mathbf{g}_{\mu}^n} - \delta\boldsymbol{\mu} \tag{78}$$

$$\nabla_{\Sigma}\mathcal{L}(\boldsymbol{\lambda}) = -\tfrac{1}{2}\sum_{n=1}^{N}\underbrace{\mathbf{g}_2^n}_{:=\mathbf{g}_S^n} - \frac{\delta(1+\lambda^{-1})}{2}\mathbf{I} + \tfrac{1}{2}\boldsymbol{\Sigma}^{-1} \tag{79}$$

Similarly, for stochastic approximation, we can sub-sampling a data point n and use MC samples to approximate $\mathbf{g}_1^n$ and $\mathbf{g}_2^n$. Plugging these stochastic gradients into

$$\boldsymbol{\Sigma}^{-1} \leftarrow (1-\beta)\boldsymbol{\Sigma}^{-1} + \beta(\delta\left(1+\lambda^{-1}\right)\mathbf{I} + N\mathbf{g}_S^n)$$
$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta\boldsymbol{\Sigma}(N\mathbf{g}_{\mu}^n + \delta\boldsymbol{\mu})$$

I. Matrix-Variate Gaussian Distribution

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

Lemma 16 *Matrix Gaussian distribution is a member of the multi-linear exponential family.*

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

$$\begin{aligned}
&\mathcal{MN}(\mathbf{Z}|\mathbf{W}, \mathbf{U}, \mathbf{V}) \\
&= (2\pi)^{-dp/2} \exp\left[-\tfrac{1}{2}\mathrm{Tr}\left(\mathbf{V}^{-1}(\mathbf{Z}-\mathbf{W})^T\mathbf{U}^{-1}(\mathbf{Z}-\mathbf{W})\right) - (d/2\log\mathrm{Det}(\mathbf{V}) + p/2\log\mathrm{Det}(\mathbf{U}))\right] \\
&= (2\pi)^{-dp/2} \exp\Big\{\mathrm{Tr}\left(\mathbf{\Lambda}_3\left(-\tfrac{1}{2}\mathbf{Z} + \mathbf{\Lambda}_1\right)^T\mathbf{\Lambda}_2\mathbf{Z}\right) \\
&\qquad - \tfrac{1}{2}\Big[\mathrm{Tr}\left(\mathbf{\Lambda}_3\mathbf{\Lambda}_1^T\mathbf{\Lambda}_2\mathbf{\Lambda}_1\right) + d\log\mathrm{Det}(\mathbf{\Lambda}_3) + p\log\mathrm{Det}(\mathbf{\Lambda}_2)\Big]\Big\}.
\end{aligned}$$

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

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

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

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

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

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

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

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

$$\begin{aligned}
\nabla_{M_1}\mathbb{E}_{q(Z|\lambda)}\left[h(\mathbf{Z})\right] &= \left(\mathbf{\Lambda}_2\right)^{-1}\nabla_W\mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right]\left(\mathbf{\Lambda}_3\right)^{-1} \\
\nabla_{M_2}\mathbb{E}_{q(Z|\lambda)}\left[h(\mathbf{Z})\right] &= \frac{-2}{p}\nabla_U\mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right] \\
\nabla_{M_3}\mathbb{E}_{q(Z|\lambda)}\left[h(\mathbf{Z})\right] &= \frac{-2}{d}\nabla_V\mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right]
\end{aligned}$$

