Appendices

Outline of the Appendix:

- Appendix A summarizes parameterizations and updates used in this work, which gives a road-map of the appendix.
- Appendix B contains more experimental results.
- Appendix C contains some useful results used in the remaining sections of the appendix.
- The rest of the appendix contains proofs of the claims and derivations of our update for examples summarized in Table 2 and Table 1.

A. Summary of Parameterizations Used in This Work

$q(\mathbf{w})$	Name	Our update in auxiliary space $\boldsymbol{\lambda}$
$\mathcal{N}(\mathbf{w}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (App. D.2)	Gaussian with covariance	See Eq (42)
$\mathcal{N}(\mathbf{w}\|\boldsymbol{\mu}, \mathbf{S}^{-1})$ (App. D.1)	Gaussian with precision	See Eq (38) for a full structure; See Eq (52) and (54) for a block triangular structure See Eq (55) for a block Heisenberg structure
$\mathcal{W}_p(\mathbf{W}\|\mathbf{S}, n)$ (App. E)	Wishart with precision	See Eq (44)
$\mathcal{MN}(\mathbf{W}\|\mathbf{E}, \mathbf{S}_U^{-1}, \mathbf{S}_V^{-1})$ (App. I)	Matrix Gaussian with Kronecker structure in precision form	See Eq (49)
$\frac{1}{K}\sum_{k=1}^K \mathcal{N}(\mathbf{w}\|\boldsymbol{\mu}_k, \mathbf{S}_k^{-1})$ (App. H)	Gaussian Mixture with precision	See Eq (47)
$B(w)\exp\left(\langle \mathbf{T}(w), \boldsymbol{\tau}\rangle - A(\boldsymbol{\tau})\right)$ (App. G)	Univariate Exponential Family	See Eq (45)

Table 1. Summary of our updates. See Table 2 for the parameterizations used in our updates.

$q(\mathbf{w})$	global $\boldsymbol{\tau}$	auxiliary $\boldsymbol{\lambda}$	local $\boldsymbol{\eta}$
$\mathcal{N}(\mathbf{w}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (App. D.1)	$\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\Sigma}\end{bmatrix} = \psi(\boldsymbol{\lambda}) = \begin{bmatrix}\boldsymbol{\mu}\\ \mathbf{A}\mathbf{A}^T\end{bmatrix}$	$\begin{bmatrix}\boldsymbol{\mu}\\ \mathbf{A}\end{bmatrix} = \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \begin{bmatrix}\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta}\\ \mathbf{A}_t\mathrm{Exp}(\frac{1}{2}\mathbf{M})\end{bmatrix}$	$\begin{bmatrix}\boldsymbol{\delta}\\ \mathbf{M}\end{bmatrix}$
$\mathcal{N}(\mathbf{w}\|\boldsymbol{\mu}, \mathbf{S}^{-1})$ (App. D.2)	$\begin{bmatrix}\boldsymbol{\mu}\\ \mathbf{S}\end{bmatrix} = \psi(\boldsymbol{\lambda}) = \begin{bmatrix}\boldsymbol{\mu}\\ \mathbf{B}\mathbf{B}^T\end{bmatrix}$	$\begin{bmatrix}\boldsymbol{\mu}\\ \mathbf{B}\end{bmatrix} = \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \begin{bmatrix}\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta}\\ \mathbf{B}_t\mathbf{h}(\mathbf{M})\end{bmatrix}$	$\begin{bmatrix}\boldsymbol{\delta}\\ \mathbf{M}\end{bmatrix}$
$\mathcal{W}_p(\mathbf{W}\|\mathbf{S}, n)$ (App. E)	$\begin{bmatrix}n\\ \mathbf{S}\end{bmatrix} = \psi(\boldsymbol{\lambda}) = \begin{bmatrix}2(f(b)+c)\\ 2(f(b)+c)\mathbf{B}\mathbf{B}^T\end{bmatrix}$ $c = \frac{p-1}{2},\ \ f(b) = \log(1+\exp(b))$	$\begin{bmatrix}b\\ \mathbf{B}\end{bmatrix} = \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \begin{bmatrix}b_t + \delta\\ \mathbf{B}_t\mathrm{Exp}(\mathbf{M})\end{bmatrix}$	$\begin{bmatrix}\delta\\ \mathbf{M}\end{bmatrix}$
general $q(\mathbf{w}\|\boldsymbol{\tau})$ (App. F)	$\boldsymbol{\tau} = \psi(\boldsymbol{\lambda}) = \boldsymbol{\lambda}$	$\boldsymbol{\lambda} = \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \boldsymbol{\lambda}_t + \boldsymbol{\eta}$	$\boldsymbol{\eta}$
$\mathcal{MN}(\mathbf{W}\|\mathbf{E}, \mathbf{S}_U^{-1}, \mathbf{S}_V^{-1}) =$ $\mathcal{N}(\mathrm{vec}(\mathbf{W})\|\mathrm{vec}(\mathbf{E}), \mathbf{S}_V^{-1}\otimes\mathbf{S}_U^{-1})$ Kronecker structure (App. I)	$\begin{bmatrix}\mathbf{E}\\ \mathbf{S}_V\\ \mathbf{S}_U\end{bmatrix} = \psi(\boldsymbol{\lambda}) = \begin{bmatrix}\mathbf{E}\\ \mathbf{A}\mathbf{A}^T\\ \mathbf{B}\mathbf{B}^T\end{bmatrix}$	$\begin{bmatrix}\mathbf{E}\\ \mathbf{A}\\ \mathbf{B}\end{bmatrix} = \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \begin{bmatrix}\mathbf{E}_t + \mathbf{B}_t^{-T}\boldsymbol{\Delta}\mathbf{A}_t^{-1}\\ \mathbf{A}_t\mathbf{h}(\mathbf{M})\\ \mathbf{B}_t\mathbf{h}(\mathbf{N})\end{bmatrix}$	$\begin{bmatrix}\boldsymbol{\Delta}\\ \mathbf{M}\\ \mathbf{N}\end{bmatrix}$
$\frac{1}{K}\sum_{k=1}^K \mathcal{N}(\mathbf{w}\|\boldsymbol{\mu}_k, \mathbf{S}_k^{-1})$ (App. H)	$\boldsymbol{\tau} = \begin{bmatrix}\boldsymbol{\mu}_k\\ \mathbf{S}_k\end{bmatrix}_{k=1}^K,\ \psi(\boldsymbol{\lambda}) = \{\psi_k(\boldsymbol{\lambda}_k)\}_{k=1}^K$ $\begin{bmatrix}\boldsymbol{\mu}_k\\ \mathbf{S}_k\end{bmatrix} = \psi_k(\boldsymbol{\lambda}_k) = \begin{bmatrix}\boldsymbol{\mu}_k\\ \mathbf{B}_k\mathbf{B}_k^T\end{bmatrix}$	$\boldsymbol{\lambda} = \begin{bmatrix}\boldsymbol{\mu}_k\\ \mathbf{B}_k\end{bmatrix}_{k=1}^K,\ \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \{\phi_{k,\boldsymbol{\lambda}_t}(\boldsymbol{\eta}_k)\}_{k=1}^K$ $\begin{bmatrix}\boldsymbol{\mu}_k\\ \mathbf{B}_k\end{bmatrix} = \phi_{k,\boldsymbol{\lambda}_t}(\boldsymbol{\eta}_k) = \begin{bmatrix}\boldsymbol{\mu}_{k,t} + \mathbf{B}_{k,t}^{-T}\boldsymbol{\delta}_k\\ \mathbf{B}_{k,t}\mathbf{h}(\mathbf{M}_k)\end{bmatrix}$	$\begin{bmatrix}\boldsymbol{\delta}_k\\ \mathbf{M}_k\end{bmatrix}_{k=1}^K$
univariate EF $q(w\|\boldsymbol{\tau})$ (App. G) $B(w)\exp\left(\langle \mathbf{T}(w), \boldsymbol{\tau}\rangle - A(\boldsymbol{\tau})\right)$	$\boldsymbol{\tau} = \psi(\boldsymbol{\lambda}) = f(\boldsymbol{\lambda})$ $f(\boldsymbol{\lambda}) = \log(1+\exp(\boldsymbol{\lambda}))$	$\boldsymbol{\lambda} = \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \boldsymbol{\lambda}_t + \boldsymbol{\eta}$	$\boldsymbol{\eta}$

Table 2. Summary of the parameterizations

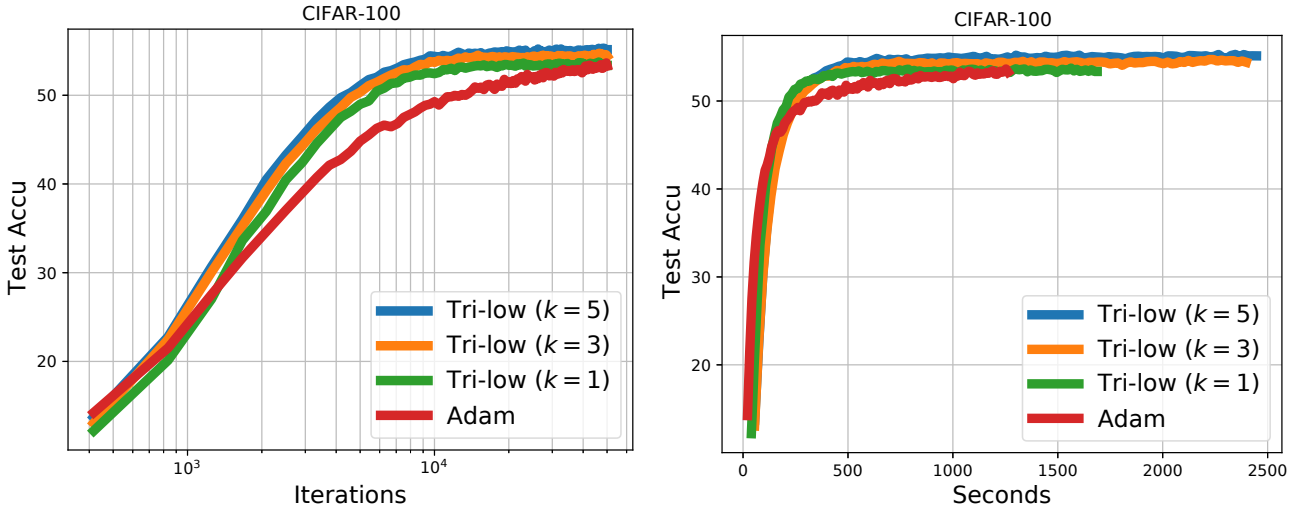
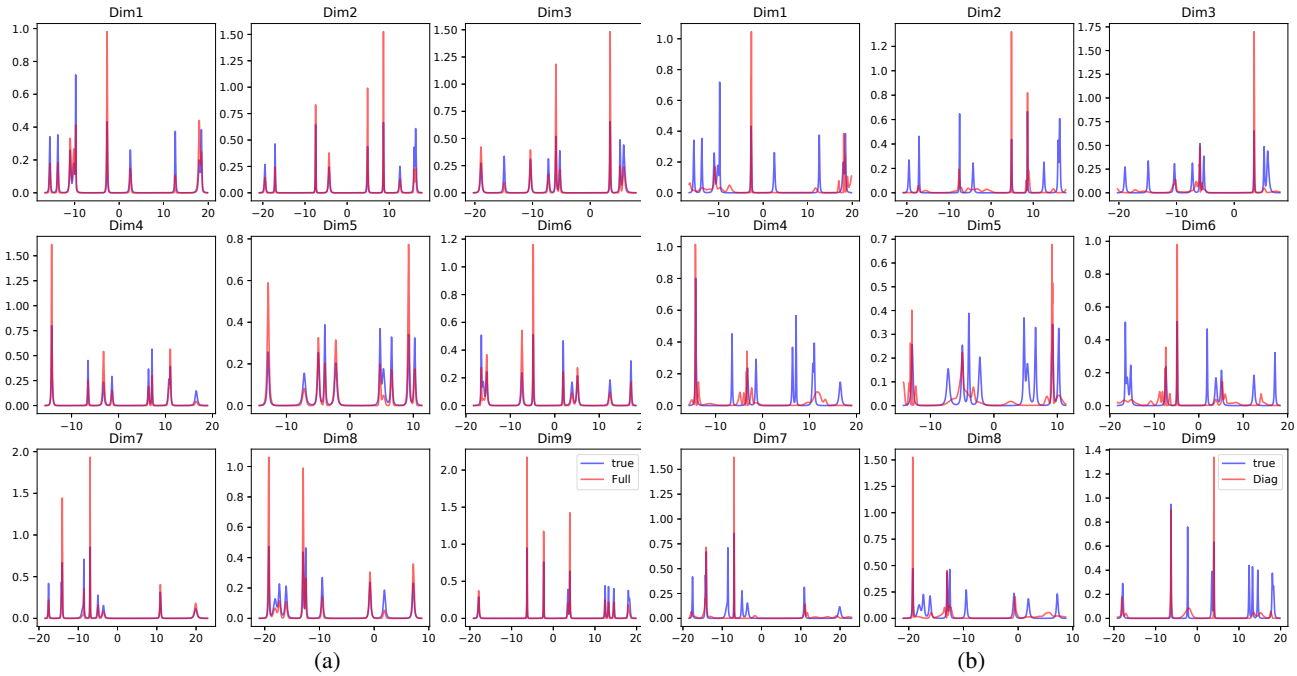


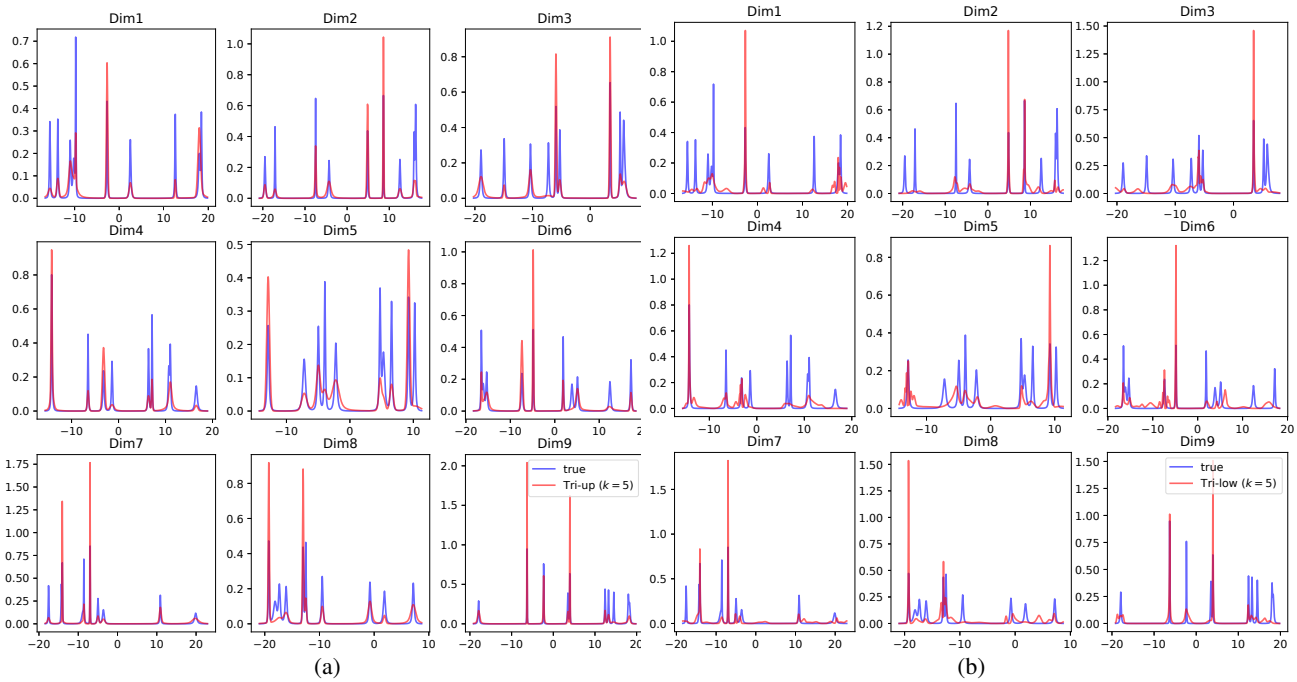
Figure 5. The performances of our updates for optimization of a CNN model on CIFAR-100 using layer-wise matrix Gaussian with low-rank structures in a Kronecker-precision form, where our updates ($O(k|\mathbf{w}|)$) have a linear iteration cost like Adam ($O(|\mathbf{w}|)$) in terms of time. For dataset "CIFAR-100", we train the model with mini-batch size 120. Our updates achieve higher test accuracy (55.2% on "CIFAR-100") than Adam (53.3% on "CIFAR-100").

B. More Results



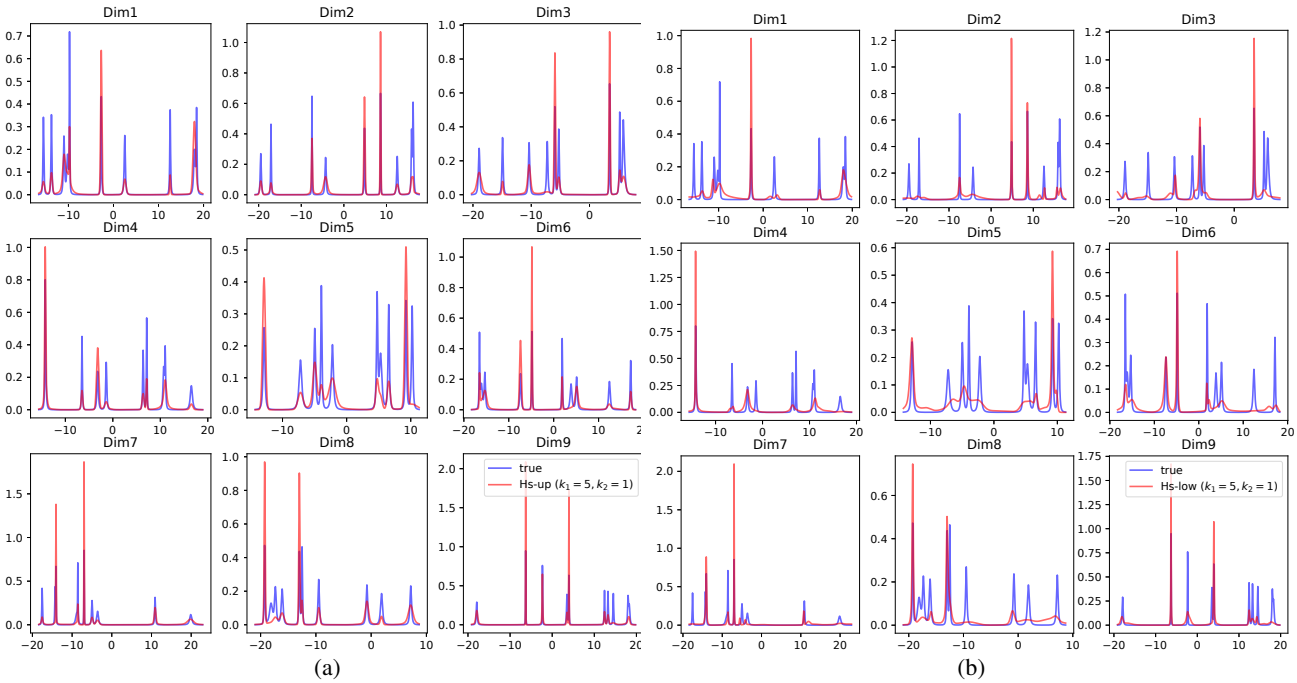
(a) (b)

Figure 6. Comparison results of structured Gaussian mixtures to fit a 80-Dim mixture of Student's t distributions with 10 components. The first 9 marginal dimensions obtained by our updates is shown in the figure, where we consider the full covariance structure and the diagonal structure.



(a) (b)

Figure 7. Comparison results of structured Gaussian mixtures to fit a 80-Dim mixture of Student's t distributions with 10 components. The first 9 marginal dimensions obtained by our updates is shown in the figure, where we consider the upper triangular structure and the lower triangular structure in the precision form. The upper triangular structure performs comparably to the full covariance structure with lower computational cost.



(a) (b)

Figure 8. Comparison results of structured Gaussian mixtures to fit a 80-Dim mixture of Student's t distributions with 10 components. The first 9 marginal dimensions obtained by our updates is shown in the figure, where we consider the upper Heisenberg structure and the lower Heisenberg structure in the precision form. The upper triangular structure performs comparably to the full covariance structure with lower computational cost.