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

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

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

$$\begin{aligned}
\nabla_W \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right] &= \mathrm{Mat}\left(\mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z})\right]\right) \\
&= \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})\right] \\
\nabla_U \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right] &= \left(\nabla_U \boldsymbol{\Sigma}\right) \nabla_\Sigma \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z})\right] \\
&= \tfrac{1}{2}\left(\nabla_U \boldsymbol{\Sigma}\right) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z^2 \hat{h}(\mathbf{z})\right] \\
&\approx \tfrac{1}{2}\left(\nabla_U \boldsymbol{\Sigma}\right) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z}) \nabla_z \hat{h}(\mathbf{z})^T\right] \qquad (80) \\
&= \tfrac{1}{2}\mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z}) \mathbf{V} \nabla_Z h(\mathbf{Z})^T\right] \\
\nabla_V \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[h(\mathbf{Z})\right] &= \left(\nabla_V \boldsymbol{\Sigma}\right) \nabla_\Sigma \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z})\right] \\
&= \tfrac{1}{2}\left(\nabla_V \boldsymbol{\Sigma}\right) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z^2 \hat{h}(\mathbf{z})\right] \\
&\approx \tfrac{1}{2}\left(\nabla_V \boldsymbol{\Sigma}\right) \mathbb{E}_{\mathcal{N}(z|\mu,\Sigma)}\left[\nabla_z \hat{h}(\mathbf{z}) \nabla_z \hat{h}(\mathbf{z})^T\right] \qquad (81) \\
&= \tfrac{1}{2}\mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})^T \mathbf{U} \nabla_Z h(\mathbf{Z})\right].
\end{aligned}$$

To avoid computation of the Hessian, we have used the Gauss-Newton approximation (Khan et al., 2018) in Eq (80) and Eq. (81).

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

$$\begin{aligned}
\boldsymbol{\Lambda}_1 &\leftarrow \boldsymbol{\Lambda}_1 - \beta_1 \left(\boldsymbol{\Lambda}_2\right)^{-1} \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})\right] \left(\boldsymbol{\Lambda}_3\right)^{-1} \\
\boldsymbol{\Lambda}_2 &\leftarrow \boldsymbol{\Lambda}_2 + \beta_2 \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z}) \mathbf{V} \nabla_Z h(\mathbf{Z})^T\right] \\
\boldsymbol{\Lambda}_3 &\leftarrow \boldsymbol{\Lambda}_3 + \beta_2 \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})^T \mathbf{U} \nabla_Z h(\mathbf{Z})\right]
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{W} &\leftarrow \mathbf{W} - \beta_1 \mathbf{U} \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})\right] \mathbf{V} \\
\left(\mathbf{U}\right)^{-1} &\leftarrow \left(\mathbf{U}\right)^{-1} + \beta_2 \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z}) \mathbf{V} \nabla_Z h(\mathbf{Z})^T\right] \\
\left(\mathbf{V}\right)^{-1} &\leftarrow \left(\mathbf{V}\right)^{-1} + \beta_2 \mathbb{E}_{\mathcal{MN}(Z|W,U,V)}\left[\nabla_Z h(\mathbf{Z})^T \mathbf{U} \nabla_Z h(\mathbf{Z})\right]
\end{aligned}$$

J. Extensions to Variational Adam

For simplicity, we consider a case when variational parameters of $q(w|\boldsymbol{\lambda}_w)$ are fixed. Since $\boldsymbol{\lambda}_w$ is fixed, using the same derivation as Khan et al. (2018), we obtain the following natural-gradient update with the natural momentum ($0 \leq m < 1$).

$$\boldsymbol{\lambda}_z^{t+1} = \frac{1}{1-m}\boldsymbol{\lambda}_z^t - \frac{m}{1-m}\boldsymbol{\lambda}_z^{t-1} + \frac{\beta}{1-m}\nabla_{m_z}\mathcal{L}(\boldsymbol{\lambda}_z)\Big|_{\boldsymbol{\lambda}_z=\boldsymbol{\lambda}_z^t} \qquad (82)$$

We assume the model prior is a Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})$ to derive extensions of the variational Adam update, where the variational distribution is a Gaussian mixture distribution such as skew Gaussian, exponentially modified Gaussian, symmetric normal inverse-Gaussian, and Student's t-distribution.

J.1. Extension for Skew Gaussian

Re-expressing the update (82) in terms of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Sigma}$ (the same derivation as Khan et al. (2018)), we obtain the following update:

$$\boldsymbol{\Sigma}_{t+1}^{-1} = \boldsymbol{\Sigma}_t^{-1} - 2\frac{\beta}{1-m}\nabla_{\Sigma_t}\mathcal{L} + \frac{m}{1-m}\left(\boldsymbol{\Sigma}_t^{-1} - \boldsymbol{\Sigma}_{t-1}^{-1}\right)$$
$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \frac{\beta}{1-m}\boldsymbol{\Sigma}_{t+1}\left(\frac{1}{1-c^2}\nabla_{\mu_t}\mathcal{L} - \frac{c}{1-c^2}\nabla_{\alpha_t}\mathcal{L}\right) + \frac{m}{1-m}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1})$$
$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \frac{\beta}{1-m}\boldsymbol{\Sigma}_{t+1}\left(\frac{1}{1-c^2}\nabla_{\alpha_t}\mathcal{L} - \frac{c}{1-c^2}\nabla_{\mu_t}\mathcal{L}\right) + \frac{m}{1-m}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t-1})$$

where we use a skew Gaussian distribution as the variational distribution, $\nabla_{\mu_t}\mathcal{L}$, $\nabla_{\alpha_t}\mathcal{L}$, and $\nabla_{\Sigma_t}\mathcal{L}$ are defined at (53) -(55).

We make use of the same approximations as Khan et al. (2018) such as the gradient-magnitude of the Hessian approximation, the square root approximation, $\boldsymbol{\Sigma}_{t-1} \approx \boldsymbol{\Sigma}_t$, and a diagonal covariance structure in $\boldsymbol{\Sigma}$ to obtain an extension of the variational skew-Adam update. Recall that $\mathbf{g}_\alpha^{\text{entropy}}$ and $\mathbf{g}_\Sigma^{\text{entropy}}$ are defined at (51) and (52) and $c = \sqrt{\frac{2}{\pi}}$. Using the same algebra manipulation used in Khan et al. (2018), we obtain the variational Adam update with Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})$, where $\boldsymbol{\Sigma}^{-1} = \text{Diag}(N\mathbf{s} + \delta)$, $v = \frac{c}{(1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})}$, and $u(\mathbf{z}) = \frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}{1+\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}$.

Skew Gaussian extension

while not converged **do**
  $\hat{\mathbf{z}} \leftarrow \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\sigma} \leftarrow 1/\sqrt{N\mathbf{s} + \delta}$
  $\mathbf{z} \leftarrow \hat{\mathbf{z}} + |w|\boldsymbol{\alpha}$, where $w \sim \mathcal{N}(0, 1)$
  Randomly sample a data example $\mathcal{D}_i$
  $\mathbf{g}_\mu \leftarrow -\nabla \log p(\mathcal{D}_i|\mathbf{z})$
  option I: $\hat{\mathbf{g}}_\alpha \leftarrow -|w|\nabla \log p(\mathcal{D}_i|\mathbf{z})$
  option II: $\hat{\mathbf{g}}_\alpha \leftarrow -\left[v\nabla \log p(\mathcal{D}_i|\hat{\mathbf{z}}) + u(\mathbf{z})\nabla \log p(\mathcal{D}_i|\mathbf{z})\right]$
  $\mathbf{g}_\alpha \leftarrow \hat{\mathbf{g}}_\alpha - \mathbf{g}_\alpha^{\text{entropy}}/N$
  $\mathbf{g}_s \leftarrow \mathbf{g}_\mu \circ \mathbf{g}_\mu - \text{diag}\left(2\mathbf{g}_\Sigma^{\text{entropy}}\right)/N + (\mathbf{s} + \delta/N)$
  $\mathbf{m}_\mu \leftarrow \gamma_1\,\mathbf{m}_\mu + (1-\gamma_1)\left(\frac{\mathbf{g}_\mu - c\mathbf{g}_\alpha}{1-c^2} + \delta\boldsymbol{\mu}/N\right)$
  $\mathbf{m}_\alpha \leftarrow \gamma_1\,\mathbf{m}_\alpha + (1-\gamma_1)\left(\frac{\mathbf{g}_\alpha - c\mathbf{g}_\mu}{1-c^2} + \delta\boldsymbol{\alpha}/N\right)$
  $\mathbf{s} \leftarrow \gamma_2\,\mathbf{s} + (1-\gamma_2)\,\mathbf{g}_s$
  $\hat{\mathbf{m}}_\mu \leftarrow \mathbf{m}_\mu/(1-\gamma_1^t), \quad \hat{\mathbf{m}}_\alpha \leftarrow \mathbf{m}_\alpha/(1-\gamma_1^t), \quad \hat{\mathbf{s}} \leftarrow (\mathbf{s} + \delta/N)/(1-\gamma_2^t)$
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta\,\hat{\mathbf{m}}_\mu/\sqrt{\hat{\mathbf{s}}}$
  $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \beta\,\hat{\mathbf{m}}_\alpha/\sqrt{\hat{\mathbf{s}}})$
  $t \leftarrow t + 1$
end while