C. Fisher information matrix and Some Useful Lemmas

The Fisher information matrix (FIM) $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau})$ of a parametric family of probability distributions $\{q_{\boldsymbol{\tau}}\}$ is expressed by $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \mathrm{Cov}_{q_{\boldsymbol{\tau}}}(\nabla_{\boldsymbol{\tau}} \log q_{\boldsymbol{\tau}}(\mathbf{w}), \nabla_{\boldsymbol{\tau}} \log q_{\boldsymbol{\tau}}(\mathbf{w}))$. Under mild regularity conditions (i.e., expectation of the score is zero and interchange of integrals with gradient operators), we have $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\nabla_{\boldsymbol{\tau}} \log q_{\boldsymbol{\tau}}(\mathbf{w})(\nabla_{\boldsymbol{\tau}} \log q_{\boldsymbol{\tau}}(\mathbf{w}))^\top\right] = -\mathbb{E}_{q_{\boldsymbol{\tau}}}[\nabla^2_{\boldsymbol{\tau}} \log q_{\boldsymbol{\tau}}(\mathbf{w})]$.

Lemma 4 *In a general case, Eq (1) can be expressed as:*

$$\mathcal{L}(\boldsymbol{\tau}) := \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\left[\ell(\mathbf{w})\right] - \gamma \mathcal{H}(q(\mathbf{w}|\boldsymbol{\tau}))$$

We have the following result:

$$\mathbf{g}_{\boldsymbol{\tau}_t} := \nabla_{\boldsymbol{\tau}} \mathcal{L}(\boldsymbol{\tau})\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t} = \nabla_{\boldsymbol{\tau}} \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\left[\ell(\mathbf{w}) + \gamma \log q(\mathbf{w}|\boldsymbol{\tau}_t)\right]\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t}$$

Therefore, we could re-define $\ell(\mathbf{w})$ to include $\gamma \log q(\mathbf{w}|\boldsymbol{\tau}_t)$ when we compute gradient $\nabla_{\boldsymbol{\tau}} \mathcal{L}(\boldsymbol{\tau})\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t}$, where $\boldsymbol{\tau}_t$ highlighted in red is considered as a constant.

The following lemma gives us an indirect approach to compute natural gradients. See Appendix G for the indirect approach and Appendix G.1 for its limitation.

Lemma 5 *(Indirect Natural-gradient Computation) If $\boldsymbol{\tau} = \boldsymbol{\psi} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta})$ is C^1-smooth w.r.t. $\boldsymbol{\eta}$, we have the following (covariant) transformation[12].*

$$\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0) = \left[\nabla_{\boldsymbol{\eta}} \boldsymbol{\tau}\right]\left[\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)\right]\left[\nabla_{\boldsymbol{\eta}} \boldsymbol{\tau}\right]^T\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}$$

where we use a layout so that $\nabla_{\eta_i} \boldsymbol{\tau}$ and $\nabla_{\boldsymbol{\eta}} \tau_j$ are a row vector and a column vector[13], respectively.

If $\hat{\boldsymbol{g}}_{\boldsymbol{\tau}_t}$ is easy to compute[14], the natural gradient $\hat{\boldsymbol{g}}_{\boldsymbol{\eta}_0}$ can be computed via the following (contravariant) transformation[15], where we assume $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)$ and the Jacobian $\left[\nabla_{\boldsymbol{\eta}} \boldsymbol{\tau}\right]\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}$ are both non-singular

$$\hat{\boldsymbol{g}}_{\boldsymbol{\eta}_0} = \left[\nabla_{\boldsymbol{\tau}} \boldsymbol{\eta}\right]^T \hat{\boldsymbol{g}}_{\boldsymbol{\tau}_t}\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t} = \left[\nabla_{\boldsymbol{\eta}} \boldsymbol{\tau}\right]^{-T} \hat{\boldsymbol{g}}_{\boldsymbol{\tau}_t}\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \tag{26}$$

where the j-th entry of the natural gradient $\hat{\boldsymbol{g}}_{\eta_0}$ can be re-expressed as $\hat{g}_{\eta_{0j}} = \sum_i \left[\nabla_{\tau_i} \eta_j\right] \hat{g}_{\tau_{ti}}$ when the Jacobian is invertible.

Therefore, $\hat{\boldsymbol{g}}_{\boldsymbol{\eta}_0}$ can be computed via a Jacobian-vector product used in forward-mode differentiation if $\hat{\boldsymbol{g}}_{\boldsymbol{\tau}_t}$ is computed beforehand and the Jacobian is invertible.

We will use the following lemmas to show that $\mathbf{h}(\cdot)$ can replace the matrix exponential map used in the main text while still keeping the natural-gradient computation tractable.

Lemma 6 *Let $\mathbf{h}(\mathbf{M}) = \mathbf{I} + \mathbf{M} + \frac{1}{2}\mathbf{M}^2$. If the matrix determinant $|\mathbf{h}(\mathbf{M})| > 0$, we have the identity:*

$$\nabla_M \log |\mathbf{h}(\mathbf{M})| = \mathbf{I} + C(\mathbf{M}),$$

where $\nabla_{M_{ij}} C(\mathbf{M})\Big|_{M=0} = \mathbf{0}$ and M_{ij} is the entry of $\mathbf{M}$ at position (i, j).

Lemma 7 *Let $\mathrm{Exp}(\mathbf{M}) := \mathbf{I} + \sum_{k=1}^{\infty} \frac{\mathbf{M}^k}{k!}$. We have a similar identity as Lemma 6:*

$$\nabla_M \log |\mathrm{Exp}(\mathbf{M})| = \mathbf{I} + C(\mathbf{M}),$$

where $\nabla_{M_{ij}} C(\mathbf{M})\Big|_{M=0} = \mathbf{0}$ and M_{ij} is the entry of $\mathbf{M}$ at position (i, j).

[12] This is the component transform for a type $(0, 2)$-tensor in Riemannian geometry.

[13] We assume $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ are vectors. For a matrix parameter, we could use the vector representation of the matrix via $\mathrm{vec}(\cdot)$.

[14] $\boldsymbol{\tau}_t$ may stay in a constrained parameter space

[15] This is the component transform for a type $(1, 0)$-tensor in Riemannian geometry.

Lemma 8 *Let* $\mathbf{f}(\mathbf{M}) = \mathbf{h}(\mathbf{M})$ *or* $\mathbf{f}(\mathbf{M}) = \mathrm{Exp}(\mathbf{M})$. *We have the following expressions:*

$$\begin{aligned}
\big[\nabla_{M_{ij}}\mathbf{f}(\mathbf{M})\big]\mathbf{f}(\mathbf{M})^T &= \big[(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M} + (\nabla_{M_{ij}}\mathbf{M})\mathbf{M}^T\big] + O(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M}) \\
\mathbf{f}(\mathbf{M})\big[\nabla_{M_{ij}}\mathbf{f}(\mathbf{M})^T\big] &= \big[(\nabla_{M_{ij}}\mathbf{M}^T) + \tfrac{1}{2}\mathbf{M}^T(\nabla_{M_{ij}}\mathbf{M}^T) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M}^T)\mathbf{M}^T + \mathbf{M}(\nabla_{M_{ij}}\mathbf{M}^T)\big] + O(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})
\end{aligned}$$

Moreover, it is obvious that $\nabla_{M_{kl}}\big[O(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})\big] = \mathbf{0}$, *where* M_{kl} *is the entry of* $\mathbf{M}$ *at position* (k, l).

C.1. Proof of Lemma 4

Proof Since $\mathcal{H}(q(\mathbf{w}|\boldsymbol{\tau})) = -\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\left[\log q(\mathbf{w}|\boldsymbol{\tau})\right]$, we can re-express $\nabla_{\boldsymbol{\tau}}\mathcal{L}(\boldsymbol{\tau})\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t}$ as

$$\begin{aligned}
\nabla_{\boldsymbol{\tau}}\mathcal{L}(\boldsymbol{\tau})\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t} &= \nabla_{\boldsymbol{\tau}}\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\left[\ell(\mathbf{w}) + \gamma\log q(\mathbf{w}|\boldsymbol{\tau})\right]\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t} \\
&= \nabla_{\boldsymbol{\tau}}\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\left[\ell(\mathbf{w}) + \gamma\log q(\mathbf{w}|\boldsymbol{\tau}_t)\right] + \gamma\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\left[\nabla_{\boldsymbol{\tau}}\log q(\mathbf{w}|\boldsymbol{\tau})\right]\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t} \quad \text{(By the chain rule)}
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\left[\nabla_{\boldsymbol{\tau}}\log q(\mathbf{w}|\boldsymbol{\tau})\right]\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t} \\
=&\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\left[\frac{\nabla_{\boldsymbol{\tau}}q(\mathbf{w}|\boldsymbol{\tau})}{q(\mathbf{w}|\boldsymbol{\tau})}\right]\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t} \\
=&\nabla_{\boldsymbol{\tau}}\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\left[1\right]\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t} \\
=&\mathbf{0}
\end{aligned} \tag{27}$$

Therefore,

$$\nabla_{\boldsymbol{\tau}}\mathcal{L}(\boldsymbol{\tau})\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t} = \nabla_{\boldsymbol{\tau}}\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\Big[\ell(\mathbf{w}) + \gamma\log q(\mathbf{w}|\overbrace{\boldsymbol{\tau}_t}^{\text{Constant}})\Big]\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t}$$

C.2. Proof of Lemma 5

Proof Let's consider an entry of the FIM $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0)$ at position (j, i).

$$\begin{aligned}
\underbrace{F_{\eta_{ji}}(\boldsymbol{\eta}_0)}_{\text{scalar}} &= \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\eta})}\big[\big[\nabla_{\eta_j}\log q(\mathbf{w}|\boldsymbol{\eta})\big]\big[\nabla_{\eta_i}\log q(\mathbf{w}|\boldsymbol{\eta})\big]\big]\big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \\
&= \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\eta})}\big[\big[\underbrace{\nabla_{\eta_j}\boldsymbol{\tau}}_{\text{row vector}}\ \underbrace{\nabla_{\boldsymbol{\tau}}\log q(\mathbf{w}|\boldsymbol{\tau})}_{\text{column vector}}\big]\big[\nabla_{\eta_i}\boldsymbol{\tau}\nabla_{\boldsymbol{\tau}}\log q(\mathbf{w}|\boldsymbol{\tau})\big]\big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \\
&= \underbrace{\big[\nabla_{\eta_j}\boldsymbol{\tau}\big]}_{\text{row vector}}\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\eta})}\big[\big[\nabla_{\boldsymbol{\tau}}\log q(\mathbf{w}|\boldsymbol{\tau})\big]\big[\nabla_{\boldsymbol{\tau}}\log q(\mathbf{w}|\boldsymbol{\tau})\big]^T\big]\underbrace{\big[\nabla_{\eta_i}\boldsymbol{\tau}\big]^T}_{\text{column vector}}\big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \\
&= \big[\nabla_{\eta_j}\boldsymbol{\tau}\big]\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau}_t)}\big[\big[\nabla_{\boldsymbol{\tau}}\log q(\mathbf{w}|\boldsymbol{\tau})\big]\big[\nabla_{\boldsymbol{\tau}}\log q(\mathbf{w}|\boldsymbol{\tau})\big]^T\big]\big[\nabla_{\eta_i}\boldsymbol{\tau}\big]^T\big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \\
&= \big[\nabla_{\eta_j}\boldsymbol{\tau}\big]\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)\big[\nabla_{\eta_i}\boldsymbol{\tau}\big]^T\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}
\end{aligned}$$

Therefore, we have $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0) = \big[\nabla_{\boldsymbol{\eta}}\boldsymbol{\tau}\big]\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)\big[\nabla_{\boldsymbol{\eta}}\boldsymbol{\tau}\big]^T\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}$.

The natural gradient $\hat{\boldsymbol{g}}_{\boldsymbol{\eta}_0}$ can be computed as follows.

$$\begin{aligned}
\hat{\boldsymbol{g}}_{\boldsymbol{\eta}_0} &= \big(\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0)\big)^{-1}\mathbf{g}_{\boldsymbol{\eta}_0}\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \\
&= \big[\nabla_{\boldsymbol{\eta}}\boldsymbol{\tau}\big]^{-T}\big(\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)\big)^{-1}\big[\nabla_{\boldsymbol{\eta}}\boldsymbol{\tau}\big]^{-1}\mathbf{g}_{\boldsymbol{\eta}_0}\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \\
&= \big[\nabla_{\boldsymbol{\tau}}\boldsymbol{\eta}\big]^{T}\big(\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)\big)^{-1}\big[\nabla_{\boldsymbol{\tau}}\boldsymbol{\eta}\big]\mathbf{g}_{\boldsymbol{\eta}_0}\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \\
&= \big[\nabla_{\boldsymbol{\tau}}\boldsymbol{\eta}\big]^{T}\big(\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)\big)^{-1}\mathbf{g}_{\boldsymbol{\tau}_t}\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0} \\
&= \big[\nabla_{\boldsymbol{\tau}}\boldsymbol{\eta}\big]^{T}\hat{\boldsymbol{g}}_{\boldsymbol{\tau}_t}\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t}
\end{aligned}$$

where $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)$ and $\nabla_{\boldsymbol{\eta}}\boldsymbol{\tau}$ are invertible by the assumption, and $\boldsymbol{\tau}_t = \psi \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}_0)$.

C.3. Proof of Lemma 6

Proof We first consider the entry M_{ij} of $\mathbf{M}$. By matrix calculus, we have the following expression.

$$\begin{aligned}
&\nabla_{M_{ij}} \log|\mathbf{h}(\mathbf{M})| \\
=&\mathrm{Tr}\big((\mathbf{h}(\mathbf{M}))^{-1}\nabla_{M_{ij}}\mathbf{h}(\mathbf{M})\big) \\
=&\mathrm{Tr}\big((\mathbf{h}(\mathbf{M}))^{-1}\big[(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M}\big]\big) \\
=&\mathrm{Tr}\big((\mathbf{h}(\mathbf{M}))^{-1}\big[\tfrac{1}{2}(\mathbf{I}+\mathbf{M})(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})(\mathbf{I}+\mathbf{M})\big]\big) \\
=&\mathrm{Tr}\big((\mathbf{h}(\mathbf{M}))^{-1}\big[\tfrac{1}{2}\mathbf{h}(\mathbf{M})(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{h}(\mathbf{M}) - \frac{1}{4}\big[\mathbf{M}^2(\nabla_{M_{ij}}\mathbf{M}) + (\nabla_{M_{ij}}\mathbf{M})\mathbf{M}^2\big]\big) \\
=&\mathrm{Tr}\big((\nabla_{M_{ij}}\mathbf{M})\big) - \frac{1}{4}\mathrm{Tr}\big((\mathbf{h}(\mathbf{M}))^{-1}\big[\mathbf{M}^2(\nabla_{M_{ij}}\mathbf{M}) + (\nabla_{M_{ij}}\mathbf{M})\mathbf{M}^2\big]\big)
\end{aligned}$$

Therefore, we can express the gradient in a matrix form.

$$\nabla_M \log|\mathbf{h}(\mathbf{M})| = \mathbf{I} - \frac{1}{4}\big(\mathbf{M}^2\big)^T\mathbf{h}(\mathbf{M})^{-T} - \frac{1}{4}\mathbf{h}(\mathbf{M})^{-T}\big(\mathbf{M}^2\big)^T$$

We will show $-\frac{1}{4}\big(\mathbf{M}^2\big)^T\mathbf{h}(\mathbf{M})^{-T} - \frac{1}{4}\mathbf{h}(\mathbf{M})^{-T}\big(\mathbf{M}^2\big)^T$ is a $C(\mathbf{M})$ function defined in our claim. We first show that

$$\nabla_{M_{ij}}\big[\mathbf{M}^2\mathbf{h}(\mathbf{M})^{-1}\big]\Big|_{M=0} = \mathbf{0}$$

By the product rule, we have

$$\begin{aligned}
&\nabla_{M_{ij}}\big[\mathbf{M}^2\mathbf{h}(\mathbf{M})^{-1}\big]\Big|_{M=0} \\
=&\big[\nabla_{M_{ij}}\mathbf{M}\big]\underbrace{\mathbf{M}}_{=\mathbf{0}}\mathbf{h}(\mathbf{M})^{-1}\Big|_{M=0} + \underbrace{\mathbf{M}}_{=\mathbf{0}}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{h}(\mathbf{M})^{-1}\Big|_{M=0} + \underbrace{\mathbf{M}^2}_{=\mathbf{0}}\big[\nabla_{M_{ij}}\mathbf{h}(\mathbf{M})^{-1}\big]\Big|_{M=0} = \mathbf{0}
\end{aligned}$$

Similarly, we can show

$$\nabla_{M_{ij}}\big[\big(\mathbf{M}^2\big)^T\mathbf{h}(\mathbf{M})^{-T}\big]\Big|_{M=0} = \mathbf{0}; \quad \nabla_{M_{ij}}\big[\mathbf{h}(\mathbf{M})^{-T}\big(\mathbf{M}^2\big)^T\big]\Big|_{M=0} = \mathbf{0}$$

Finally, we obtain the result as $\nabla_M \log|\mathbf{h}(\mathbf{M})| = \mathbf{I} + C(\mathbf{M})$, where $C(\mathbf{M}) = -\frac{1}{4}\big(\mathbf{M}^2\big)^T\mathbf{h}(\mathbf{M})^{-T} - \frac{1}{4}\mathbf{h}(\mathbf{M})^{-T}\big(\mathbf{M}^2\big)^T$

C.4. Proof of Lemma 7

Proof First of all, $|\mathrm{Exp}(\mathbf{M})| > 0$ and $(\mathrm{Exp}(\mathbf{M}))^{-1} = \mathrm{Exp}(-\mathbf{M})$. We consider the following expressions.

$$\mathrm{Exp}(-\mathbf{M}) = \mathbf{I} - \mathbf{M} + \underbrace{O(\mathbf{M}^2)}_{\text{remaining higher-order terms}}$$

$$\mathrm{Exp}(\mathbf{M}) = \mathbf{I} + \mathbf{M} + \tfrac{1}{2}\mathbf{M}^2 + \underbrace{O(\mathbf{M}^3)}_{\text{remaining higher-order terms}}$$

$$\nabla_{M_{ij}}\mathrm{Exp}(\mathbf{M}) = (\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M} + \underbrace{O(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})}_{\text{remaining higher-order terms}}$$

By matrix calculus, we have the following expression.

$$\begin{aligned}
&\nabla_{M_{ij}} \log|\mathrm{Exp}(\mathbf{M})| \\
=&\mathrm{Tr}\big(\mathrm{Exp}(-\mathbf{M})\nabla_{M_{ij}}\mathrm{Exp}(\mathbf{M})\big) \\
=&\mathrm{Tr}\big(\mathrm{Exp}(-\mathbf{M})\big[(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M} + O(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})\big]\big) \\
=&\mathrm{Tr}\big(\mathrm{Exp}(-\mathbf{M})\big[(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M} + O(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})\big]\big) \\
=&\mathrm{Tr}\big(\big(\mathbf{I} - \mathbf{M} + O(\mathbf{M}^2)\big)\big[(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M} + O(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})\big]\big) \\
=&\mathrm{Tr}\big((\nabla_{M_{ij}}\mathbf{M})\big) + \mathrm{Tr}\big(-\tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M} + O(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})\big)
\end{aligned}$$

Therefore, we have

$$\nabla_M \log|\mathrm{Exp}(\mathbf{M})| = \mathbf{I} - \tfrac{1}{2}\mathbf{M}^T + \tfrac{1}{2}\mathbf{M}^T + O(\mathbf{M}^2) = \mathbf{I} + O(\mathbf{M}^2)$$

Now, we show that the remaining $O(\mathbf{M}^2)$ term is a $C(\mathbf{M})$ function defined in our claim. Note that

$$\nabla_{M_{ij}} O(\mathbf{M}^2)\Big|_{M=0} = \mathrm{Tr}(\underbrace{O(\mathbf{M})}_{=\mathbf{0}}\big[\nabla_{M_{ij}}\mathbf{M}\big])\Big|_{M=0} = \mathbf{0}$$

where $O(\mathbf{M})$ contains at least the first order term of $\mathbf{M}$.

Therefore, the remaining $O(\mathbf{M}^2)$ term is a $C(\mathbf{M})$ function.

C.5. Proof of Lemma 8

Proof

First note that

$$\begin{aligned}
\mathbf{f}(\mathbf{M})^T &= \mathbf{I} + \mathbf{M}^T + O(\mathbf{M}^2) \\
\mathbf{f}(\mathbf{M}) &= \mathbf{I} + \mathbf{M} + \tfrac{1}{2}\mathbf{M}^2 + D(\mathbf{M}^3) \\
\nabla_{M_{ij}}\mathbf{f}(\mathbf{M}) &= (\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M} + D(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})
\end{aligned}$$

where $D(\mathbf{M}^3) = O(\mathbf{M}^3)$ and $D(\mathbf{M}^2) = O(\mathbf{M}^2)$ when $\mathbf{f}(\mathbf{M}) = \mathrm{Exp}(\mathbf{M})$ while $D(\mathbf{M}^3) = \mathbf{0}$ and $D(\mathbf{M}^2) = \mathbf{0}$ when $\mathbf{f}(\mathbf{M}) = \mathbf{h}(\mathbf{M})$.