J.2. Extension for Exponentially Modified Gaussian

Similarly, re-expressing the update (82) in terms of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Sigma}$, we obtain the following update:

$$\boldsymbol{\Sigma}_{t+1}^{-1} = \boldsymbol{\Sigma}_t^{-1} - 2\frac{\beta}{1-m}\nabla_{\Sigma_t}\mathcal{L} + \frac{m}{1-m}\left(\boldsymbol{\Sigma}_t^{-1} - \boldsymbol{\Sigma}_{t-1}^{-1}\right)$$
$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \frac{\beta}{1-m}\boldsymbol{\Sigma}_{t+1}\left(2\nabla_{\mu_t}\mathcal{L} - \nabla_{\alpha_t}\mathcal{L}\right) + \frac{m}{1-m}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1})$$
$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \frac{\beta}{1-m}\boldsymbol{\Sigma}_{t+1}\left(\nabla_{\alpha_t}\mathcal{L} - \nabla_{\mu_t}\mathcal{L}\right) + \frac{m}{1-m}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t-1})$$

where we use an exponentially modified Gaussian distribution as the variational distribution, $\nabla_{\mu_t}\mathcal{L}$, $\nabla_{\alpha_t}\mathcal{L}$, and $\nabla_{\Sigma_t}\mathcal{L}$ are defined at (65) -(67).

Likewise, we can obtain the variational Adam update with Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})$ as shown below, where $\mathbf{g}_\alpha^{\text{entropy}}$ and $\mathbf{g}_\Sigma^{\text{entropy}}$ are defined at (63) and (64), $\boldsymbol{\Sigma}^{-1} = \text{Diag}(N\mathbf{s} + \delta)$, $v = \frac{1}{(\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha})}$, and $u(\mathbf{z}) = \frac{(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} - 1}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}}$

Exponentially Modified Gaussian extension

while not converged **do**
  $\hat{\mathbf{z}} \leftarrow \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\sigma} \leftarrow 1/\sqrt{N\mathbf{s} + \delta}$
  $\mathbf{z} \leftarrow \hat{\mathbf{z}} + w\boldsymbol{\alpha}$, where $w \sim \text{Exp}(1)$
  Randomly sample a data example $\mathcal{D}_i$
  $\mathbf{g}_\mu \leftarrow -\nabla \log p(\mathcal{D}_i|\mathbf{z})$
  option I: $\hat{\mathbf{g}}_\alpha \leftarrow -w\nabla \log p(\mathcal{D}_i|\mathbf{z})$
  option II: $\hat{\mathbf{g}}_\alpha \leftarrow -\left[v\nabla \log p(\mathcal{D}_i|\hat{\mathbf{z}}) + u(\mathbf{z})\nabla \log p(\mathcal{D}_i|\mathbf{z})\right]$
  $\mathbf{g}_\alpha \leftarrow \hat{\mathbf{g}}_\alpha - \mathbf{g}_\alpha^{\text{entropy}}/N$
  $\mathbf{g}_s \leftarrow \mathbf{g}_\mu \circ \mathbf{g}_\mu - \text{diag}\left(2\mathbf{g}_\Sigma^{\text{entropy}}\right)/N + (\mathbf{s} + \delta/N)$
  $\mathbf{m}_\mu \leftarrow \gamma_1 \mathbf{m}_\mu + (1-\gamma_1)\left(2\mathbf{g}_\mu - \mathbf{g}_\alpha + \delta\boldsymbol{\mu}/N\right)$
  $\mathbf{m}_\alpha \leftarrow \gamma_1 \mathbf{m}_\alpha + (1-\gamma_1)\left(\mathbf{g}_\alpha - \mathbf{g}_\mu + \delta\boldsymbol{\alpha}/N\right)$
  $\mathbf{s} \leftarrow \gamma_2 \mathbf{s} + (1-\gamma_2)\mathbf{g}_s$
  $\hat{\mathbf{m}}_\mu \leftarrow \mathbf{m}_\mu/(1-\gamma_1^t), \quad \hat{\mathbf{m}}_\alpha \leftarrow \mathbf{m}_\alpha/(1-\gamma_1^t), \quad \hat{\mathbf{s}} \leftarrow (\mathbf{s} + \delta/N)/(1-\gamma_2^t)$
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta\ \hat{\mathbf{m}}_\mu/\sqrt{\hat{\mathbf{s}}}$
  $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \beta\ \hat{\mathbf{m}}_\alpha/\sqrt{\hat{\mathbf{s}}}$
  $t \leftarrow t + 1$
end while

J.3. Extension for Student's t-distribution

Likewise, re-expressing the update (82) in terms of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, we obtain the following update:

$$\boldsymbol{\Sigma}_{t+1}^{-1} = \boldsymbol{\Sigma}_t^{-1} - 2\frac{\beta}{1-m}\nabla_{\Sigma_t}\mathcal{L} + \frac{m}{1-m}\left(\boldsymbol{\Sigma}_t^{-1} - \boldsymbol{\Sigma}_{t-1}^{-1}\right)$$
$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \frac{\beta}{1-m}\boldsymbol{\Sigma}_{t+1}\nabla_{\mu_t}\mathcal{L} + \frac{m}{1-m}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1})$$

where we use a Student's t-distribution with fixed $\alpha > 1$ as the variational distribution, $\nabla_{\mu_t}\mathcal{L}$ and $\nabla_{\Sigma_t}\mathcal{L}$ are defined at (83) - (84).

Now, we consider the following lower bound ($\mathbf{z} \in \mathcal{R}^d$).

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(z|\lambda)}\left[\sum_{n=1}^{N}\underbrace{\log p(\mathcal{D}_n|\mathbf{z})}_{-f_n(\mathbf{z})} + \log\mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I}) - \log q(\mathbf{z}|\boldsymbol{\lambda})\right].$$

where

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

We use the results from Kotz & Nadarajah (2004).

$$\mathbb{E}_{q(z|\lambda)}\left[-\log q(\mathbf{z}|\boldsymbol{\lambda})\right] = \tfrac{1}{2}\log|\boldsymbol{\Sigma}| + \log\frac{(2a\pi)^{d/2}}{\Gamma(d/2)} + \log\frac{\Gamma(d/2)\Gamma(a)}{\Gamma(d/2+a)} + (a + d/2)\left(\psi\left(a + d/2\right) - \psi\left(a\right)\right)$$
$$\mathbb{E}_{q(z|\lambda)}\left[\log\mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})\right] = -\frac{d}{2}\log(2\pi) + \frac{d\log(\delta)}{2} - \frac{\delta}{2}\boldsymbol{\mu}^T\boldsymbol{\mu} - \frac{\delta}{2}\frac{a}{a-1}\text{Tr}\left(\boldsymbol{\Sigma}\right)$$

where $\psi(\cdot)$ is the digamma function.