We will show the first identity.

$$\begin{aligned}
&\big[\nabla_{M_{ij}}\mathbf{f}(\mathbf{M})\big]\mathbf{f}(\mathbf{M})^T \\
=&\big[(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M} + D(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})\big]\mathbf{f}(\mathbf{M})^T \\
=&\big[(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M} + D(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})\big]\big(\mathbf{I} + \mathbf{M}^T + O(\mathbf{M}^2)\big) \\
=&\big[(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}\mathbf{M}(\nabla_{M_{ij}}\mathbf{M}) + \tfrac{1}{2}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M} + (\nabla_{M_{ij}}\mathbf{M})\mathbf{M}^T\big] + O(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M}),
\end{aligned}$$

where $\mathbf{M}(\nabla_{M_{ij}}\mathbf{M})\mathbf{M}^T, (\nabla_{M_{ij}}\mathbf{M})\mathbf{M}\mathbf{M}^T \in O(\mathbf{M}^2)(\nabla_{M_{ij}}\mathbf{M})$.

Similarly, we can show the second expression holds.

D. Gaussian Distribution

D.1. Gaussian with square-root precision structure

Let's consider a global parameterization $\boldsymbol{\tau} = \{\boldsymbol{\mu}, \mathbf{S}\}$, where $\mathbf{S}$ is the precision and $\boldsymbol{\mu}$ is the mean. We use the following parameterizations:

$$\begin{aligned}\boldsymbol{\tau} &:= \left\{\boldsymbol{\mu} \in \mathbb{R}^p,\ \ \mathbf{S} \in \mathcal{S}_{++}^{p\times p}\right\}\\ \boldsymbol{\lambda} &:= \left\{\boldsymbol{\mu} \in \mathbb{R}^p,\ \ \mathbf{B} \in \mathcal{R}_{++}^{p\times p}\right\}\\ \boldsymbol{\eta} &:= \left\{\boldsymbol{\delta} \in \mathbb{R}^p,\ \ \mathbf{M} \in \mathcal{S}^{p\times p}\right\}.\end{aligned}$$

and maps:

$$\begin{aligned}\begin{Bmatrix}\boldsymbol{\mu}\\ \mathbf{S}\end{Bmatrix} = \boldsymbol{\psi}(\boldsymbol{\lambda}) &:= \begin{Bmatrix}\boldsymbol{\mu}\\ \mathbf{B}\mathbf{B}^\top\end{Bmatrix}\\ \begin{Bmatrix}\boldsymbol{\mu}\\ \mathbf{B}\end{Bmatrix} = \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) &:= \begin{Bmatrix}\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta}\\ \mathbf{B}_t\mathbf{h}(\mathbf{M})\end{Bmatrix}.\end{aligned}$$

Under this local parametrization, we can re-expressed the negative logarithm of the Gaussian P.D.F. as below.

$$-\log q(\mathbf{w}|\boldsymbol{\eta}) = -\log|\mathbf{B}_t\mathbf{h}(\mathbf{M})| + \tfrac{1}{2}(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})^T\mathbf{B}_t\mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T\mathbf{B}_t^T(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w}) + C$$

where C is a constant number and $\boldsymbol{\lambda}_t = \{\boldsymbol{\mu}_t, \mathbf{B}_t\}$ is the auxiliary parameterization evaluated at iteration t.

Lemma 9 *Under this local parametrization $\boldsymbol{\eta}$, $\mathbf{F}_{\boldsymbol{\eta}}$ is block diagonal with two blocks–the $\boldsymbol{\delta}$ block and the $\mathbf{M}$ block. The claim holds even when $\mathbf{M}$ is not symmetric.*

Proof Any cross term of $\mathbf{F}_{\boldsymbol{\eta}}$ between these two blocks is zero as shown below.

$$\begin{aligned}&-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}}\nabla_\delta \log q(\mathbf{w}|\boldsymbol{\eta})\right]\\ =&\mathbb{E}_{q(w|\boldsymbol{\eta})}\Big[\nabla_{M_{ij}}\Big(\mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T\mathbf{B}_t^T(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})\Big)\Big]\\ =&\nabla_{M_{ij}}\big(\mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T\big)\Big(\mathbf{B}_t^T\mathbb{E}_{q(w|\boldsymbol{\eta})}\Big[\underbrace{(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})}_{=\mathbf{0}}\Big]\Big)\\ =&\mathbf{0}\end{aligned}$$

where $\mathbb{E}_{q(w|\boldsymbol{\eta})}\big[\mathbf{w}\big] = \boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta}$ and M_{ij} denotes the element of the matrix $\mathbf{M}$ at (i,j).

Lemma 10 *The FIM w.r.t. block $\boldsymbol{\delta}$ denoted by $\mathbf{F}_\delta$ is $\mathbf{I}_\delta$ when we evaluate it at $\boldsymbol{\eta}_0 = \{\boldsymbol{\delta}_0, \mathbf{M}_0\} = \mathbf{0}$. The claim holds even when $\mathbf{M}$ is not symmetric.*

Proof

$$\begin{aligned}\mathbf{F}_\delta(\boldsymbol{\eta}_0) &= -\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_\delta^2 \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\ &= \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_\delta\Big(\mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T\mathbf{B}_t^T(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})\Big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\ &= \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_\delta\Big(\boldsymbol{\delta} + \mathbf{B}_t^T(\boldsymbol{\mu}_t - \mathbf{w})\Big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\ &= \mathbf{I}_\delta\end{aligned}$$

where we use the fact that $\mathbf{h}(\mathbf{M}) = \mathbf{I}$ when $\mathbf{M} = \mathbf{0}$ to move from step 2 to step 3.

Now, we discuss how to compute the FIM w.r.t. $\mathbf{M}$, where the following expressions hold even when $\mathbf{M}$ is not symmetric since we deliberately do not make use the symmetric constraint. The only requirement for $\mathbf{M}$ is $|\mathbf{h}(\mathbf{M})| > 0$ due to Lemma 6.

Let $\mathbf{Z} = \mathbf{B}_t^T(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})^T\mathbf{B}_t$. By matrix calculus, we have the following expression.

$$\begin{aligned}
&\tfrac{1}{2}\nabla_{M_{ij}}\left[(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})^T\mathbf{B}_t\mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T\mathbf{B}_t^T(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})\right] \\
=&\tfrac{1}{2}\nabla_{M_{ij}}\mathrm{Tr}\big(\mathbf{Z}\mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T\big) \\
=&\tfrac{1}{2}\mathrm{Tr}\big(\mathbf{Z}\big[\nabla_{M_{ij}}\mathbf{h}(\mathbf{M})\big]\mathbf{h}(\mathbf{M})^T + \mathbf{Z}\mathbf{h}(\mathbf{M})\nabla_{M_{ij}}\big[\mathbf{h}(\mathbf{M})^T\big]\big)
\end{aligned}$$

By Lemma 8, we obtain a simplified expression.

$$\begin{aligned}
&\tfrac{1}{2}\nabla_{M}\left[(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})^T\mathbf{B}_t\mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T\mathbf{B}_t^T(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})\right] \\
=&\tfrac{1}{2}\left[2\mathbf{Q} + \mathbf{Q}\mathbf{M}^T + \mathbf{M}^T\mathbf{Q} + 2\mathbf{Q}\mathbf{M}\right] + O(\mathbf{M}^2)\mathbf{Z} \\
=&\mathbf{Z} + (\mathbf{Z}\mathbf{M}^T + \mathbf{M}^T\mathbf{Z})/2 + \mathbf{Z}\mathbf{M} + O(\mathbf{M}^2)\mathbf{Z}
\end{aligned}$$

where $\mathbf{Q} = (\mathbf{Z}^T + \mathbf{Z})/2 = \mathbf{Z}$

By Lemma 6, we can re-express the gradient w.r.t. $\mathbf{M}$ as

$$-\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta}) = \underbrace{-\mathbf{I} - C(\mathbf{M})}_{-\nabla_M \log|\mathbf{h}(\mathbf{M})|} + \mathbf{Z} + (\mathbf{Z}\mathbf{M}^T + \mathbf{M}^T\mathbf{Z})/2 + \mathbf{Z}\mathbf{M} + O(\mathbf{M}^2)\mathbf{Z} \tag{28}$$

Finally, we have the following lemma to compute the FIM w.r.t. $\mathbf{M}$ (denoted by $\mathbf{F}_M$) evaluated at $\boldsymbol{\eta}_0 = \mathbf{0}$.

Lemma 11 $-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}}\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = \nabla_{M_{ij}}\Big(\mathbf{M} + \mathbf{M}^T\Big)$. *The claim holds even when* $\mathbf{M}$ *is not symmetric as long as* $|\mathbf{h}(\mathbf{M})| > 0$ *or* $|-\mathbf{h}(\mathbf{M})| > 0$.

Proof

$$\begin{aligned}
&-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}}\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}}\Big(-\mathbf{I} - C(\mathbf{M}) + \mathbf{Z} + (\mathbf{Z}\mathbf{M}^T + \mathbf{M}^T\mathbf{Z})/2 + \mathbf{Z}\mathbf{M} + O(\mathbf{M}^2)\mathbf{Z}\Big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \quad \text{(by Eq 28)} \\
=&\left[\nabla_{M_{ij}}\Big((\mathbf{M}^T + \mathbf{M}^T)/2 + \mathbf{M} + O(\mathbf{M}^2)\Big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} - \underbrace{\nabla_{M_{ij}}C(\mathbf{M})\Big|_{\boldsymbol{\eta}=\mathbf{0}}}_{=\mathbf{0}} \\
=&\nabla_{M_{ij}}\Big(\mathbf{M} + \mathbf{M}^T\Big) + O(\mathbf{M})\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&\nabla_{M_{ij}}\Big(\mathbf{M} + \mathbf{M}^T\Big)
\end{aligned} \tag{29}$$

where we use the fact that $\mathbb{E}_{q(w|\boldsymbol{\eta})}[\mathbf{Z}] = \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\mathbf{B}_t^T(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})(\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} - \mathbf{w})^T\mathbf{B}_t\right] = \mathbf{I}$ evaluated at $\boldsymbol{\eta} = \mathbf{0}$ to move from step 2 to step 3.

Now, we discuss the symmetric constraint in $\mathbf{M} \in \mathcal{S}^{p\times p}$. The constraint is essential since the FIM can be singular without a proper constraint.

D.1.1. Symmetric Constraint $\mathcal{S}^{p\times p}$ in $\mathbf{M}$

Instead of directly using the symmetric property of $\mathbf{M}$ to simplify Eq (29), we present a general approach so that we can deal with asymmetric $\mathbf{M}$ discussed in Appendix J. The key idea is to decomposition $\mathbf{M}$ as a sum of special matrices so that the FIM computation is simple. We also numerically verify the following computation of FIM by Auto-Diff.

First of all, we consider a symmetric constraint in $\mathbf{M}$. We will show that this constraint ensures the FIM is non-singular, which implies that we can use Lemma 11 in this case.

Lemma 12 *When* $\mathbf{M}$ *is symmetric,* $|\mathbf{h}(\mathbf{M})| > 0$.

Proof

$$\begin{aligned}\mathbf{h}(\mathbf{M}) &= \mathbf{I} + \mathbf{M} + \tfrac{1}{2}\mathbf{M}^2 \\ &= \tfrac{1}{2}(\mathbf{I} + (\mathbf{I}+\mathbf{M})(\mathbf{I}+\mathbf{M})) \\ &= \tfrac{1}{2}(\mathbf{I} + (\mathbf{I}+\mathbf{M})(\mathbf{I}+\mathbf{M})^T) \ \text{(since } \mathbf{M} \text{ is symmetric)} \\ &\succ \mathbf{0} \ \text{(positive-definite)}\end{aligned}$$

Therefore, $|\mathbf{h}(\mathbf{M})| > 0$.

Since $\mathbf{M}$ is symmetric, we can re-express the matrix $\mathbf{M}$ as follows.

$$\mathbf{M} = \mathbf{M}_{\text{low}} + \mathbf{M}_{\text{low}}^T + \mathbf{M}_{\text{diag}},$$

where $\mathbf{M}_{\text{low}}$ contains the lower-triangular half of $\mathbf{M}$ excluding the diagonal elements, and $\mathbf{M}_{\text{diag}}$ contains the diagonal entries of $\mathbf{M}$.

$$\mathbf{M}_{\text{low}} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ M_{21} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ M_{d1} & M_{d2} & \cdots & 0 \end{bmatrix} \quad \mathbf{M}_{\text{diag}} = \begin{bmatrix} M_{11} & 0 & \cdots & 0 \\ 0 & M_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & M_{dd} \end{bmatrix}$$

By Eq. 28 and the chain rule, we have the following expressions, where $i > j$.

$$\begin{aligned} -\nabla_{M_{\text{low}\,ij}} \log q(\mathbf{w}|\boldsymbol{\eta}) &= -\text{Tr}\big(\underbrace{\left[\nabla_{M_{\text{low}\,ij}}\mathbf{M}\right]}_{\mathbf{I}_{ij}+\mathbf{I}_{ji}} \left[\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big) \\ -\nabla_{M_{\text{diag}\,ii}} \log q(\mathbf{w}|\boldsymbol{\eta}) &= -\text{Tr}\big(\underbrace{\left[\nabla_{M_{\text{diag}\,ii}}\mathbf{M}\right]}_{\mathbf{I}_{ii}} \left[\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big) \end{aligned}$$

Therefore, we have

$$-\nabla_{M_{\text{low}}} \log q(\mathbf{w}|\boldsymbol{\eta}) = -\text{Low}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta}) + \nabla_M^T \log q(\mathbf{w}|\boldsymbol{\eta})\big) \tag{30}$$

$$-\nabla_{M_{\text{diag}}} \log q(\mathbf{w}|\boldsymbol{\eta}) = -\tfrac{1}{2}\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta}) + \nabla_M^T \log q(\mathbf{w}|\boldsymbol{\eta})\big) = -\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big) \tag{31}$$

where we define the $\text{Diag}(\cdot)$ function that returns a diagonal matrix with the same structure as $\mathbf{M}_{\text{diag}}$ and the $\text{Low}(\cdot)$ function that returns a lower-triangular matrix with the same structure as $\mathbf{M}_{\text{low}}$.

By Lemma 9, the FIM $\mathbf{F}_{\boldsymbol{\eta}}$ is block-diagonal with two blocks—the $\boldsymbol{\delta}$ block and the $\mathbf{M}$ block. We have the following lemma for $\mathbf{F}_M$

Lemma 13 *The $\mathbf{M}$ block of the FIM denoted by $\mathbf{F}_M$ is also block-diagonal with two block— the diagonal block denoted by non-zero entries in $\mathbf{M}_{diag}$, and the lower-triangular block denoted by non-zero entries in $\mathbf{M}_{low}$.*

Proof We will prove this lemma by showing any cross term of the FIM between the non-zero entries in $\mathbf{M}_{\text{low}}$ and the non-zero entries in $\mathbf{M}_{\text{diag}}$ is also zero.

Notice that we only consider non-zero entries in $\mathbf{M}_{\text{low}}$, which implies that $i > j$ in the following expression. Therefore, any

cross term can be expressed as below.

$$
\begin{aligned}
&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{low}\,ij}}\nabla_{M_{\text{diag}}}\log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = -\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{low}\,ij}}\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \text{ (by Eq. 31)}\\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\Big[\sum_{k,l}\left[\nabla_{M_{\text{low}\,ij}}M_{kl}\right]\nabla_{M_{kl}}\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\Big]\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\Big[\underbrace{\left[\nabla_{M_{\text{low}\,ij}}M_{ij}\right]}_{=1}\nabla_{M_{ij}}\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big) + \underbrace{\left[\nabla_{M_{\text{low}\,ij}}M_{ji}\right]}_{=1}\nabla_{M_{ji}}\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\Big]\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}}\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big) + \nabla_{M_{ji}}\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\
=&- \text{Diag}\big(\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}}\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta}) + \nabla_{M_{ji}}\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big)\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\
=&\text{Diag}\big(\underbrace{\nabla_{M_{ij}}(\mathbf{M}+\mathbf{M}^T)}_{\mathbf{I}_{ij}+\mathbf{I}_{ji}} + \underbrace{\nabla_{M_{ji}}(\mathbf{M}+\mathbf{M}^T)}_{\mathbf{I}_{ij}+\mathbf{I}_{ji}}\big) = \mathbf{0} \text{ (by Lemma 11)}
\end{aligned}
$$

where $M_{\text{low}ij}$ denotes the entry of M_{low} at position (i,j), we use $\mathbf{M} = \mathbf{M}_{\text{low}} + \mathbf{M}_{\text{low}}^T + \mathbf{M}_{\text{diag}}$ to move from step 2 to step 3, and obtain the last step since $i > j$ and $\text{Diag}(\mathbf{I}_{ij}) = \mathbf{0}$

To compute the FIM w.r.t a symmetric $\mathbf{M}$, we can consider the FIM w.r.t. the non-zero entries in both $\mathbf{M}_{\text{low}}$ and $\mathbf{M}_{\text{diag}}$ separately due to the block-diagonal structure of the FIM. Now, we compute the FIM w.r.t. $\mathbf{M}_{\text{diag}}$ and $\mathbf{M}_{\text{low}}$.

By the chain rule, we have

$$
\begin{aligned}
&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{diag}\,ii}}\nabla_{M_{\text{diag}}}\log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{diag}\,ii}}\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \text{ (by Eq. 31)}\\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\sum_{j,k}\left[\nabla_{M_{\text{diag}\,ii}}M_{jk}\right]\nabla_{M_{jk}}\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\underbrace{\left[\nabla_{M_{\text{diag}\,ii}}M_{ii}\right]}_{=1}\nabla_{M_{ii}}\text{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\
=&- \text{Diag}\big(\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ii}}\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big)\Big|_{\boldsymbol{\eta}=\mathbf{0}}
\end{aligned}
$$

By Lemma 11, the FIM w.r.t. $\mathbf{M}_{\text{low}}$ is

$$
- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{diag}\,ii}}\nabla_{M_{\text{diag}}}\log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = -\text{Diag}\big(\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ii}}\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big)\Big|_{\boldsymbol{\eta}=\mathbf{0}} = \text{Diag}\big(\nabla_{M_{ii}}\left(\mathbf{M}+\mathbf{M}^T\right)\big) = 2\text{Diag}(\mathbf{I}_{ii}) \tag{32}
$$

Now, we compute the FIM w.r.t. $\mathbf{M}_{\text{low}}$. By the chain rule, we have

$$
\begin{aligned}
&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{low}\,ij}}\nabla_{M_{\text{low}}}\log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}}\\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{low}\,ij}}\text{Low}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta}) + \nabla_M^T \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \text{ (by Eq. 30)}
\end{aligned}
$$

We will first consider the following term.

$$
\begin{aligned}
&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{low}\,ij}} \text{Low}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\sum_{k,l}\left[\nabla_{M_{\text{low}\,ij}} M_{kl}\right]\nabla_{M_{kl}} \text{Low}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\underbrace{\left[\nabla_{M_{\text{low}\,ij}} M_{ji}\right]}_{=1}\nabla_{M_{ji}} \text{Low}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big) + \underbrace{\left[\nabla_{M_{\text{low}\,ij}} M_{ij}\right]}_{=1}\nabla_{M_{ij}} \text{Low}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \text{Low}\big(\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ji}}\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta}) + \nabla_{M_{ij}}\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big)\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&\text{Low}\big(\underbrace{\nabla_{M_{ji}}\left[\mathbf{M}+\mathbf{M}^T\right]}_{=\mathbf{I}_{ji}+\mathbf{I}_{ij}} + \underbrace{\nabla_{M_{ij}}\left[\mathbf{M}+\mathbf{M}^T\right]}_{=\mathbf{I}_{ij}+\mathbf{I}_{ji}}\big) = 2\mathbf{I}_{ij} \quad \text{(By Lemma 11)}
\end{aligned}
$$

where we obtain the last step by Eq 29 and the fact that $\mathbf{M}$ is symmetric.

Similarly, we can show

$$
-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{low}\,ij}} \text{Low}\big(\nabla_M^T \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = 2\mathbf{I}_{ij} \tag{33}
$$

Therefore, the FIM w.r.t. $\mathbf{M}_{\text{low}}$ is

$$
-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{low}\,ij}}\nabla_{M_{\text{low}}} \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = 4\mathbf{I}_{ij} \tag{34}
$$

.

Now, we discuss how to compute the Euclidean gradients. Recall that

$$
\begin{aligned}
\boldsymbol{\mu} &= \boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} \\
\mathbf{S} &= \mathbf{B}_t \mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T \mathbf{B}_t^T
\end{aligned}
$$

Let $\mathcal{L} := \mathbb{E}_{q(\mathbf{w})}\left[\ell(\mathbf{w})\right] - \gamma\mathcal{H}(q(\mathbf{w}))$. By the chain rule, we have

$$
\begin{aligned}
\nabla_{\delta_i}\mathcal{L} &= \left[\nabla_{\delta_i}\boldsymbol{\mu}\right]^T\nabla_\mu\mathcal{L} + \text{Tr}\big(\overbrace{\left[\nabla_{\delta_i}\mathbf{S}\right]}^{=0}\nabla_S\mathcal{L}\big) \\
&= \left[\nabla_{\delta_i}\boldsymbol{\delta}\right]^T\mathbf{B}^{-1}\nabla_\mu\mathcal{L} \\
\nabla_{M_{ij}}\mathcal{L} &= \underbrace{\left[\nabla_{M_{ij}}\boldsymbol{\mu}\right]^T}_{=0}\nabla_\mu\mathcal{L} + \text{Tr}\big(\left[\nabla_{M_{ij}}\mathbf{S}\right]\nabla_S\mathcal{L}\big) \\
&= \text{Tr}\big(\left[\nabla_{M_{ij}}\mathbf{S}\right]\nabla_S\mathcal{L}\big) \\
&= -\text{Tr}\big(\left[\nabla_{M_{ij}}\mathbf{S}\right]\boldsymbol{\Sigma}\left[\nabla_\Sigma\mathcal{L}\right]\boldsymbol{\Sigma}\big) \\
&= -\text{Tr}\big(\left[\mathbf{B}\big\{\left[\nabla_{M_{ij}}\mathbf{h}(\mathbf{M})\right]\mathbf{h}(\mathbf{M})^T + \mathbf{h}(\mathbf{M})\left[\nabla_{M_{ij}}\mathbf{h}(\mathbf{M})^T\right]\big\}\mathbf{B}^T\right]\boldsymbol{\Sigma}\left[\nabla_\Sigma\mathcal{L}\right]\boldsymbol{\Sigma}\big)
\end{aligned}
$$

where $\boldsymbol{\Sigma} = \mathbf{S}^{-1}$ and we use the gradient identity $\nabla_S\mathcal{L} = -\boldsymbol{\Sigma}\left[\nabla_\Sigma\mathcal{L}\right]\boldsymbol{\Sigma}$.

Therefore, when we evaluate the gradient at $\boldsymbol{\eta}_0 = \{\boldsymbol{\delta}_0, \mathbf{M}_0\} = \mathbf{0}$, we have

$$
\begin{aligned}
\nabla_{\delta_i}\mathcal{L}\big|_{\boldsymbol{\eta}=0} &= \left[\nabla_{\delta_i}\boldsymbol{\delta}\right]^T\mathbf{B}_t^{-1}\nabla_\mu\mathcal{L} \\
\nabla_{M_{ij}}\mathcal{L}\big|_{\boldsymbol{\eta}=0} &= -\text{Tr}\big(\big[\mathbf{B}_t\big(\left[\nabla_{M_{ij}}\mathbf{h}(\mathbf{M})\right]\underbrace{\mathbf{h}(\mathbf{0})^T}_{=\mathbf{I}} + \underbrace{\mathbf{h}(\mathbf{0})}_{=\mathbf{I}}\left[\nabla_{M_{ij}}\mathbf{h}(\mathbf{M})^T\right]\big)\mathbf{B}_t^T\big]\underbrace{\boldsymbol{\Sigma}_t}_{\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}}\left[\nabla_\Sigma\mathcal{L}\right]\boldsymbol{\Sigma}_t\big) \\
&= -\text{Tr}\big(\big[\mathbf{B}_t\big(\left[\nabla_{M_{ij}}\mathbf{h}(\mathbf{M})\right] + \left[\nabla_{M_{ij}}\mathbf{h}(\mathbf{M})^T\right]\big)\mathbf{B}_t^T\big]\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\left[\nabla_\Sigma\mathcal{L}\right]\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\big) \\
&= -\text{Tr}\big(\big(\left[\nabla_{M_{ij}}\mathbf{M}\right] + \left[\nabla_{M_{ij}}\mathbf{M}^T\right]\big)\mathbf{B}_t^{-1}\left[\nabla_\Sigma\mathcal{L}\right]\mathbf{B}_t^{-T}\big) \\
&= -\text{Tr}\big(\left[\nabla_{M_{ij}}\left(\mathbf{M}+\mathbf{M}^T\right)\right]\mathbf{B}_t^{-1}\left[\nabla_\Sigma\mathcal{L}\right]\mathbf{B}_t^{-T}\big)
\end{aligned} \tag{35}
$$

where note that $\mathbf{h}(\mathbf{M}) = \mathbf{I} + \mathbf{M} + O(\mathbf{M}^2)$ and its gradient evaluated at $\boldsymbol{\eta} = \mathbf{0}$ can be simplified as

$$\nabla_{M_{ij}} \mathbf{h}(\mathbf{M})\big|_{\boldsymbol{\eta}=0} = \nabla_{M_{ij}} \mathbf{M} + \underbrace{O(\mathbf{M})}_{=\mathbf{0}} \big[\nabla_{M_{ij}} \mathbf{M}\big]\big|_{\boldsymbol{\eta}=0} = \nabla_{M_{ij}} \mathbf{M}$$

Let's denote $\mathbf{G}_M = -2\mathbf{B}_t^{-1}\big[\nabla_\Sigma \mathcal{L}\big]\mathbf{B}_t^{-T}$. Therefore, we can show that

$$\nabla_{M_{\text{diag}}} \mathcal{L}\big|_{\boldsymbol{\eta}=0} = \text{Diag}(\mathbf{G}_M); \quad \nabla_{M_{\text{low}}} \mathcal{L}\big|_{\boldsymbol{\eta}=0} = \text{Low}\big(\mathbf{G}_M + \mathbf{G}_M^T\big) = 2\text{Low}(\mathbf{G}_M)$$

The FIM is block-diagonal w.r.t. three blocks, the $\boldsymbol{\delta}$ block, the $\mathbf{M}_{\text{diag}}$ block, and the $\mathbf{M}_{\text{low}}$ block

Recall that the FIM w.r.t. $\boldsymbol{\delta}$, $\mathbf{M}_{\text{diag}}$ and $\mathbf{M}_{\text{low}}$ are $\mathbf{I}$, $2\mathbf{I}$, $4\mathbf{I}$, respectively. The above statement implies that Assumption 1 is satisfied.

The natural gradients w.r.t. $\mathbf{M}_{\text{diag}}$ and $\mathbf{M}_{\text{low}}$ are $\frac{1}{2}\text{Diag}(\mathbf{G}_M)$ and $\frac{1}{2}\text{Low}(\mathbf{G}_M)$.

Therefore, the natural gradients w.r.t. $\boldsymbol{\delta}$ and w.r.t. $\mathbf{M}$ are

$$\hat{\boldsymbol{g}}_\delta = \mathbf{B}_t^{-1}\nabla_\mu \mathcal{L}, \quad \hat{\boldsymbol{g}}_M = \tfrac{1}{2}\mathbf{G}_M = -\mathbf{B}_t^{-1}\big[\nabla_\Sigma \mathcal{L}\big]\mathbf{B}_t^{-T} \tag{36}$$

Now, we show that Assumption 2 is also satisfied. We will use the inverse function theorem to show this.

Recall that we have shown that Assumption 1 is satisfied by using the lower-triangular half (i.e., $\mathbf{M}_{\text{low}}$ and $\mathbf{M}_{\text{diag}}$) of $\mathbf{M}$ since $\mathbf{M}$ is symmetric. Let's consider the vector representation of the non-zero entries of the *lower-triangular* part of $\mathbf{M}$ denoted by $\mathbf{m}_{\text{vec}}$. We consider the following function denoted by $\text{Mat}(\mathbf{m}_{\text{vec}})$ to obtain $\mathbf{M}$ given the vector. It is easy to see that this function is linear and therefore it is C^1-smooth w.r.t. $\mathbf{m}_{\text{vec}}$. Consider the vector representation of the local parameter $\boldsymbol{\eta}_{\text{vec}} = \{\boldsymbol{\mu}, \mathbf{m}_{\text{vec}}\}$. Assumption 1 implies that the FIM $\mathbf{F}_{\boldsymbol{\eta}_{\text{vec}}}(\mathbf{0})$ is non-singular at $\boldsymbol{\eta}_0 = \mathbf{0}$.

Note that $\mathbf{S}$ is a symmetric positive-definite matrix and it can be represented by using a (lower-triangular) Cholesky factor $\mathbf{L}$ such as $\mathbf{S} = \mathbf{L}\mathbf{L}^T$. We denote the vector representation of the non-zero entries of $\mathbf{L}$ denoted by $\text{vec}(\mathbf{L})$. Moreover, the length of $\mathbf{m}_{\text{vec}}$ is the same as the length $\text{vec}(\mathbf{L})$. Indeed, this length is the (effective) degrees of freedom of the local parameter.

Now, consider a new global parameterization $\boldsymbol{\tau}_{\text{new}} = \{\boldsymbol{\mu}, \text{vec}(\mathbf{L})\}$ and the new map $\boldsymbol{\tau}_{\text{new}} = \boldsymbol{\psi}_{\text{new}} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}_{\text{vec}})$.

$$\begin{bmatrix} \boldsymbol{\mu} \\ \text{vec}(\mathbf{L}) \end{bmatrix} = \boldsymbol{\psi}_{\text{new}} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}\Big(\begin{bmatrix} \boldsymbol{\delta} \\ \mathbf{m}_{\text{vec}} \end{bmatrix} \Big) = \begin{bmatrix} \boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} \\ \text{vec}(\text{Chol}(\mathbf{B}_t\mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T\mathbf{B}_t^T)) \end{bmatrix} \tag{37}$$

where $\mathbf{M} = \text{Mat}(\mathbf{m}_{\text{vec}})$.

It is obvious that Jacobian matrix $\nabla_{\boldsymbol{\eta}_{\text{vec}}} \boldsymbol{\tau}_{\text{new}}$ is a square matrix. Moreover, since $\mathbf{S} = \mathbf{L}\mathbf{L}^T$, this new FIM under this parameterization remains the same, denoted by $\mathbf{F}_{\boldsymbol{\eta}_{\text{new}}}(\mathbf{0})$. It is non-singular at $\boldsymbol{\eta}_{\text{vec}} = \mathbf{0}$ due to Assumption 1.

By Lemma 5, we know that

$$\mathbf{F}_{\boldsymbol{\eta}_{\text{new}}}(\mathbf{0}) = \big[\nabla_{\boldsymbol{\eta}_{\text{vec}}} \boldsymbol{\tau}_{\text{new}}\big]\big[\mathbf{F}_{\boldsymbol{\tau}_{\text{new}}}(\boldsymbol{\tau}_{\text{new}_t})\big]\big[\nabla_{\boldsymbol{\eta}_{\text{vec}}} \boldsymbol{\tau}_{\text{new}}\big]^T\Big|_{\boldsymbol{\eta}_{\text{new}}=\mathbf{0}}$$

Since $\mathbf{F}_{\boldsymbol{\eta}_{\text{new}}}(\mathbf{0})$ is non-singular and the Jacobian matrix $\nabla_{\boldsymbol{\eta}_{\text{vec}}} \boldsymbol{\tau}_{\text{new}}$ is a square matrix, the Jacobian matrix is non-singular at $\boldsymbol{\eta}_{\text{vec}} = \mathbf{0}$.

Notice that the Cholesky decomposition $\text{Chol}(\mathbf{X})$ is C^1-smooth w.r.t. $\mathbf{X}$. The smoothness of the Cholesky decomposition is used by Sun et al. (2009); Salimbeni et al. (2018). We can see that this map $\boldsymbol{\tau}_{\text{new}} = \boldsymbol{\psi}_{\text{new}} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}_{\text{vec}})$ is C^1-smooth w.r.t. $\boldsymbol{\eta}_{\text{vec}}$.

By the inverse function theorem, we know that there exist a (local) inverse function of $\{\boldsymbol{\mu}, \text{vec}(\mathbf{L})\} = \boldsymbol{\psi}_{\text{new}} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}_{\text{vec}})$ at an open neighborhood of $\boldsymbol{\eta}_{\text{vec}} = \mathbf{0}$, which is also C^1-smooth.

Since $\mathbf{S} = \mathbf{L}\mathbf{L}^T$, we know that $\boldsymbol{\tau} = \{\boldsymbol{\mu}, \mathbf{S}\}$ and $\boldsymbol{\eta} = \{\boldsymbol{\delta}, \mathbf{M}\}$ are locally C^1-diffeomorphic at an open neighborhood of $\boldsymbol{\eta}_0$.

D.1.2. Connection to Newton's Method

In Eq (1), we consider the following problem.

$$\min_{q(\mathbf{w})\in\mathcal{Q}} \mathbb{E}_{q(\mathbf{w})}\left[\ell(\mathbf{w})\right] - \gamma\mathcal{H}(q(\mathbf{w}))$$

Note that we assume $\gamma = 0$ in Eq (2) for simplicity.

By Eq (36), our update in the auxiliary parameter space with step-size β is

$$\begin{aligned}
\boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t + \mathbf{B}_t^{-T}(-\beta)\mathbf{B}_t^{-1}\mathbf{g}_\mu = \boldsymbol{\mu}_t - \beta\overbrace{\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}}^{\mathbf{S}_t^{-1}}\mathbf{g}_\mu \\
\mathbf{B}_{t+1} &\leftarrow \mathbf{B}_t\mathbf{h}(\beta\mathbf{B}_t^{-1}\left[\mathbf{g}_\Sigma\right]\mathbf{B}_t^{-T})
\end{aligned} \tag{38}$$

When $\gamma \geq 0$, due to Stein's identities, we have

$$\mathbf{g}_\mu = \mathbb{E}_{q(\mathbf{w}|\mu,\Sigma)}\left[\nabla_w \ell(\mathbf{w})\right], \quad \mathbf{g}_\Sigma = \tfrac{1}{2}\left(\mathbb{E}_{q(\mathbf{w}|\mu,\Sigma)}\left[\nabla_w^2 \ell(\mathbf{w})\right] - \gamma\boldsymbol{\Sigma}^{-1}\right)$$

Let $\mathbf{G}_t = \mathbb{E}_q\left[\nabla_w^2\ell(\mathbf{w})\right] - \gamma\boldsymbol{\Sigma}_t^{-1} = \mathbb{E}_q\left[\nabla_w^2\ell(\mathbf{w})\right] - \gamma\mathbf{S}_t$

Therefore, our update in $\mathbf{S}$ is

$$\begin{aligned}
\mathbf{S}_{t+1} &= \mathbf{B}_{t+1}\mathbf{B}_{t+1}^T = \mathbf{B}_t\mathbf{h}(\beta\mathbf{B}_t^{-1}\left[\mathbf{g}_{\Sigma_t}\right]\mathbf{B}_t^{-T})\mathbf{h}(\beta\mathbf{B}_t^{-1}\left[\mathbf{g}_{\Sigma_t}\right]\mathbf{B}_t^{-T})^T\mathbf{B}_t^T \\
&= \mathbf{B}_t\left[\mathbf{I} + 2\left(\beta\mathbf{B}_t^{-1}\left[\mathbf{g}_{\Sigma_t}\right]\mathbf{B}_t^{-T}\right) + 2\left(\beta\mathbf{B}_t^{-1}\left[\mathbf{g}_{\Sigma_t}\mathbf{B}_t^{-T}\right)^2 + O(\beta^3)\right]\mathbf{B}_t^T \\
&= \mathbf{B}_t\left[\mathbf{I} + \beta\mathbf{B}_t^{-1}\mathbf{G}_t\mathbf{B}_t^{-T} + \frac{\beta^2}{2}\mathbf{B}_t^{-1}\mathbf{G}_t\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{G}_t\mathbf{B}_t^{-T} + O(\beta^3)\right]\mathbf{B}_t^T \\
&= \mathbf{S}_t + \beta\mathbf{G}_t + \frac{\beta^2}{2}\mathbf{G}_t\mathbf{S}_t^{-1}\mathbf{G}_t + O(\beta^3)
\end{aligned} \tag{39}$$

where we use the following result when $\mathbf{X}$ is symmetric

$$\mathbf{h}(\mathbf{X})\mathbf{h}(\mathbf{X})^T = \mathbf{h}(\mathbf{X})\mathbf{h}(\mathbf{X}) = (\mathbf{I} + \mathbf{X} + \tfrac{1}{2}\mathbf{X}^2)(\mathbf{I} + \mathbf{X} + \tfrac{1}{2}\mathbf{X}^2) = \mathbf{I} + 2\mathbf{X} + 2\mathbf{X}^2 + O(\mathbf{X}^3)$$

When $\gamma = 1$, we obtain the update proposed by Lin et al. (2020) if we ignore the $O(\beta^3)$ term.

$$\begin{aligned}
\mathbf{S}_{t+1} &= \mathbf{S}_t + \beta\mathbf{G}_t + \frac{\beta^2}{2}\mathbf{G}_t\mathbf{S}_t^{-1}\mathbf{G}_t + O(\beta^3) \\
&= (1-\beta)\mathbf{S}_t + \beta\mathbb{E}_q\left[\nabla_w^2\ell(\mathbf{w})\right] + \frac{\beta^2}{2}\mathbf{G}\mathbf{S}_t^{-1}\mathbf{G}_t + O(\beta^3)
\end{aligned}$$

where $\mathbf{G}_t = \mathbb{E}_q\left[\nabla_w^2\ell(\mathbf{w})\right] - \mathbf{S}_t$

D.1.3. Unconstrained $\mathbf{M}$

In Appendix D.1.1, we show that if $\mathbf{M}$ is symmetric, the FIM $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0)$ is non-singular. Unfortunately, if $\mathbf{M} \in \mathbb{R}^{p\times p}$ is unconstrained, the FIM is indeed singular. In this appendix, we consider the square-root case for the precision. It is easy to show that the following result is also true for the square-root case of the covariance discussed in Appendix D.2.

To see why the FIM is indeed singular, we will use the vector representation of $\mathbf{M}$ as $\mathbf{v} = \mathrm{vec}(\mathbf{M})$. Let's consider these two entries M_{ij} and M_{ji}, where $i \neq j$. Unlike the symmetric case, M_{ij} and M_{ji} are *distinct* parameters in the unconstrained case. In our vector representation, we use v_{k_1} and v_{k_2} to uniquely represent M_{ij} and M_{ji}, respectively, where $k_1 \neq k_2$ since $i \neq j$.

First of all, since $\mathbf{v} = \mathrm{vec}(\mathbf{M})$, we have the following identity.

$$-\nabla_{\mathbf{v}} \log q(\mathbf{w}|\boldsymbol{\eta}) = \mathrm{vec}(-\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta}))$$

Recall that FIM is block-diagonal with two blocks—the $\boldsymbol{\delta}$ block and the $\mathbf{M}$ block. To show that the FIM is singular, we will show that the $\mathbf{M}$ block contains two identical columns/rows. For simplicity, we will instead show that the FIM w.r.t. $\mathbf{v}$ contains two identical columns/rows, where $\mathbf{v}$ is the vector representation of $\mathbf{M}$.

Let's consider the following row/column of the FIM for the $\mathbf{M}$ block.

$$
\begin{aligned}
&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{v_{k_1}}\left(\nabla_{\mathbf{v}} \log q(\mathbf{w}|\boldsymbol{\eta})\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{v_{k_1}} \mathrm{vec}\left(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\sum_{l,m}\left[\nabla_{v_{k_1}} M_{lm}\right]\nabla_{M_{lm}} \mathrm{vec}\left(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\underbrace{\left[\nabla_{v_{k_1}} M_{ij}\right]}_{=1}\nabla_{M_{ij}} \mathrm{vec}\left(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}}
\end{aligned}
$$

we obtain the last step since v_{k_1} uniquely represents M_{ij}.

Similarly, we can show

$$
-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{v_{k_1}}\left(\nabla_{\mathbf{v}} \log q(\mathbf{w}|\boldsymbol{\eta})\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = -\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\underbrace{\left[\nabla_{v_{k_2}} M_{ji}\right]}_{=1}\nabla_{M_{ji}} \mathrm{vec}\left(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}}
$$

According to Eq 29, we have

$$
\begin{aligned}
&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{v_{k_1}}\left(\nabla_{\mathbf{v}} \log q(\mathbf{w}|\boldsymbol{\eta})\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\underbrace{\left[\nabla_{v_{k_1}} M_{ij}\right]}_{=1}\nabla_{M_{ij}} \mathrm{vec}\left(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathrm{vec}\left(\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}}\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\right)\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&\mathrm{vec}\left(\nabla_{M_{ij}}\left(\mathbf{M}+\mathbf{M}^T\right)\right) \\
=&\mathrm{vec}\left(\mathbf{I}_{ij}+\mathbf{I}_{ji}\right)
\end{aligned}
$$

Similarly, we have

$$
-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{v_{k_2}}\left(\nabla_{v} \log q(\mathbf{w}|\boldsymbol{\eta})\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = \mathrm{vec}\left(\nabla_{M_{ji}}\left(\mathbf{M}+\mathbf{M}^T\right)\right) = \mathrm{vec}\left(\mathbf{I}_{ji}+\mathbf{I}_{ij}\right)
$$

Therefore, the FIM of the $\mathbf{M}$ block contains two identical columns/rows and it must be singular.