The remaining thing is to compute the gradients about $\mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right]$. To compute the gradients, the reparametrization trick can be used. However, we can do better by the extended Bonnet's and Price's theorems for Student's t-distribution (Lin et al., 2019). Assuming that $f_n(\mathbf{z})$ satisfies the assumptions needed for these two theorems, we obtain the following gradient

expression:

$$\begin{aligned}\mathbf{g}_1^n &:= \nabla_\mu \mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right] = \mathbb{E}_{q(z)}\left[\nabla_z f_n(\mathbf{z})\right] \approx \nabla_z f_n(\mathbf{z}) \\ \mathbf{g}_2^n &:= 2\nabla_\Sigma \mathbb{E}_{q(z)}\left[f_n(\mathbf{z})\right] \\ &= \mathbb{E}_{q(z)}\left[u(\mathbf{z})\nabla_z^2 f_n(\mathbf{z})\right] \approx u(\mathbf{z})\nabla_z^2 f_n(\mathbf{z}) \\ &= \mathbb{E}_{q(w,z)}\left[w\nabla_z^2 f_n(\mathbf{z})\right] \approx w\nabla_z^2 f_n(\mathbf{z})\end{aligned}$$

where $\mathbf{z} \in \mathcal{R}^d$ is generated from $q(\mathbf{z})$, w is generated from $q(w)$, and

$$u(\mathbf{z}) := \frac{a + \frac{1}{2}\left(\mathbf{z}-\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right)}{(a + d/2 - 1)}$$

The gradients of $\mathcal{L}(\boldsymbol{\lambda})$ can be expressed as

$$\nabla_\mu \mathcal{L}(\boldsymbol{\lambda}) = -\sum_{n=1}^N \mathbf{g}_1^n - \delta\boldsymbol{\mu} \tag{83}$$

$$\nabla_\Sigma \mathcal{L}(\boldsymbol{\lambda}) = -\tfrac{1}{2}\sum_{n=1}^N \mathbf{g}_2^n - \frac{\delta}{2}\frac{a}{a-1}\mathbf{I} + \tfrac{1}{2}\boldsymbol{\Sigma}^{-1} \tag{84}$$

Likewise, we can obtain the variational Adam update with Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})$ as shown below, where $\mathbf{z} \in \mathcal{R}^d$, $\boldsymbol{\Sigma}^{-1} = \text{Diag}(N\mathbf{s} + \frac{\alpha\delta}{\alpha-1})$, and $u(\mathbf{z}) = \frac{a+\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})}{(a+d/2-1)}$.

Student's t ($\alpha > 1$) extension

while not converged **do**
  $\mathbf{z} \leftarrow \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$, where $w \sim \mathcal{IG}(\alpha, \alpha)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\sigma} \leftarrow \sqrt{w/(N\mathbf{s} + \frac{\alpha\delta}{\alpha-1})}$
  Randomly sample a data example $\mathcal{D}_i$
  $\mathbf{g}_\mu \leftarrow -\nabla \log p(\mathcal{D}_i|\mathbf{z})$
  option I: $\mathbf{g}_s \leftarrow w\mathbf{g}_\mu \circ \mathbf{g}_\mu$
  option II: $\mathbf{g}_s \leftarrow u(\mathbf{z})\mathbf{g}_\mu \circ \mathbf{g}_\mu$
  $\mathbf{m}_\mu \leftarrow \gamma_1 \mathbf{m}_\mu + (1-\gamma_1)\left(\mathbf{g}_\mu + \delta\boldsymbol{\mu}/N\right)$
  $\mathbf{s} \leftarrow \gamma_2 \mathbf{s} + (1-\gamma_2)\mathbf{g}_s$
  $\hat{\mathbf{m}}_\mu \leftarrow \mathbf{m}_\mu/(1-\gamma_1^t), \quad \hat{\mathbf{s}} \leftarrow (\mathbf{s} + \frac{\alpha\delta}{N(\alpha-1)})/(1-\gamma_2^t)$
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta\, \hat{\mathbf{m}}_\mu/\sqrt{\hat{\mathbf{s}}}$
  $t \leftarrow t+1$
end while