D.2. Gaussian with square-root covariance structure

Let's consider a global parameterization $\boldsymbol{\tau} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, where $\boldsymbol{\Sigma}$ is the covariance and $\boldsymbol{\mu}$ is the mean. We use the following Parameterizations:

$$
\begin{aligned}
\boldsymbol{\tau} &:= \left\{\boldsymbol{\mu} \in \mathbb{R}^p,\ \ \boldsymbol{\Sigma} \in \mathcal{S}_{++}^{p\times p}\right\} \\
\boldsymbol{\lambda} &:= \left\{\boldsymbol{\mu} \in \mathbb{R}^p,\ \ \mathbf{A} \in \mathcal{R}_{++}^{p\times p}\right\} \\
\boldsymbol{\eta} &:= \left\{\boldsymbol{\delta} \in \mathbb{R}^p,\ \ \mathbf{M} \in \mathcal{S}^{p\times p}\right\}.
\end{aligned}
$$

and maps:

$$\left\{ \begin{array}{c} \boldsymbol{\mu} \\ \boldsymbol{\Sigma} \end{array} \right\} = \boldsymbol{\psi}(\boldsymbol{\lambda}) := \left\{ \begin{array}{c} \boldsymbol{\mu} \\ \mathbf{A}\mathbf{A}^\top \end{array} \right\}$$
$$\left\{ \begin{array}{c} \boldsymbol{\mu} \\ \mathbf{A} \end{array} \right\} = \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) := \left\{ \begin{array}{c} \boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} \\ \mathbf{A}_t \mathrm{Exp}(\frac{1}{2}\mathbf{M}) \end{array} \right\}.$$

Now, we will use the fact that that $\mathbf{M}$ is symmetric. Under this local parametrization, we can re-expressed the negative logarithm of the Gaussian P.D.F. as below.

$$-\log q(\mathbf{w}|\boldsymbol{\eta}) = \log|\mathbf{A}_t\mathrm{Exp}(\tfrac{1}{2}\mathbf{M})| + \tfrac{1}{2}(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w})^T\mathbf{A}_t^{-T}\mathrm{Exp}(-\mathbf{M})\mathbf{A}_t^{-1}(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w}) + C$$

where C is a constant number and $\boldsymbol{\lambda}_t = \{\boldsymbol{\mu}_t, \mathbf{A}_t\}$ is the auxiliary parameterization evaluated at iteration t.

Like Sec D.1, we can show the FIM w.r.t. $\boldsymbol{\eta}$ is block-diagonal w.r.t. two blocks— the $\boldsymbol{\delta}$ block and the $\mathbf{M}$ block.

Now, we show that the FIM w.r.t. block $\boldsymbol{\delta}$ denoted by $\mathbf{F}_\delta$ is $\mathbf{I}_\delta$ when we evaluate it at $\boldsymbol{\eta}_0 = \{\boldsymbol{\delta}_0, \mathbf{M}_0\} = \mathbf{0}$.

$$\begin{aligned}
\mathbf{F}_\delta(\boldsymbol{\eta}_0) &= -\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla^2_\delta \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
&= \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_\delta\Big(\mathrm{Exp}(-\mathbf{M})\mathbf{A}_t^{-1}(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w})\Big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
&= \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_\delta\Big(\boldsymbol{\delta} + \mathbf{A}_t^{-1}(\boldsymbol{\mu}_t - \mathbf{w})\Big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
&= \mathbf{I}_\delta
\end{aligned}$$

where we use the fact that $\mathrm{Exp}(-\mathbf{M}) = \mathbf{I}$ when $\mathbf{M} = \mathbf{0}$ to move from step 2 to step 3.

Now, we discuss how to compute the FIM w.r.t. $\mathbf{M}$, where we explicitly use the fact that $\mathbf{M}$ is symmetric.

Let $\mathbf{Z} = \mathbf{A}_t^{-1}(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w})(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w})^T\mathbf{A}_t^{-T}$. By matrix calculus, we have the following expression.

$$\begin{aligned}
&\tfrac{1}{2}\nabla_{M_{ij}}\left[(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w})^T\mathbf{A}_t^{-T}\mathrm{Exp}(-\mathbf{M})\mathbf{A}_t^{-1}(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w})\right] \\
=&\tfrac{1}{2}\nabla_{M_{ij}}\mathrm{Tr}\big(\mathbf{Z}\mathrm{Exp}(-\mathbf{M})\big) \\
=&\tfrac{1}{2}\mathrm{Tr}\big(\mathbf{Z}\nabla_{M_{ij}}(-\mathbf{M} + \tfrac{1}{2}\mathbf{M}^2 + O(\mathbf{M}^3))\big)
\end{aligned}$$

Therefore, we have

$$\tfrac{1}{2}\nabla_M\left[(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w})^T\mathbf{A}_t^{-T}\mathrm{Exp}(-\mathbf{M})\mathbf{A}_t^{-1}(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w})\right] = -\tfrac{1}{2}\mathbf{Z} + \frac{1}{4}(\mathbf{Z}\mathbf{M} + \mathbf{M}\mathbf{Z}) + O(\mathbf{M}^2)\mathbf{Z}$$

By Lemma 7, we can re-express the gradient w.r.t. $\mathbf{M}$ as

$$-\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta}) = \underbrace{\tfrac{1}{2}(\mathbf{I} + C(\mathbf{M}))}_{\nabla_M \log|\frac{1}{2}\mathrm{Exp}(\mathbf{M})|} -\tfrac{1}{2}\mathbf{Z} + \frac{1}{4}(\mathbf{Z}\mathbf{M} + \mathbf{M}\mathbf{Z}) + O(\mathbf{M}^2)\mathbf{Z} \tag{40}$$

Finally, we have the following lemma to compute the FIM w.r.t. $\mathbf{M}$ (denoted by $\mathbf{F}_M$) evaluated at $\boldsymbol{\eta}_0 = \mathbf{0}$.

Lemma 14 $-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}}\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = \frac{1}{2}\nabla_{M_{ij}}\mathbf{M}$. *The claim assumes* $\mathbf{M}$ *is symmetric.*

Proof

$$
\begin{aligned}
&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}} \nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}}\left(\tfrac{1}{2}(\mathbf{I}+C(\mathbf{M})) - \tfrac{1}{2}\mathbf{Z} + \frac{1}{4}(\mathbf{ZM}+\mathbf{MZ}) + O(\mathbf{M}^2)\mathbf{Z}\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \text{ (by Eq 40)} \\
=&\left[\nabla_{M_{ij}}\left(\tfrac{1}{2}\mathbf{M} + O(\mathbf{M}^2)\right)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} + \tfrac{1}{2}\underbrace{\nabla_{M_{ij}} C(\mathbf{M})\Big|_{\boldsymbol{\eta}=\mathbf{0}}}_{=\mathbf{0}} \\
=&\tfrac{1}{2}\nabla_{M_{ij}}\Big(\mathbf{M}\Big) + O(\mathbf{M})\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&\tfrac{1}{2}\nabla_{M_{ij}}\Big(\mathbf{M}\Big)
\end{aligned} \tag{41}
$$

where we use the fact that $\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\mathbf{Z}\right] = \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\mathbf{A}_t^{-1}(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w})(\boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} - \mathbf{w})^T\mathbf{A}_t^{-T}\right] = \mathbf{I}$ evaluated at $\boldsymbol{\eta} = \mathbf{0}$ to move from step 2 to step 3.

Therefore, $\mathbf{F}_M(\boldsymbol{\eta}_0) = \frac{1}{2}\mathbf{I}_M$.

Now, we discuss how to compute the Euclidean gradients. Recall that

$$
\begin{aligned}
\boldsymbol{\mu} &= \boldsymbol{\mu}_t + \mathbf{A}_t\boldsymbol{\delta} \\
\boldsymbol{\Sigma} &= \mathbf{A}_t \mathrm{Exp}(\mathbf{M})\mathbf{A}_t^T
\end{aligned}
$$

Let $\mathcal{L} := \mathbb{E}_{q(\mathbf{w})}\left[\ell(\mathbf{w})\right] - \gamma\mathcal{H}(q(\mathbf{w}))$. By the chain rule, we have

$$
\begin{aligned}
\nabla_{\delta_i}\mathcal{L} &= \left[\nabla_{\delta_i}\boldsymbol{\mu}\right]^T\nabla_\mu\mathcal{L} + \mathrm{Tr}\big(\overbrace{\left[\nabla_{\delta_i}\boldsymbol{\Sigma}\right]}^{=0}\nabla_\Sigma\mathcal{L}\big) \\
&= \left[\nabla_{\delta_i}\boldsymbol{\delta}\right]^T\mathbf{A}_t^T\nabla_\mu\mathcal{L} \\
\nabla_{M_{ij}}\mathcal{L} &= \underbrace{\left[\nabla_{M_{ij}}\boldsymbol{\mu}\right]^T}_{=0}\nabla_\mu\mathcal{L} + \mathrm{Tr}\big(\left[\nabla_{M_{ij}}\boldsymbol{\Sigma}\right]\nabla_\Sigma\mathcal{L}\big) \\
&= \mathrm{Tr}\big(\left[\nabla_{M_{ij}}\boldsymbol{\Sigma}\right]\nabla_\Sigma\mathcal{L}\big) \\
&= \mathrm{Tr}\big(\mathbf{A}_t\left[\nabla_{M_{ij}}\mathrm{Exp}(\mathbf{M})\right]\mathbf{A}_t^T\nabla_\Sigma\mathcal{L}\big)
\end{aligned}
$$

Therefore, when we evaluate the gradient at $\boldsymbol{\eta}_0 = \{\boldsymbol{\delta}_0, \mathbf{M}_0\} = \mathbf{0}$, we have

$$
\begin{aligned}
\nabla_{\delta_i}\mathcal{L}\big|_{\boldsymbol{\eta}=0} &= \left[\nabla_{\delta_i}\boldsymbol{\delta}\right]^T\mathbf{A}_t^T\nabla_\mu\mathcal{L} \\
\nabla_{M_{ij}}\mathcal{L}\big|_{\boldsymbol{\eta}=0} &= \mathrm{Tr}\big(\mathbf{A}_t\left[\nabla_{M_{ij}}\mathrm{Exp}(\mathbf{M})\right]\mathbf{A}_t^T\nabla_\Sigma\mathcal{L}\big)\big|_{\boldsymbol{\eta}=0} \\
&= \mathrm{Tr}\big(\mathbf{A}_t\left[\nabla_{M_{ij}}\mathbf{M}\right]\mathbf{A}_t^T\nabla_\Sigma\mathcal{L}\big)
\end{aligned}
$$

where note that $\mathrm{Exp}(\mathbf{M}) = \mathbf{I} + \mathbf{M} + O(\mathbf{M}^2)$ and its gradient evaluated at $\boldsymbol{\eta} = \mathbf{0}$ can be simplified as

$$
\nabla_{M_{ij}}\mathrm{Exp}(\mathbf{M})\big|_{\boldsymbol{\eta}=0} = \nabla_{M_{ij}}\mathbf{M} + \underbrace{O(\mathbf{M})}_{=\mathbf{0}}\left[\nabla_{M_{ij}}\mathbf{M}\right]\big|_{\boldsymbol{\eta}=0} = \nabla_{M_{ij}}\mathbf{M}
$$

Therefore,

$$
\begin{aligned}
\nabla_\delta\mathcal{L}\big|_{\boldsymbol{\eta}=0} &= \mathbf{A}_t^T\nabla_\mu\mathcal{L} \\
\nabla_{M_{ij}}\mathcal{L}\big|_{\boldsymbol{\eta}=0} &= \mathbf{A}_t^T\left[\nabla_\Sigma\mathcal{L}\right]\mathbf{A}_t
\end{aligned}
$$

Recall that the FIM w.r.t. $\boldsymbol{\delta}$ and $\mathbf{M}$ are $\mathbf{I}$ and $\frac{1}{2}\mathbf{I}$, respectively. In other words,

$$\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0) = \begin{bmatrix} \mathbf{I}_\delta & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{I}_M \end{bmatrix},$$

which implies that Assumption 1 is satisfied.

Therefore, the natural gradient w.r.t. $\boldsymbol{\delta}$ is $\hat{\boldsymbol{g}}_\delta = \mathbf{A}_t^T \nabla_\mu \mathcal{L}$. The natural-gradient w.r.t. $\mathbf{M}$ as $\hat{\boldsymbol{g}}_M = 2\mathbf{A}_t^T \big[\nabla_\Sigma \mathcal{L}\big] \mathbf{A}_t$.

Therefore, our update in the auxiliary parameter space is

$$\begin{aligned} \boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t - \beta \mathbf{S}_t^{-1} \mathbf{g}_\mu \\ \mathbf{A}_{t+1} &\leftarrow \mathbf{A}_t \mathrm{Exp}\big(-\beta \mathbf{A}_t^T \mathbf{g}_\Sigma \mathbf{A}_t\big) \end{aligned} \tag{42}$$

recall that $\mathbf{A} = \mathbf{A}_t \mathrm{Exp}\big(-\beta \frac{1}{2} \hat{\boldsymbol{g}}_M\big)$.

Now, we show that Assumption 2 is also satisfied. Since $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\} = \boldsymbol{\tau} = \boldsymbol{\psi} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\{\boldsymbol{\delta}, \mathbf{M}\})$, where $\boldsymbol{\lambda}_t = \{\boldsymbol{\mu}_t, \mathbf{A}_t\}$, we have

$$\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\Sigma} \end{bmatrix} = \boldsymbol{\psi} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}\Big(\begin{bmatrix} \boldsymbol{\delta} \\ \mathbf{M} \end{bmatrix} \Big) = \begin{bmatrix} \boldsymbol{\mu}_t + \mathbf{A}_t \boldsymbol{\delta} \\ \mathbf{A}_t \mathrm{Exp}(\mathbf{M}) \mathbf{A}_t^T \end{bmatrix}$$

It is easy to see that $\boldsymbol{\psi} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta})$ is C^1-smooth w.r.t. $\boldsymbol{\eta}$.

Since we have shown Assumption 1 is satisfied, we have $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0)$ is non-singular. By Lemma 5, we know that both $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)$ and the Jacobian matrix $\nabla_{\boldsymbol{\eta}} \boldsymbol{\tau}$ evaluated at $\boldsymbol{\eta}_0$ are non-singular. By the inverse function theorem, we know that there exist a (local) inverse function of $\boldsymbol{\psi} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta})$ at an open neighborhood of $\boldsymbol{\eta}_0$, which is also C^1-smooth.

Therefore, we know that $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\} = \boldsymbol{\tau} = \boldsymbol{\psi} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\{\boldsymbol{\delta}, \mathbf{M}\})$ is locally C^1-diffeomorphic at an open neighborhood of $\boldsymbol{\eta}_0$.

D.3. Our NG Updates for the 1-Dim Bayesian Logistic Regression

Now, we consider the following parameterization $\boldsymbol{\tau} = \{\mu \in \mathbb{R}, \log \sigma \in \mathbb{R}\}$ for a Gaussian distribution q, where σ^2 is the variance and $\sigma > 0$. The FIM under this parameterization is

$$\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & 2 \end{bmatrix}$$

The standard NGD using this (global) parameterization $\boldsymbol{\tau}$ with step-size $\beta > 0$ is

$$\mu \leftarrow \mu - \beta \sigma^2 g_\mu$$

$$\log \sigma \leftarrow \log \sigma - \beta \tfrac{1}{2} g_{\log \sigma} = \log \sigma - \beta \tfrac{1}{2} \overbrace{(2\sigma^2 g_{\sigma^2})}^{g_{\log \sigma}} = \log \sigma - \beta \sigma^2 g_{\sigma^2}$$

Recall that our local-parameter approach also includes the standard NGD as a special case shown in Appendix F. We can also similarly show that the standard NGD on parameterization $\boldsymbol{\tau} = \{\mu, \log \sigma^2\}$ obtain an equivalent update.

For our local-parameter approach, consider the following parameterizations:

$$\begin{aligned} \boldsymbol{\tau} &= \{\mu \in \mathbb{R}, \sigma^{-2} > 0\} \\ \boldsymbol{\lambda} &= \{\mu \in \mathbb{R}, b \in \mathbb{R} \setminus \{0\}\} \\ \boldsymbol{\eta} &= \{\delta \in \mathbb{R}, m \in \mathbb{R}\} \\ \begin{bmatrix} \mu \\ b \end{bmatrix} &= \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \begin{bmatrix} \mu_t + b_t^{-1} \delta \\ b_t \exp(m) \end{bmatrix} \end{aligned}$$

where $\sigma^2 = b^{-2}$ is the variance.

Our NGD update (see (38)) under these parameterizations is

$$\mu \leftarrow \mu - \beta b^{-2} g_\mu = \mu - \beta \sigma^2 g_\mu$$
$$b \leftarrow b \exp(\beta b^{-2} g_{\sigma^2}) \iff \underbrace{\log b}_{-\log \sigma} \leftarrow \log b + \beta \sigma^2 g_{\sigma^2}, \quad \text{we assume } b > 0 \text{ for } \log(b) \text{ otherwise we use } \log(-b)$$

where we use the exponential map.

Consider another set of parameterizations for our approach:

$$\boldsymbol{\tau} = \{\mu \in \mathbb{R}, \sigma^2 > 0\}$$
$$\boldsymbol{\lambda} = \{\mu \in \mathbb{R}, a \in \mathbb{R} \setminus \{0\}\}$$
$$\boldsymbol{\eta} = \{\delta \in \mathbb{R}, m \in \mathbb{R}\}$$
$$\begin{bmatrix}\mu \\ a\end{bmatrix} = \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \begin{bmatrix}\mu_t + a_t \delta \\ a_t \exp(\frac{1}{2} m)\end{bmatrix}$$

where $\sigma^2 = a^2$ and the red term $\frac{1}{2}$ appears since we use the same parameterizations as Glasmachers et al. (2010).

Our NGD update (see (12)) under these parameterizations is

$$\mu \leftarrow \mu - \beta a^2 g_\mu = \mu - \beta \sigma^2 g_\mu$$
$$a \leftarrow a \exp(-\beta a^2 g_{\sigma^2}) \iff \underbrace{\log(a)}_{\log \sigma} \leftarrow \log a - \beta \sigma^2 g_{\sigma^2}, \quad \text{we assume } a > 0 \text{ for } \log(a) \text{ otherwise we use } \log(-a)$$

Therefore, we can see our NG updates including standard NGD in global parameterization $\boldsymbol{\tau} = \{\mu, \log \sigma^2\}$ in this univariate case are all equivalent under these parameterizations and maps. We could also use map $h(\cdot)$ defined in Sec.3.5. As shown in (16), this map matches the first two order and in practice, there is no difference between these two maps in terms of performance.

For (Euclidean) gradient descent (GD), it is not invariant to these parameterizations. Let's consider a unconstrained parameterization $\{\mu, \log \sigma^2\}$. The GD update under parameterization $\{\mu, \log \sigma^2\}$ with step size $\beta > 0$ is

$$\mu \leftarrow \mu - \beta g_\mu$$
$$\log \sigma^2 \leftarrow \log \sigma^2 - \beta g_{\log \sigma^2} = \log \sigma^2 - \beta(\sigma^2 g_{\sigma^2})$$

Now, we consider another unconstrained parameterization $\{\mu, \log \sigma\}$. The GD update with parameterization $\{\mu, \log \sigma\}$ step size $\beta > 0$ is

$$\mu \leftarrow \mu - \beta g_\mu$$
$$\log \sigma \leftarrow \log \sigma - \beta g_{\log \sigma} = \log \sigma - \beta(2\sigma^2 g_{\sigma^2}) \iff \log \sigma^2 \leftarrow \log \sigma^2 - 4\beta(\sigma^2 g_{\sigma^2})$$

Clearly, GD is not invariant to the change of parameterizations and its performance depends on the parameterization even in this simple case.

D.4. Difficulties of the standard NGD involving structured covariance/precision

Before we discuss issues in structured cases, we first revisit cases with full covariance, where we have a Kronecker structure. This Kronecker structure plays a key role for computational reduction. Unfortunately, this structure could be missing in structured covariance/precision cases.

D.4.1. Cases with full covariance

Let's consider the following parameterization $\boldsymbol{\tau} = \{\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma})\}$, where $\boldsymbol{\Sigma}$ is the covariance and $\boldsymbol{\mu}$ is the mean The negative-log Gaussian distribution is $-\log q(\mathbf{w}|\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma})) = \frac{1}{2}\left[\log|\boldsymbol{\Sigma}| + \text{Tr}(\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})(\mathbf{w}-\boldsymbol{\mu})^T)\right]$. The FIM under

this parameterization is

$$\begin{aligned}
\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}) &= -\mathbb{E}_q\left[\nabla^2_{\boldsymbol{\tau}} \log q(\mathbf{w}|\boldsymbol{\tau})\right] \\
&= \mathbb{E}_q\begin{bmatrix} \boldsymbol{\Sigma}^{-1} & \nabla_{\mathrm{vec}(\Sigma)}\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu}) \\ \nabla^T_{\mathrm{vec}(\Sigma)}\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu}) & \frac{1}{2}\nabla^2_{\mathrm{vec}(\Sigma)}\left[\log|\boldsymbol{\Sigma}| + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})(\mathbf{w}-\boldsymbol{\mu})^T)\right] \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Sigma}^{-1} & \left[\nabla_{\mathrm{vec}(\Sigma)}\boldsymbol{\Sigma}^{-1}\right]\mathbb{E}_q\left[(\mathbf{w}-\boldsymbol{\mu})\right] \\ \left[\nabla^T_{\mathrm{vec}(\Sigma)}\boldsymbol{\Sigma}^{-1}\right]\mathbb{E}_q\left[(\mathbf{w}-\boldsymbol{\mu})\right] & \frac{1}{2}\left(\left[\nabla^2_{\mathrm{vec}(\Sigma)}\log|\boldsymbol{\Sigma}|\right] + \mathrm{Tr}(\left[\nabla^2_{\mathrm{vec}(\Sigma)}\boldsymbol{\Sigma}^{-1}\right]\mathbb{E}_q\left[(\mathbf{w}-\boldsymbol{\mu})(\mathbf{w}-\boldsymbol{\mu})^T\right])\right) \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathrm{Hess}(f(\boldsymbol{\Sigma})) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{F}_{\mu}(\boldsymbol{\tau}) & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_{\mathrm{vec}(\Sigma)}(\boldsymbol{\tau}) \end{bmatrix}
\end{aligned}$$

where $\mathbf{V}_0 = \mathbb{E}_q\left[(\mathbf{w}-\boldsymbol{\mu})(\mathbf{w}-\boldsymbol{\mu})^T\right] = \boldsymbol{\Sigma}$ is considered as a constant,

$$\begin{aligned}
f(\mathbf{X}) &:= \tfrac{1}{2}\left[\log|\mathbf{X}| + \mathrm{Tr}(\mathbf{X}^{-1}\mathbf{V}_0)\right] \\
\mathrm{Hess}(f(\boldsymbol{\Sigma})) &:= \nabla^2_{\mathrm{vec}(\Sigma)} f(\boldsymbol{\Sigma})
\end{aligned}$$

Similarly, let's consider another parameterization $\boldsymbol{\tau} = \{\boldsymbol{\mu}, \mathrm{vec}(\mathbf{S})\}$, where $\mathbf{S}$ is the precision. The FIM under this parameterization is

$$\begin{aligned}
\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}) &= \begin{bmatrix} \mathbf{F}_{\mu}(\boldsymbol{\tau}) & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_{\mathrm{vec}(S)}(\boldsymbol{\tau}) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \nabla^2_{\mathrm{vec}(S)} f(\mathbf{S}^{-1}) \end{bmatrix}
\end{aligned}$$

where $\mathbf{V}_0$ is a constant used in function $f(\cdot)$ defined above and the value of $\mathbf{V}_0 = \mathbf{S}^{-1}$.

Let's denote a Euclidean gradient of $\mathbb{E}_q\left[\ell(\mathbf{w})\right]$ w.r.t. $\boldsymbol{\Sigma}$ by $\mathbf{G}_{\Sigma}$, where $\ell(\mathbf{w})$ is a model loss function and $q(\mathbf{w}) := \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We also denote the corresponding natural-gradient w.r.t. $\boldsymbol{\Sigma}$ by $\hat{\mathbf{G}}_{\Sigma}$.

Since the FIM is block-diagonal, we see the FIM block for the vector form of this precision $\mathrm{vec}(\mathbf{S})$ is

$$\mathbf{F}_{\mathrm{vec}(S)}(\boldsymbol{\tau}) := \nabla^2_{\mathrm{vec}(S)} f(\mathbf{S}^{-1})$$

Note that this FIM block has a Kronecker form as $\mathbf{F}_{\mathrm{vec}(S)}(\boldsymbol{\tau}) = \frac{1}{2}(\mathbf{S}^{-1} \otimes \mathbf{S}^{-1})$ for $\mathrm{vec}(\mathbf{S})$. The natural gradient for $\mathrm{vec}(\mathbf{S})$ is

$$\mathrm{vec}(\hat{\mathbf{G}}_S) = \hat{\mathbf{g}}_{\mathrm{vec}(S)} = \big(\mathbf{F}_{\mathrm{vec}(S)}(\boldsymbol{\tau})\big)^{-1}\mathrm{vec}(\mathbf{G}_S) = 2\big(\mathbf{S} \otimes \mathbf{S}\big)\mathrm{vec}(\mathbf{G}_S)$$

where $\mathrm{vec}(\mathbf{G}_S) = \mathbf{g}_{\mathrm{vec}(S)}$ is the Euclidean gradient w.r.t. $\mathrm{vec}(\mathbf{S})$.

Exploiting the Kronecker structure, we can convert this vector form of natural-gradient in a matrix form as

$$\begin{aligned}
\mathrm{Mat}(\hat{\mathbf{g}}_{\mathrm{vec}(S)}) = \mathrm{Mat}(2\big(\mathbf{S} \otimes \mathbf{S}\big)\mathrm{vec}(\mathbf{G}_S)) &= 2\mathbf{S}\big(\mathbf{G}_S\big)\mathbf{S} \quad \text{(exploiting the Kronecker structure)} \\
&= -2\mathbf{G}_{S^{-1}} \quad \text{(using matrix calculus)} \\
&= -2\mathbf{G}_{\Sigma} \\
&= -\mathbb{E}_{q(w)}\left[\nabla^2_w \ell(\mathbf{w})\right] \quad \text{(using Stein's identity)},
\end{aligned}$$

which is the natural-gradient for the precision matrix $\mathbf{S}$.

D.4.2. Issue involving structured cases

In low-rank Gaussian cases, as an example, consider the following parameterization $\boldsymbol{\tau} = \{\boldsymbol{\mu}, \boldsymbol{\alpha}\}$ where $\boldsymbol{\alpha} := \begin{bmatrix} \mathbf{v} \\ \mathbf{d} \end{bmatrix}$ and $\boldsymbol{\Sigma} := \mathbf{v}\mathbf{v}^T + \mathrm{Diag}(\mathbf{d}^2)$. The FIM under this parameterization is

$$\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathrm{Hess}(h(\boldsymbol{\alpha})) \end{bmatrix}$$

where $\mathbf{V}_0 = \mathbf{\Sigma}$ is considered as a constant, $\mathbf{\Sigma}$ is considered as a function of $\boldsymbol{\alpha}$, and

$$h(\boldsymbol{\alpha}) := f(\mathbf{\Sigma}(\boldsymbol{\alpha}))$$
$$\mathrm{Hess}(h(\boldsymbol{\alpha})) := \nabla^2_\alpha f(\mathbf{\Sigma}(\boldsymbol{\alpha}))$$

There are several issues about NGD for structured Gaussian cases, which lead to a case-by-case derivation for structures.

- One issue is that $\mathbf{F}_\alpha(\boldsymbol{\tau})$ can be singular for an arbitrary structure as shown in Appendix J.1.6.
- A critical issue in that $\mathbf{F}_\alpha(\boldsymbol{\tau}) = \mathrm{Hess}(h(\boldsymbol{\alpha}))$ may not have a Kronecker form exploited in full Gaussian cases. Without the Kronecker from, a computational challenge is how to efficiently compute

$$\hat{\mathbf{g}}_\alpha = \big(\mathbf{F}_\alpha(\boldsymbol{\tau})\big)^{-1}\mathbf{g}_\alpha = \mathrm{Hess}(h(\boldsymbol{\alpha}))^{-1}\mathbf{g}_\alpha$$

- Note that $\mathbf{g}_\alpha$ depends on $\mathbf{G}_\Sigma = \frac{1}{2}\mathbb{E}_q\left[\nabla^2_w \ell(\mathbf{w})\right]$. If we want to make use of second-order information via Stein's identity, another computational challenge is about how to re-express $\mathrm{Hess}(h(\boldsymbol{\alpha}))^{-1}\mathbf{g}_\alpha$ in terms of $\mathbf{G}_\Sigma$ and how to efficiently compute natural-gradients for $\boldsymbol{\alpha}$ without computing the whole Hessian $\nabla^2_w \ell(\mathbf{w})$.

E. Wishart distribution with square-root precision structure

Let's consider a global parameterization $\boldsymbol{\tau} = \{\mathbf{S}, n\}$. The P.D.F. of a Wishart distribution under this parameterization is

$$q(\mathbf{W}|\boldsymbol{\tau}) = \exp\{-\tfrac{1}{2}\mathrm{Tr}(\mathbf{SW}) + \frac{n-p-1}{2}\log|\mathbf{W}| - \frac{np}{2}\log 2 + \frac{n}{2}\log|\mathbf{S}| - \log\Gamma_p(\frac{n}{2})\}$$

where $\mathbf{W}$ is a p-by-p positive-definite matrix. The parameterization constraint for Wishart distribution is $n > p-1$ and $\mathbf{S} \in \mathcal{S}^{p\times p}_{++}$, where $\mathcal{S}^{p\times p}_{++}$ denotes the set of p-by-p positive-definite matrices.

We start by specifying the parameterization,

$$\boldsymbol{\tau} := \left\{n \in \mathbb{R},\ \ \mathbf{S} \in \mathcal{S}^{p\times p}_{++} \ \mid\ n > p-1\right\},$$
$$\boldsymbol{\lambda} := \left\{b \in \mathbb{R},\ \ \mathbf{B} \in \mathcal{R}^{p\times p}_{++}\right\},$$
$$\boldsymbol{\eta} := \left\{\delta \in \mathbb{R},\ \ \mathbf{M} \in \mathcal{S}^{p\times p}\right\},$$

and their respective maps defined at $\boldsymbol{\lambda}_t := \{b_t, \mathbf{B}_t\}$

$$\begin{Bmatrix} n \\ \mathbf{S} \end{Bmatrix} = \boldsymbol{\psi}(\boldsymbol{\lambda}) := \begin{Bmatrix} 2f(b)+p-1 \\ (2f(b)+p-1)\mathbf{B}\mathbf{B}^\top \end{Bmatrix},$$
$$\begin{Bmatrix} b \\ \mathbf{B} \end{Bmatrix} = \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) := \begin{Bmatrix} b_t+\delta \\ \mathbf{B}_t\mathrm{Exp}\left(\mathbf{M}\right) \end{Bmatrix}.$$

where $f(b) = \log(1+\exp(b))$ is the soft-plus function.

For simplicity, we assume $\mathbf{M}$ is symmetric. We can also exploit structures in the Wishart case.

Under this local parameterization, we have the following result.

$$\begin{aligned}-\log q(\mathbf{W}|\boldsymbol{\eta}) = {} & (f(b_t+\delta)+c)\mathrm{Tr}(\mathbf{B}_t\mathrm{Exp}(\mathbf{M})\mathrm{Exp}(\mathbf{M})^T\mathbf{B}_t^T\mathbf{W}) - (f(b_t+\delta)-1)\log|\mathbf{W}| \\ & -(f(b_t+\delta)+c)p\log(f(b_t+\delta)+c) - 2(f(b_t+\delta)+c)(\log|\mathrm{Exp}(\mathbf{M})| + \log|\mathbf{B}_t|) \\ & +\log\Gamma_p(f(b_t+\delta)+c)\end{aligned}$$

where $c = \frac{p-1}{2}$.

Lemma 15 *Under this local parametrization $\boldsymbol{\eta}$, $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0)$ is block diagonal with two blocks–the $\boldsymbol{\delta}$ block and the $\mathbf{M}$ block.*

Proof The cross term at $\boldsymbol{\eta}_0 = \mathbf{0}$ is

$$\begin{aligned}
&-\mathbb{E}_{q(W|\boldsymbol{\eta})}\big[\nabla_\delta \nabla_M \log q(\mathbf{W}|\boldsymbol{\eta})\big]\big|_{\boldsymbol{\eta}=0} \\
=&\nabla_\delta 2(f(b_t+\delta)+c)\mathbb{E}_{q(w|\boldsymbol{\eta})}\big[\mathbf{B}_t^T\mathbf{W}\mathbf{B}_t - \mathbf{I}\big]\big|_{\boldsymbol{\eta}=0} \\
=&\nabla_\delta 2(f(b_t+\delta)+c)\underbrace{\big[\mathbf{I}-\mathbf{I}\big]}_{=\mathbf{0}}\big|_{\boldsymbol{\eta}=0} \\
=&\mathbf{0}
\end{aligned}$$

where we have the fact that $\mathbb{E}_{q(W|\boldsymbol{\eta})}\big[\mathbf{W}\big]\big|_{\boldsymbol{\eta}=0} = \mathbf{B}_t^{-T}\mathbf{B}_t^{-1}$.

Let $\mathbf{Z} = \mathbf{B}_t^T\mathbf{W}\mathbf{B}_t$ First, we consider the following result.

$$\begin{aligned}
\nabla_{M_{ij}}\mathrm{Tr}(\mathbf{B}_t\mathrm{Exp}(\mathbf{M})\mathrm{Exp}(\mathbf{M})^T\mathbf{B}_t^T\mathbf{W}) &= \nabla_{M_{ij}}\mathrm{Tr}(\mathbf{Z}\mathrm{Exp}(\mathbf{M})\mathrm{Exp}(\mathbf{M})^T) \\
&= \mathrm{Tr}\big(\mathbf{Z}\big[\nabla_{M_{ij}}\mathrm{Exp}(\mathbf{M})\big]\mathrm{Exp}(\mathbf{M})^T + \mathbf{Z}\mathrm{Exp}(\mathbf{M})\nabla_{M_{ij}}\big[\mathrm{Exp}(\mathbf{M})^T\big]\big)
\end{aligned}$$

By Lemma 8, we obtain a simplified expression.

$$\nabla_M\big[\mathbf{B}_t\mathrm{Exp}(\mathbf{M})\mathrm{Exp}(\mathbf{M})^T\mathbf{B}_t^T\mathbf{W}\big] = 2\mathbf{Z} + (\mathbf{Z}\mathbf{M}^T + \mathbf{M}^T\mathbf{Z}) + 2\mathbf{Z}\mathbf{M} + \mathbf{Z}O(\mathbf{M}^2)$$

By Lemma 7, we have

$$-\nabla_M \log|\mathrm{Exp}(\mathbf{M})| = -\mathbf{I} - C(\mathbf{M}) \tag{43}$$

Now, we can compute the FIM w.r.t. block $\mathbf{M}$ as follows. Note that we also numerically verify the following computation of FIM by Auto-Diff.

$$\begin{aligned}
&-\mathbb{E}_{q(W|\boldsymbol{\eta})}\big[\nabla_M^2 \log q(\mathbf{W}|\boldsymbol{\eta})\big]\big|_{\boldsymbol{\eta}=0} \\
=&\mathbb{E}_{q(W|\boldsymbol{\eta})}\big[(f(b_t+\delta)+c)\nabla_M\big[2\mathbf{Z} + (\mathbf{Z}\mathbf{M}^T + \mathbf{M}^T\mathbf{Z}) + 2\mathbf{Z}\mathbf{M} + \mathbf{Z}O(\mathbf{M}^2) - 2\mathbf{I} - 2C(\mathbf{M})\big]\big]\big|_{\boldsymbol{\eta}=0} \\
=&\big[(f(b_t+\delta)+c)\nabla_M\big[2\mathbf{I} + 2\mathbf{M}^T + 2\mathbf{M} - 2\mathbf{I} + O(\mathbf{M}^2)\big]\big]\big|_{\boldsymbol{\eta}=0} - 2\big[(f(b_t+\delta)+c)\big]\underbrace{\nabla_M\big[C(\mathbf{M})\big]\big|_{\boldsymbol{\eta}=0}}_{=\mathbf{0}} \\
=&\big[(f(b_t+\delta)+c)\nabla_M\big[2\mathbf{M}^T + 2\mathbf{M} + O(\mathbf{M}^2)\big]\big]\big|_{\boldsymbol{\eta}=0} \\
=&2(f(b_t)+c)\nabla_M\big(\mathbf{M}^T + \mathbf{M}\big)
\end{aligned}$$

where we use the fact that $\mathbb{E}_{q(W|\boldsymbol{\eta})}[\mathbf{Z}] = \mathbb{E}_{q(W|\boldsymbol{\eta})}\left[\mathbf{B}_t^T\mathbf{W}\mathbf{B}_t\right] = \mathbf{I}$ evaluated at $\boldsymbol{\eta} = \mathbf{0}$ to move from step 2 to step 3.

When $\mathbf{M}$ is symmetric, we have $\mathbf{F}_M(\boldsymbol{\eta}_0) = 4(f(b_t)+c)\mathbf{I} = 2n_t\mathbf{I}$.

Next, we discuss how to compute the FIM w.r.t. δ. Let $z(\delta) := \big[\mathrm{Tr}(\mathbf{B}_t\mathrm{Exp}(\mathbf{M})\mathrm{Exp}(\mathbf{M})^T\mathbf{B}_t^T\mathbf{W}) - \log|\mathbf{W}| - p\log(f(b_t+\delta)+c) - p - 2(\log|\mathrm{Exp}(\mathbf{M})| + \log|\mathbf{B}_t|) + \psi_p(f(b_t+\delta)+c)\big]$, where $\psi_p(x) := \nabla_x \log\Gamma_p(x)$ is the multivariate digamma function.

First, let's observe that

$$-\nabla_\delta \log q(\mathbf{W}|\boldsymbol{\eta}) = z(\delta)\frac{\exp(b_t+\delta)}{1+\exp(b_t+\delta)}$$

Similarly, we have

$$-\nabla_\delta^2 \log q(\mathbf{W}|\boldsymbol{\eta}) = z(\delta)\big[\nabla_\delta \frac{\exp(b_t+\delta)}{1+\exp(b_t+\delta)}\big] + \big[\nabla_\delta z(\delta)\big]\frac{\exp(b_t+\delta)}{1+\exp(b_t+\delta)}$$

Let's consider the first term in the above expression.

$$z(\delta)\big[\nabla_\delta \frac{\exp(b_t+\delta)}{1+\exp(b_t+\delta)}\big] = -\big[\nabla_\delta \log q(\mathbf{W}|\boldsymbol{\eta})\big]\frac{1+\exp(b_t+\delta)}{\exp(b_t+\delta)}\big[\nabla_\delta \frac{\exp(b_t+\delta)}{1+\exp(b_t+\delta)}\big]$$

Note that $\boldsymbol{\eta}_0 = \{\mathbf{M}_0, \delta_0\} = \mathbf{0}$. We have the following result.

$$\begin{aligned}
&\mathbb{E}_{q(W|\boldsymbol{\eta})}\big[z(\delta)\big[\nabla_\delta \frac{\exp(b_t+\delta)}{1+\exp(b_t+\delta)}\big]\big]\big|_{\boldsymbol{\eta}=0} \\
&= -\underbrace{\mathbb{E}_{q(W|\boldsymbol{\eta})}\big[\nabla_\delta \log q(\mathbf{W}|\boldsymbol{\eta})\big]\big|_{\boldsymbol{\eta}=0}}_{=0 \text{ (see Eq (27))}}\Big(\frac{1+\exp(b_t+\delta)}{\exp(b_t+\delta)}\big[\nabla_\delta \frac{\exp(b_t+\delta)}{1+\exp(b_t+\delta)}\big]\Big)\big|_{\boldsymbol{\eta}=0} \\
&= 0
\end{aligned}$$

Now, we consider the second term. Note that

$$\big[\nabla_\delta z(\delta)\big] = \frac{\exp(b_t+\delta)}{1+\exp(b_t+\delta)}\big(-\frac{p}{f(b_t+\delta)+c} + D_{\psi,p}\big(f(b_t+\delta)+c\big)\big)$$

where $D_{\psi,p}(x) = \nabla\psi_p(x)$ is the multivariate trigamma function.