J.4. Extension for Symmetric Normal Inverse-Gaussian Distribution

Re-expressing the update (82) in terms of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (the same derivation as Khan et al. (2018)), we obtain the following update:

$$\begin{aligned}\boldsymbol{\Sigma}_{t+1}^{-1} &= \boldsymbol{\Sigma}_t^{-1} - 2\frac{\beta}{1-m}\nabla_{\Sigma_t}\mathcal{L} + \frac{m}{1-m}\left(\boldsymbol{\Sigma}_t^{-1} - \boldsymbol{\Sigma}_{t-1}^{-1}\right) \\ \boldsymbol{\mu}_{t+1} &= \boldsymbol{\mu}_t + \frac{\beta}{1-m}\boldsymbol{\Sigma}_{t+1}\nabla_{\mu_t}\mathcal{L} + \frac{m}{1-m}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t-1})\end{aligned}$$

where we use a symmetric normal inverse-Gaussian distribution with fixed $\lambda > 0$ as the variational distribution, $\nabla_{\mu_t}\mathcal{L}$ and $\nabla_{\Sigma_t}\mathcal{L}$ are defined at (78) -(79).

Likewise, we can obtain the variational Adam update with Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \delta^{-1}\mathbf{I})$ as shown below. where

$\mathbf{z} \in \mathcal{R}^d$, $\mathbf{\Sigma}^{-1} = \mathrm{Diag}(N\mathbf{s} + \delta(1+\lambda^{-1}))$, and

$$u(\mathbf{z}) = \sqrt{\frac{(\mathbf{z}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) + \lambda}{\lambda}} \frac{\mathcal{K}_{\frac{d-3}{2}}\left(\sqrt{\lambda\left((\mathbf{z}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) + \lambda\right)}\right)}{\mathcal{K}_{\frac{d-1}{2}}\left(\sqrt{\lambda\left((\mathbf{z}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) + \lambda\right)}\right)}$$

Recall that the ratio about the Bessel functions can be approximated by Eq (71) when $d \geq 2$.

Symmetric Normal Inverse-Gaussian ($\lambda > 0$) extension

while not converged **do**
  $\mathbf{z} \leftarrow \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$, where $w \sim \mathrm{InvGauss}(1, \lambda)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\sigma} \leftarrow \sqrt{1/\left[w(N\mathbf{s} + \delta(1+\lambda^{-1}))\right]}$
  Randomly sample a data example $\mathcal{D}_i$
  $\mathbf{g}_\mu \leftarrow -\nabla \log p(\mathcal{D}_i|\mathbf{z})$
  option I: $\mathbf{g}_s \leftarrow w^{-1}\mathbf{g}_\mu \circ \mathbf{g}_\mu$
  option II: $\mathbf{g}_s \leftarrow u(\mathbf{z})\mathbf{g}_\mu \circ \mathbf{g}_\mu$
  $\mathbf{m}_\mu \leftarrow \gamma_1 \mathbf{m}_\mu + (1-\gamma_1)\left(\mathbf{g}_\mu + \delta\boldsymbol{\mu}/N\right)$
  $\mathbf{s} \leftarrow \gamma_2 \mathbf{s} + (1-\gamma_2)\mathbf{g}_s$
  $\hat{\mathbf{m}}_\mu \leftarrow \mathbf{m}_\mu/(1-\gamma_1^t), \quad \hat{\mathbf{s}} \leftarrow (\mathbf{s} + \frac{\delta(1+\lambda^{-1})}{N})/(1-\gamma_2^t)$
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \beta\, \hat{\mathbf{m}}_\mu/\sqrt{\hat{\mathbf{s}}}$
  $t \leftarrow t+1$
end while