Therefore, we can compute the FIM w.r.t. δ as follows.

$$\begin{aligned}
\mathbf{F}_\delta(\boldsymbol{\eta}_0) &= -\mathbb{E}_{q(W|\boldsymbol{\eta})}\big[\nabla_\delta^2 \log q(\mathbf{W}|\boldsymbol{\eta})\big]\big|_{\boldsymbol{\eta}=0} = \big(\frac{\exp(b_t)}{1+\exp(b_t)}\big)^2\big(-\frac{2p}{2f(b_t)+p-1} + D_{\psi,p}\big(f(b_t)+\frac{p-1}{2}\big)\big) \\
&= \big(\frac{\exp(b_t)}{1+\exp(b_t)}\big)^2\big(-\frac{2p}{n_t} + D_{\psi,p}\big(\frac{n_t}{2}\big)\big)
\end{aligned}$$

Now, we discuss how to compute the Euclidean gradients. First note that

$$\begin{aligned}
n &:= 2(f(b_t+\delta)+\frac{p-1}{2}) \\
\mathbf{V}^{-1} &:= \mathbf{S} = 2(f(b_t+\delta)+\frac{p-1}{2})\mathbf{B}_t\mathrm{Exp}(\mathbf{M})\mathrm{Exp}(\mathbf{M})^T\mathbf{B}_t^T
\end{aligned}$$

where we will evaluate n and $\mathbf{V}$ at $\delta = 0$ and $\mathbf{M} = 0$.

Let $\mathcal{L} := \mathbb{E}_{q(\mathbf{w})}\left[\ell(\mathbf{w})\right] - \gamma\mathcal{H}(q(\mathbf{w}))$. By the chain rule, we have

$$\begin{aligned}
\nabla_\delta\mathcal{L} &:= \mathrm{Tr}\big(\big[\nabla_V\mathcal{L}\big]\big[\nabla_\delta\mathbf{V}\big]\big) + \big[\nabla_n\mathcal{L}\big]\big[\nabla_\delta n\big] \\
\nabla_{M_{ij}}\mathcal{L} &:= \mathrm{Tr}\big(\big[\nabla_V\mathcal{L}\big]\big[\nabla_{M_{ij}}\mathbf{V}\big]\big) + \big[\nabla_n\mathcal{L}\big]\overbrace{\big[\nabla_{M_{ij}} n\big]}^{=\mathbf{0}} \\
&= \mathrm{Tr}\big(\big[\nabla_V\mathcal{L}\big]\big[\nabla_{M_{ij}}\mathbf{V}\big]\big) \\
&= -\mathrm{Tr}\big(\big[\nabla_V\mathcal{L}\big]\mathbf{V}\big[\nabla_{M_{ij}}\mathbf{V}^{-1}\big]\mathbf{V}\big)
\end{aligned}$$

Note that

$$\begin{aligned}
\nabla_\delta \mathcal{L}\big|_{\boldsymbol{\eta}=0} &:= \mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\big[\nabla_\delta \mathbf{V}\big]\big) + \big[\nabla_n \mathcal{L}\big]\big[\nabla_\delta n\big]\big|_{\boldsymbol{\eta}=0} \\
&= \frac{-1}{2(f(b_t)+\frac{p-1}{2})^2}\frac{\exp(b_t)}{1+\exp(b_t)}\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\big) + \frac{2\exp(b_t)}{1+\exp(b_t)}\big[\nabla_n \mathcal{L}\big] \\
&= \frac{2\exp(b_t)}{1+\exp(b_t)}\Big(\frac{-1}{4(f(b_t)+\frac{p-1}{2})^2}\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\big) + \big[\nabla_n \mathcal{L}\big]\Big) \\
&= \frac{2\exp(b_t)}{1+\exp(b_t)}\Big(\frac{-1}{n_t^2}\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\underbrace{\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}}_{=n_t\mathbf{V}_t}\big) + \big[\nabla_n \mathcal{L}\big]\Big) \\
&= \frac{2\exp(b_t)}{1+\exp(b_t)}\Big(-\frac{\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\mathbf{V}_t\big)}{n_t} + \big[\nabla_n \mathcal{L}\big]\Big)
\end{aligned}$$

$$\begin{aligned}
\nabla_{M_{ij}} \mathcal{L}\big|_{\boldsymbol{\eta}=0} &:= -\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\mathbf{V}\big[\nabla_{M_{ij}}\mathbf{V}^{-1}\big]\mathbf{V}\big)\big|_{\boldsymbol{\eta}=0} \\
&= -n_t\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\mathbf{V}_t\big[\mathbf{B}_t\nabla_{M_{ij}}(\mathbf{M}+\mathbf{M}^T)\mathbf{B}_t^T\big]\mathbf{V}_t\big) \\
&= -n_t\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\underbrace{n_t^{-1}\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}}_{=\mathbf{V}_t}\big[\mathbf{B}_t\nabla_{M_{ij}}(\mathbf{M}+\mathbf{M}^T)\mathbf{B}_t^T\big]\underbrace{n_t^{-1}\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}}_{=\mathbf{V}_t}\big) \\
&= -n_t^{-1}\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\mathbf{B}_t^{-T}\big[\nabla_{M_{ij}}(\mathbf{M}+\mathbf{M}^T)\big]\mathbf{B}_t^{-1}\big)
\end{aligned}$$

when $\mathbf{M}$ is symmetric, we have

$$\begin{aligned}
\nabla_M \mathcal{L}\big|_{\boldsymbol{\eta}=0} &:= -\frac{2}{n_t}\mathrm{Tr}\big(\mathbf{B}_t^{-1}\big[\nabla_V \mathcal{L}\big]\mathbf{B}_t^{-T}\big) \\
\nabla_\delta \mathcal{L}\big|_{\boldsymbol{\eta}=0} &:= \frac{2\exp(b_t)}{1+\exp(b_t)}\Big[\frac{-\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\mathbf{V}_t\big)}{n_t} + \big[\nabla_n \mathcal{L}\big]\Big]
\end{aligned}$$

where we use the fact that $\big[\nabla_V \mathcal{L}\big]$ is symmetric.

In the symmetric case, the FIM w.r.t. $\boldsymbol{\eta}$ at $\boldsymbol{\eta}_0$ is

$$\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0) = \begin{bmatrix} 2n_t\mathbf{I}_M & \mathbf{0} \\ \mathbf{0} & \big(\frac{\exp(b_t)}{1+\exp(b_t)}\big)^2\big(-\frac{2p}{n_t} + D_{\psi,p}\big(\frac{n_t}{2}\big)\big) \end{bmatrix},$$

which implies that Assumption 1 is satisfied.

The natural gradients are

$$\begin{aligned}
\hat{\mathbf{g}}_M &:= \frac{1}{2n_t}\mathbf{G} = -\frac{1}{n_t^2}\mathbf{B}_t^{-1}\big[\nabla_V \mathcal{L}\big]\mathbf{B}_t^{-T} \\
\hat{\mathbf{g}}_\delta &:= \frac{2(1+\exp(b_t))}{\exp(b_t)}\Big(-\frac{2p}{n_t} + D_{\psi,p}\big(\frac{n_t}{2}\big)\Big)^{-1}\Big[\frac{-\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\mathbf{V}_t\big)}{n_t} + \big[\nabla_n \mathcal{L}\big]\Big]
\end{aligned}$$

where $\nabla_V \mathcal{L}$ and $\nabla_n \mathcal{L}$ can be computed by the implicit reparametrization trick in the following section.

Therefore, our update with step-size β is

$$\begin{aligned}
\mathbf{B}_{t+1} &\leftarrow \mathbf{B}_t\mathrm{Exp}(0-\beta\hat{\mathbf{g}}_M) = \mathbf{B}_t\mathrm{Exp}\big(\frac{\beta}{n_t^2}\mathbf{B}_t^{-1}\big[\nabla_V \mathcal{L}\big]\mathbf{B}_t^{-T}\big) \\
b_{t+1} &\leftarrow b_t + (0-\beta\hat{\mathbf{g}}_\delta) = b_t - \frac{2\beta(1+\exp(b_t))}{\exp(b_t)}\Big(-\frac{2p}{n_t} + D_{\psi,p}\big(\frac{n_t}{2}\big)\Big)^{-1}\Big[\frac{-\mathrm{Tr}\big(\big[\nabla_V \mathcal{L}\big]\mathbf{V}_t\big)}{n_t} + \big[\nabla_n \mathcal{L}\big]\Big] \qquad (44)
\end{aligned}$$

We can similarly show that Assumption 2 is also satisfied by the inverse function theorem as discussed in Gaussian cases (see Appendix D.1) since the soft-plus function $f(b)$ and $\mathrm{Exp}(\mathbf{M})$ are both C^1-smooth.

E.1. Reparametrizable Gradients

Recall that we can generate a Wishart random variable $\mathbf{W}$ due to the Bartlett decomposition as shown below. $\mathbf{W} = \mathbf{L}\boldsymbol{\Omega}\boldsymbol{\Omega}^T\mathbf{L}^T$, where $\mathbf{L}$ is the lower-triangular Cholesky factor of $\mathbf{S}^{-1} = \mathbf{V}$ and $\boldsymbol{\Omega}$ is the random lower-triangular matrix defined according to the Bartlett decomposition as follows

$$\boldsymbol{\Omega} = \begin{bmatrix} c_1 & 0 & 0 & \cdots & 0 \\ n_{21} & c_2 & 0 & \cdots & 0 \\ n_{31} & n_{32} & c_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{d1} & n_{d2} & n_{d3} & \cdots & c_d \end{bmatrix}$$

where the square of diagonal entry c_i^2 is independently generated from Gamma distribution with shape $\frac{n-i+1}{2}$ and rate $\frac{1}{2}$, and other non-zero entries n_{ij} are independently drawn from standard normal distribution.

Let $\mathcal{L}_1 = \mathbb{E}_q\left[\ell(\mathbf{W})\right]$. According to this sampling scheme, we can clearly see that Wishart distribution is reparametrizable. The gradient w.r.t. $\mathbf{V}$ can be computed as

$$\nabla_V \mathcal{L}_1 = \mathbb{E}_{q(\Omega)}\left[\nabla_W \ell(\mathbf{W}) \nabla_V \left(\mathbf{L}\boldsymbol{\Omega}\boldsymbol{\Omega}^T\mathbf{L}^T\right)\right]$$

Since Gamma distribution is implicitly re-parametrizable, we can also compute the gradient $\nabla_n \mathcal{L}_1$ thanks to the implicit reparametrization trick (Figurnov et al., 2018; Lin et al., 2019b) for Gamma distribution.

E.2. Riemannian Gradient Descent at $\mathbf{U}$

$$\min_{Z \in \mathcal{S}_{++}^{p\times p}} \ell(\mathbf{Z})$$

Instead of optimizing $\mathbf{Z}$, we optimize $\mathbf{U} = \mathbf{Z}^{-1}$. A Riemannian gradient (Hosseini & Sra, 2015; Lin et al., 2020) in the manifold $\mathcal{S}_{++}^{p\times p}$ is $\hat{\boldsymbol{G}} = \mathbf{U}\big(\nabla_U \ell\big)\mathbf{U}$. The RGD update with retraction and step-size β_1 is

$$\mathbf{U} \leftarrow \mathbf{U} - \beta_1 \hat{\boldsymbol{G}} + \frac{\beta_1^2}{2}\hat{\boldsymbol{G}}(\mathbf{U})^{-1}\hat{\boldsymbol{G}}.$$

Due to matrix calculus, we have $\nabla_Z \ell = -\mathbf{U}\big(\nabla_U \ell\big)\mathbf{U}$. We can re-express the RGD update as

$$\mathbf{U} \leftarrow \mathbf{U} + \beta_1 \nabla_Z \ell + \frac{\beta_1^2}{2}\big[\nabla_Z \ell\big]\mathbf{U}^{-1}\big[\nabla_Z \ell\big].$$

E.3. Gradients Evaluated at the Mean

Recall that the mean of the Wishart distribution as $\mathbf{Z} = \mathbb{E}_q\left[\mathbf{W}\right] = n\mathbf{S}^{-1} = n\mathbf{V}$. We can approximate the Euclidean gradients as below.

$$\nabla_{V_{ij}} \mathbb{E}_{q(W)}\left[\ell(\mathbf{W})\right] \approx \mathrm{Tr}\big(\nabla_Z \ell(\mathbf{Z}) \nabla_{V_{ij}}\big(n\mathbf{V}\big)\big) = n\nabla_{Z_{ij}}\ell(\mathbf{Z})$$
$$\nabla_n \mathbb{E}_{q(W)}\left[\ell(\mathbf{W})\right] \approx \mathrm{Tr}\big(\nabla_Z \ell(\mathbf{Z}) \nabla_n\big(n\mathbf{V}\big)\big) = \mathrm{Tr}(\nabla_Z \ell(\mathbf{Z})\mathbf{V})$$

where $\mathbf{Z} = n\mathbf{V}$.

Therefore,

$$\mathbf{G}_{\mathbf{V}_t} \approx n_t \nabla \ell(\mathbf{Z}_t), \quad g_{n_t} \approx \mathrm{Tr}\left[\nabla \ell(\mathbf{Z}_t)\mathbf{V}_t\right]$$

F. Standard NGD is a Special Case

The standard NGD in a global parameter $\boldsymbol{\tau}$ is a special case of using a local parameter $\boldsymbol{\eta}$. For simplicity, we assume $\boldsymbol{\tau}$ is unconstrained and the FIM is non-singular for $\boldsymbol{\tau} \in \Omega_\tau$. In this case, we choose the auxiliary parameter $\boldsymbol{\lambda}$ to be the same as $\boldsymbol{\tau}$. The map $\boldsymbol{\psi} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta})$ is chosen to be

$$\boldsymbol{\tau} = \boldsymbol{\psi}(\boldsymbol{\lambda}) := \boldsymbol{\lambda}; \quad \boldsymbol{\lambda} = \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) := \boldsymbol{\lambda}_t + \boldsymbol{\eta}.$$

Theorem 1 *Let $\mathbf{F}_{\boldsymbol{\eta}}$ and $\mathbf{F}_{\boldsymbol{\tau}}$ be the FIM under the local parameter $\boldsymbol{\eta}$ and the global parameter $\boldsymbol{\tau}$, respectively.*

$$\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0) = \mathbf{F}_{\boldsymbol{\eta}}(\mathbf{0}) = \mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)$$

It is obvious that Assumption 2 is satisfied since the map is linear. Since $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)$ is non-singular, we know that Assumption 1 is satisfied due to Theorem 1. Since $\boldsymbol{\tau} = \boldsymbol{\psi} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \boldsymbol{\lambda}_t + \boldsymbol{\eta}$, by the chain rule, we have $\mathbf{g}_{\boldsymbol{\eta}_0} = \big[\nabla_{\boldsymbol{\eta}} \boldsymbol{\tau}\big] \mathbf{g}_{\boldsymbol{\tau}_t} = \mathbf{g}_{\boldsymbol{\tau}_t}$

Therefore, the NGD update with step-size β in this local parameterization is

$$\boldsymbol{\eta}^{\text{new}} = \mathbf{0} - \beta \mathbf{F}_{\boldsymbol{\eta}}(\mathbf{0})^{-1} \mathbf{g}_{\boldsymbol{\eta}_0} = -\beta \mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)^{-1} \mathbf{g}_{\boldsymbol{\tau}_t}$$

Finally, we re-express the update in the global parameter as:

$$\boldsymbol{\tau}_{t+1} = \boldsymbol{\psi} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}^{\text{new}}) = \boldsymbol{\tau}_t + \boldsymbol{\eta}^{\text{new}} = \boldsymbol{\tau}_t - \beta \mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)^{-1} \mathbf{g}_{\boldsymbol{\tau}_t}$$

which is exactly the NGD update in $\boldsymbol{\tau}$.

F.1. Proof of theorem 1

Note that $\boldsymbol{\tau} = \psi \circ \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \boldsymbol{\lambda}_t + \boldsymbol{\eta} = \boldsymbol{\tau}_t + \boldsymbol{\eta}$. Now, we will show that the FIM under the local parameter $\boldsymbol{\eta}$ can be computed as

$$\begin{aligned}
\mathbf{F}_{\boldsymbol{\eta}}(\mathbf{0}) &= -\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\eta})}\big[\nabla^2_{\boldsymbol{\eta}} \log q(\mathbf{w}|\boldsymbol{\eta})\big]\big|_{\boldsymbol{\eta}=0} \\
&= -\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\eta})}\big[\nabla_{\boldsymbol{\eta}}\big[\underbrace{\nabla_{\boldsymbol{\eta}} \boldsymbol{\tau}}_{\mathbf{I}} \nabla_{\boldsymbol{\tau}} \log q(\mathbf{w}|\boldsymbol{\tau})\big]\big]\big|_{\boldsymbol{\eta}=0} \\
&= -\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\eta})}\big[\nabla_{\boldsymbol{\eta}}\big[\nabla_{\boldsymbol{\tau}} \log q(\mathbf{w}|\boldsymbol{\tau})\big]\big]\big|_{\boldsymbol{\eta}=0} \\
&= -\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\eta})}\big[\big[\nabla_{\boldsymbol{\eta}} \boldsymbol{\tau}\big] \nabla_{\boldsymbol{\tau}}\big[\nabla_{\boldsymbol{\tau}} \log q(\mathbf{w}|\boldsymbol{\tau})\big]\big]\big|_{\boldsymbol{\eta}=0} \\
&= -\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\eta})}\big[\nabla_{\boldsymbol{\tau}}\big[\nabla_{\boldsymbol{\tau}} \log q(\mathbf{w}|\boldsymbol{\tau})\big]\big]\big|_{\boldsymbol{\eta}=0} \\
&= -\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\big[\nabla_{\boldsymbol{\tau}}\big[\nabla_{\boldsymbol{\tau}} \log q(\mathbf{w}|\boldsymbol{\tau})\big]\big]\big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_t} \\
&= \mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)
\end{aligned}$$

G. Univariate Minimal Exponential Family Distributions

Using Lemma 5, we can generalize the indirect method of Salimbeni et al. (2018) to compute natural-gradients for univariate *minimal* EF distributions using a local parameterization. Salimbeni et al. (2018) only consider the method for multivariate Gaussian cases using a global parameterization.

Note that the main issue to perform the standard NGD update in the global parameter space is that the NGD update in $\boldsymbol{\tau}$ may violate a parameter constraint. However, we can perform a NGD update in an unconstrained space (e.g., the auxiliary space of $\boldsymbol{\lambda}$) if the natural gradient computation in the space of unconstrained space of $\boldsymbol{\lambda}$ is simple. Salimbeni et al. (2018) suggest using the indirect method to compute natural gradients via Auto-Differentiation (Auto-Diff).

For univariate minimal EF distributions, we can also use this indirect method to compute natural gradients. We consider a class of univariate EF distributions. We make the following assumptions for the class of distributions: (A) Each distribution in the class contains separable natural parameter blocks so that each parameter constraint only appears once in a block and each block only contains a scalar parameter. (B) The natural gradient w.r.t. the natural parameterization is easy to compute.

We choose the natural parameterization as a global parameterization $\boldsymbol{\tau}$ with K blocks: $q(w|\boldsymbol{\tau}) = B(w)\exp(\langle \mathbf{T}(w), \boldsymbol{\tau}\rangle - A(\boldsymbol{\tau}))$, where $B(w)$ is the base measure, $A(\boldsymbol{\tau})$ is the log partition function[16], and $\mathbf{T}(w)$ is the sufficient statistics. A common parameter constraint in $\boldsymbol{\tau}$ is the scalar positivity constraint denoted by $\mathcal{S}^1_{++}$. For simplicity, we assume $\mathcal{S}^1_{++}$ is the only parameter constraint. Common univariate EF distributions such as Bernoulli, exponential, Pareto, Weibull, Laplace, Wald, univariate Gaussian, Beta, and Gamma distribution all satisfy Assumption A. Assumption B is also valid for these univariate EF distributions since we can either compute the natural gradient $\hat{\boldsymbol{g}}_{\boldsymbol{\tau}_t}$ via the Euclidean gradient w.r.t. the expectation parameter (Khan & Lin, 2017) or use the direct natural gradient computation when K is small ($K < 3$ in common cases).

Given a distribution in the class, we consider the following parameterizations:

$$\boldsymbol{\tau} := \begin{bmatrix} \tau_1 \in \mathcal{S}^1_{++} \\ \cdots \\ \tau_K \in \mathcal{S}^1_{++} \end{bmatrix}, \quad \boldsymbol{\lambda} := \begin{bmatrix} \lambda_1 \\ \cdots \\ \lambda_K \end{bmatrix} \in \mathbb{R}^K, \quad \boldsymbol{\eta} := \begin{bmatrix} \eta_1 \\ \cdots \\ \eta_K \end{bmatrix} \in \mathbb{R}^K$$

and maps:

$$\boldsymbol{\tau} = \psi(\boldsymbol{\lambda}) := \begin{bmatrix} f(\lambda_1) \\ \cdots \\ f(\lambda_K) \end{bmatrix}, \quad \boldsymbol{\lambda} = \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) := \boldsymbol{\lambda}_t + \boldsymbol{\eta} = \begin{bmatrix} \lambda_{1,t} + \eta_1 \\ \cdots \\ \lambda_{K,t} + \eta_K \end{bmatrix}$$

where $f(b) := \log(1 + \exp(b))$ is the soft-plus function[17] and $\boldsymbol{\tau}$ is the natural parameterization.

In this case, we can easily compute the Jacobian, where $\nabla f(b) := \frac{\exp(b)}{1+\exp(b)}$.

$$\nabla_{\boldsymbol{\eta}} \boldsymbol{\tau}\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0=\mathbf{0}} = \mathrm{Diag}\Big(\begin{bmatrix} \nabla f(\lambda_{1,t}) \\ \cdots \\ \nabla f(\lambda_{K,t}) \end{bmatrix} \Big)$$

By Lemma 5, we have

$$\hat{\boldsymbol{g}}_{\boldsymbol{\eta}_0} = \left[\nabla_{\boldsymbol{\eta}} \boldsymbol{\tau}\right]^{-T} \hat{\boldsymbol{g}}_{\boldsymbol{\tau}_t}\Big|_{\boldsymbol{\eta}=\mathbf{0}}$$

where natural-gradient $\hat{\boldsymbol{g}}_{\boldsymbol{\tau}_t}$ can be computed via the Euclidean gradient w.r.t. its expectation parameter or via direct inverse FIM computation as below

$$\begin{aligned} \hat{\boldsymbol{g}}_{\boldsymbol{\tau}_t} &= \left(\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t)\right)^{-1} \mathbf{g}_{\boldsymbol{\tau}_t} \\ &= \left(\nabla_{\boldsymbol{\tau}} \mathbf{m}\right)^{-1} \mathbf{g}_{\boldsymbol{\tau}_t} \\ &= \mathbf{g}_{\mathbf{m}} \end{aligned}$$

where $\mathbf{m} = \mathbb{E}_q\left[\mathbf{T}(w)\right] = \nabla_{\boldsymbol{\tau}} A(\boldsymbol{\tau})$ is the expectation parameter and $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\tau}_t) = \nabla^2_{\boldsymbol{\tau}} A(\boldsymbol{\tau}_t)$ is the FIM which is non-singular due to the minimality of the distribution.

Our update in the auxiliary parameter space is

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t + (-\beta \hat{\boldsymbol{g}}_{\boldsymbol{\eta}_0}) \tag{45}$$

Since $\boldsymbol{\lambda} = \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \boldsymbol{\lambda}_t + \boldsymbol{\eta}$, we can easily show that $\hat{\boldsymbol{g}}_{\boldsymbol{\eta}_0} = \hat{\boldsymbol{g}}_{\boldsymbol{\lambda}_t}$. In other words, our update recovers the standard NGD update in an unconstrained space of $\boldsymbol{\lambda}$.

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t - \beta \hat{\boldsymbol{g}}_{\boldsymbol{\lambda}_t},$$

which recovers the method proposed by Salimbeni et al. (2018) in multivariate Gaussian cases.

Therefore, by choosing $\boldsymbol{\lambda} = \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) = \boldsymbol{\lambda}_t + \boldsymbol{\eta}$, Lemma 5 generalizes the indirect method proposed by Salimbeni et al. (2018).

[16] $\exp(\cdot)$ is the scalar exponential function and do not confuse it with the matrix exponential function $\mathrm{Exp}(\cdot)$. $A(\boldsymbol{\tau})$ is C^2-smooth w.r.t. $\boldsymbol{\tau}$ as shown in Johansen (1979).

[17] We use the soft-plus function instead of the scalar exponential map for numerical stability.

G.1. Discussion about the Indirect Method

Salimbeni et al. (2018) propose an indirect method to compute natural-gradients via Auto-Differentiation (Auto-Diff) for multivariate Gaussian with full covariance structure via a unconstrained parameter transform. We have shown that this method is a special case of our approach by using a particular local parameterization and have extended it to univariate *minimal* EF distributions by using Lemma 5.

The indirect approach requires us to first define one parameterization $\boldsymbol{\tau}$ so that natural-gradient $\hat{\boldsymbol{g}}_{\boldsymbol{\tau}}$ is easy to compute under this parameterization. To compute natural-gradient in a new parameterization $\boldsymbol{\eta}$, the indirect method avoids computing the FIM $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ by computing the Jacobian $\left[\nabla_{\boldsymbol{\tau}}\boldsymbol{\eta}\right]$ instead. Unfortunately, the Jacobian matrix computation can be very complicated when it comes to a matrix parameter. Salimbeni et al. (2018) suggest using Auto-Diff to track non-zero terms in the Jacobian matrix $\left[\nabla_{\boldsymbol{\tau}}\boldsymbol{\eta}\right]$ (e.g., $\boldsymbol{\eta}$ can be a Cholesky factor of $\mathbf{S}$ and $\boldsymbol{\tau} = \mathbf{S}$ is the precision matrix in Gaussian cases with a constant mean) and to perform the Jacobian-vector product as shown in Lemma 5.

However, this indirect method has several limitations when it comes to a structured matrix parameter $\boldsymbol{\eta}$ such as structured Gaussian and Wishart cases.

- The parameterization transform used in this indirect approach often requires the Jacobian matrix $\left[\nabla_{\boldsymbol{\tau}}\boldsymbol{\eta}\right]$ to be square and invertible (see Lemma 5). For a new structured parameter $\boldsymbol{\eta}$, the Jacobian between $\boldsymbol{\tau}$ and $\boldsymbol{\eta}$ can be a non-square matrix and therefore the classical parameter transform rule fails (e.g., Lemma 5). Furthermore, it is difficult to automatically verify whether the Jacobian is invertible or not even when the Jacobian is a square matrix.

- The existing Auto-Diff implementation of the Jacobian-vector product requires us to compute a dense natural-gradient $\hat{\boldsymbol{g}}_{\boldsymbol{\tau}}$ (e.g., $\mathbf{g}_{\Sigma}$ has to compute the Hessian matrix in Gaussian cases with a constant mean) beforehand, which is not efficient for a sparse structured parameter $\boldsymbol{\eta}$.

- For a structured Gaussian NGD with second-order information, the Auto-Diff system has to first record non-zero entries in the Jacobian matrix from a structured parameterization $\boldsymbol{\eta}$ to the precision $\boldsymbol{\tau} = \mathbf{S}$ and then query the corresponding entries of natural gradient $\hat{\boldsymbol{g}}_{\boldsymbol{\tau}}$ for the precision (which can be expressed in terms of $\mathbf{G}_{S^{-1}} = \frac{1}{2}\mathbb{E}_q\left[\nabla^2_w \ell(\mathbf{w})\right]$ via Stein's identity (Khan et al., 2018)). Since Auto-Diff does not know how to organize the required entries in $\mathbf{G}_{S^{-1}}$ in *a compact and structural way*, Auto-Diff may perform too many Hessian-vector products to obtain the entries in $\mathbf{G}_{S^{-1}}$ even when we allow Auto-Diff to compute $\hat{\boldsymbol{g}}_{\boldsymbol{\tau}}$ on the fly.

- It is also unclear whether the Jacobian matrix $\left[\nabla_{\boldsymbol{\tau}}\boldsymbol{\eta}\right]$ is sparse even when the parameter $\boldsymbol{\eta}$ is sparse.

- As demonstrated by Lin et al. (2020), the indirect method via Auto-Diff could be inefficient and numerically unstable for matrix parameters such as multivariate Gaussian cases with full precision $\boldsymbol{\tau} = \mathbf{S}$.

The flexibility of our approach allows us to freely use either the indirect method (see Eq (26)) or the direct method (Eq (7)) to compute natural gradients. By using a proper local parameterization, we can directly compute the natural-gradient $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0)^{-1}\hat{\boldsymbol{g}}_{\boldsymbol{\eta}_0}$ without computing the Jacobian matrix. As shown in the main text, our update recovers the direct method suggested by Lin et al. (2020). Moreover, we can easily exploit a sparse structure in a matrix parameter as discussed in Sec. 4 of the main text. Our structured updates also reduce the number of Hessian-vector products.

The indirect method is also related to the Riemannian trivialization method (Lezcano Casado, 2019), where the unconstrained transform is considered as a push-forward map. In the trivialization method, the authors suggest doing a unconstrained transform and then performing *Euclidean gradient descent* in the trivialized (unconstrained) space. Unfortunately, the update via a trivialization (e.g., Euclidean gradient descent in a unconstrained space) can converge very slowly as shown in our experiments (see Figure 3a in the main text). In variational inference, the Riemannian trivialization method is known as the black-box variational inference (Ranganath et al., 2014). Khan & Lin (2017); Lin et al. (2019a) demonstrate that natural-gradient variational inference converges faster than block-box variational inference.

The Riemannian trivialization method is different from the natural-gradient transform method suggested by Salimbeni et al. (2018). In the method of Salimbeni et al. (2018), the authors suggest using a unconstrained global parameterization and then performing *natural gradient descent* in the unconstrained space. As shown in this paper, our approach contains the method of Salimbeni et al. (2018) as a special case.

H. Finite Mixture of Gaussians

In this appendix, we consider the following Gaussian mixture distribution q with K components.

$$q(\mathbf{w}|\boldsymbol{\tau}) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_k, \mathbf{S}_k^{-1})$$

where $\boldsymbol{\tau} = \{\boldsymbol{\mu}_k, \mathbf{S}_k\}_{k=1}^{K=1}$ and $\mathbf{S}_k$ is the precision matrix of the k-th Gaussian component.

As discussed in Lin et al. (2019a), the FIM of $q(\mathbf{w}|\boldsymbol{\eta})$ can be singular. Therefore, Assumption 1 is not satisfied.

We define $\lambda_{z_k} = \log(\frac{\pi_k}{\pi_K}) = 0$, where $\pi_k = \frac{1}{K}$. However, we can consider the Gaussian mixture as the marginal distribution of the following joint distribution such that $\int q(\mathbf{w}, z|\boldsymbol{\tau})dz = q(\mathbf{w}|\boldsymbol{\tau})$.

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

where $B(\boldsymbol{\mu}_k, \mathbf{S}_k) = \frac{1}{2}\left[\boldsymbol{\mu}_k^T\mathbf{S}_k\boldsymbol{\mu}_k - \log|\mathbf{S}_k/(2\pi)|\right]$, $A_w(\boldsymbol{\tau}, z) = \sum_{k=1}^{K}\mathbb{I}(z=k)B(\boldsymbol{\mu}_k, \mathbf{S}_k)$, $A_z(\boldsymbol{\lambda}_z) = \log(1 + \sum_{k=1}^{K-1}\exp(\lambda_{z_k}))$.

As discussed in Lin et al. (2019a), the FIM of the joint distribution $q(\mathbf{w}, z|\boldsymbol{\tau})$ is not singular. To solve a variational inference problem, Lin et al. (2019a) consider the following problem with $\gamma = 1$ in Eq (1).

$$\min_{q(\mathbf{w},z)\in\mathcal{Q}} \mathbb{E}_{q(\mathbf{w},z)}\left[\ell(\mathbf{w})\right] - \gamma\mathcal{H}(q(\mathbf{w})),$$

where we use the entropy of the marginal distribution $q(\mathbf{w})$. This approach has been studied by Agakov & Barber (2004).

This formalization allows us to relax Assumption 1 and use the joint FIM instead. Lin et al. (2019a) further show that the joint FIM is block-diagonal for each component.

Therefore, we use the following parameterizations:

$$\begin{aligned}
\boldsymbol{\tau} &:= \left\{\boldsymbol{\mu}_k \in \mathbb{R}^p,\ \ \mathbf{S}_k \in \mathcal{S}_{++}^{p\times p}\right\}_{k=1}^{K} \\
\boldsymbol{\lambda} &:= \left\{\boldsymbol{\mu}_k \in \mathbb{R}^p,\ \ \mathbf{B}_k \in \mathcal{R}_{++}^{p\times p}\right\}_{k=1}^{K} \\
\boldsymbol{\eta} &:= \left\{\boldsymbol{\delta}_k \in \mathbb{R}^p,\ \ \mathbf{M}_k \in \mathcal{S}^{p\times p}\right\}_{k=1}^{K}.
\end{aligned}$$

and maps are defined as

$$\begin{aligned}
\boldsymbol{\psi}(\boldsymbol{\lambda}) &= \{\boldsymbol{\psi}_k(\boldsymbol{\lambda}_k)\}_{k=1}^{K} \\
\boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) &= \left\{\boldsymbol{\phi}_{k,\boldsymbol{\lambda}_t}(\boldsymbol{\eta}_k)\right\}_{k=1}^{K} \\
\begin{Bmatrix}\boldsymbol{\mu}_k \\ \mathbf{S}_k\end{Bmatrix} = \boldsymbol{\psi}_k(\boldsymbol{\lambda}_k) &:= \begin{Bmatrix}\boldsymbol{\mu}_k \\ \mathbf{B}_k\mathbf{B}_k^\top\end{Bmatrix} \\
\begin{Bmatrix}\boldsymbol{\mu}_k \\ \mathbf{B}_k\end{Bmatrix} = \boldsymbol{\phi}_{k,\boldsymbol{\lambda}_t}(\boldsymbol{\eta}_k) &:= \begin{Bmatrix}\boldsymbol{\mu}_{k,t} + \mathbf{B}_{k,t}^{-T}\boldsymbol{\delta}_k \\ \mathbf{B}_{k,t}\mathbf{h}(\mathbf{M}_k)\end{Bmatrix}.
\end{aligned}$$

where $\mathbf{B}_{k,t}$ denotes the value of $\mathbf{B}_k$ at iteration t and $\boldsymbol{\lambda}_t = \left\{\boldsymbol{\mu}_{k,t}, \mathbf{B}_{k,t}\right\}_{k=1}^{K}$.

We can show that Assumption 2 is also satisfied as discussed in Gaussian cases (see Appendix D.1).

Natural gradients w.r.t. $\boldsymbol{\delta}_k$ and $\mathbf{M}_k$ can be computed as below, which is similar to (36).

$$\hat{\boldsymbol{g}}_{\delta_k} = \frac{1}{\pi_k}\mathbf{B}_{k,t}^{-1}\nabla_{\mu_k}\mathcal{L}, \quad \hat{\boldsymbol{g}}_{M_k} = -\frac{1}{\pi_k}\mathbf{B}_{k,t}^{-1}\left[\nabla_{\Sigma_k}\mathcal{L}\right]\mathbf{B}_{k,t}^{-T} \tag{46}$$

where $\mathcal{L} := \mathbb{E}_{q(\mathbf{w},z)}\left[\ell(\mathbf{w})\right] - \gamma\mathcal{H}(q(\mathbf{w}))$ and $\pi_k = \frac{1}{K}$.

Therefore, our update for the k Gaussian component is

$$\begin{aligned}\boldsymbol{\mu}_{k,t+1} &\leftarrow \boldsymbol{\mu}_{k,t} - \frac{\beta}{\pi_k}\mathbf{B}_{k,t}^{-T}\mathbf{B}_{k,t}^{-1}\nabla_{\mu_k}\mathcal{L} \\ \mathbf{B}_{k,t+1} &\leftarrow \mathbf{B}_{k,t}\mathbf{h}(\frac{\beta}{\pi_k}\mathbf{B}_{k,t}^{-1}\left[\nabla_{\Sigma_k}\mathcal{L}\right]\mathbf{B}_{k,t}^{-T})\end{aligned} \tag{47}$$

where $\pi_k = \frac{1}{K}$.

Euclidean gradients $\nabla_{\mu_k}\mathcal{L}$ and $\nabla_{\Sigma_k}\mathcal{L}$ can be computed as suggested by Lin et al. (2019a), where we use second-order information to compute $\nabla_{\Sigma_k}\mathcal{L}$. Lin et al. (2020) also show that we can compute $\nabla_{\Sigma_k}\mathcal{L}$ by first-order information if second-order information is not available.

$$\begin{aligned}\nabla_{\mu_k}\mathcal{L} &= \mathbb{E}_{q(w)}[\pi_k\delta_k\nabla_w b(\mathbf{w})] \\ \nabla_{\Sigma_k}\mathcal{L} &= \tfrac{1}{2}\mathbb{E}_{q(w)}\left[\pi_k\delta_k\nabla_w^2 b(\mathbf{w})\right] \\ &= \tfrac{1}{2}\mathbb{E}_{q(w)}\left[\pi_k\delta_k\mathbf{S}_k(\mathbf{w}-\boldsymbol{\mu}_k)\nabla_w^T b(\mathbf{w})\right]\end{aligned}$$

where $\delta_k := \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_k,\mathbf{S}_k)/\sum_{c=1}^K \pi_c\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_c,\mathbf{S}_c)$, $b(\mathbf{w}) := \ell(\mathbf{w}) + \gamma\log q(\mathbf{w}|\boldsymbol{\tau})$.

I. Matrix Gaussian for Matrix Weights in Deep Learning

In this appendix, we consider a matrix Gaussian for layer-wise matrix weights in a neural network, where a precision form will be used.

$$\mathcal{MN}(\mathbf{W}|\mathbf{E},\mathbf{S}_U^{-1},\mathbf{S}_V^{-1}) := \mathcal{N}(\text{vec}(\mathbf{W})|\text{vec}(\mathbf{E}),\mathbf{S}^{-1})$$

where the precision $\mathbf{S} = \mathbf{S}_V \otimes \mathbf{S}_U$ has a Kronecker form, $\mathbf{W} \in \mathcal{R}^{d\times p}$ is a matrix, $\mathbf{S}_V \in \mathcal{S}_{++}^{p\times p}$, $\mathbf{S}_U \in \mathcal{S}_{++}^{d\times d}$, and $\otimes$ denotes the Kronecker product.

In this case, Assumption 1 is not satisfied since the FIM of a matrix Gaussian is singular due to the cross terms between $\mathbf{S}_U$ and $\mathbf{S}_V$ in the FIM. However, a block-diagonal approximation for the FIM is non-singular. This approximation has been used in many works such as Tran et al. (2020); Glasmachers et al. (2010); Lin et al. (2019a). Therefore, we relax Assumption 1 and use the block-diagonal approximation of the FIM instead. The update is known as simultaneous block coordinate (natural-gradient) descent in optimization.

We consider the following optimization problem for NNs with L_2 regularization.

$$\min_{\boldsymbol{\tau}\in\Omega_\tau} \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\tau})}\left[\ell(\mathbf{W}) + \frac{\alpha}{2}\text{Tr}(\mathbf{W}^T\mathbf{W})\right] - \gamma\mathcal{H}(q(\mathbf{W}|\boldsymbol{\tau}))$$

where $q(\mathbf{W}) = \prod_l q(\mathbf{W}_l)$ and for each layer l, $q(\mathbf{W}_l)$ is a matrix Gaussian distribution with precision matrix $\mathbf{S}_l = \mathbf{S}_{l,V} \otimes \mathbf{S}_{l,U}$.

For simplicity, we only consider one layer and drop the layer index l.

Let's consider a global parameterization $\boldsymbol{\tau} = \{\mathbf{E},\mathbf{S}_U,\mathbf{S}_V\}$ We use the following parameterizations:

$$\begin{aligned}\boldsymbol{\tau} &:= \left\{\mathbf{E}\in\mathbb{R}^{d\times p},\ \ \mathbf{S}_V\in\mathcal{S}_{++}^{p\times p},\ \ \mathbf{S}_U\in\mathcal{S}_{++}^{d\times d}\right\} \\ \boldsymbol{\lambda} &:= \left\{\mathbf{E}\in\mathbb{R}^{d\times p},\ \ \mathbf{A}\in\mathcal{R}_{++}^{p\times p},\ \ \mathbf{B}\in\mathcal{R}_{++}^{d\times d}\right\} \\ \boldsymbol{\eta} &:= \left\{\boldsymbol{\Delta}\in\mathbb{R}^{d\times p},\ \ \mathbf{M}\in\mathcal{S}^{p\times p},\ \ \mathbf{N}\in\mathcal{S}^{d\times d}\right\}.\end{aligned}$$

and maps:

$$\left\{\begin{array}{c}\mathbf{E}\\ \mathbf{S}_V\\ \mathbf{S}_U\end{array}\right\} = \boldsymbol{\psi}(\boldsymbol{\lambda}) := \left\{\begin{array}{c}\mathbf{E}\\ \mathbf{A}\mathbf{A}^\top\\ \mathbf{B}\mathbf{B}^\top\end{array}\right\}$$
$$\left\{\begin{array}{c}\mathbf{E}\\ \mathbf{A}\\ \mathbf{B}\end{array}\right\} = \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) := \left\{\begin{array}{c}\mathbf{E}_t + \mathbf{B}_t^{-T}\boldsymbol{\Delta}\mathbf{A}_t^{-1}\\ \mathbf{A}_t\mathbf{h}(\mathbf{M})\\ \mathbf{B}_t\mathbf{h}(\mathbf{N})\end{array}\right\}.$$

Thanks to this parameterization, it is also easy to generate samples from a matrix Gaussian $\mathcal{MN}(\mathbf{W}|\mathbf{E}, \mathbf{S}_U^{-1}, \mathbf{S}_V^{-1})$ as

$$\mathbf{W} = \mathbf{E} + \mathbf{B}^{-T}\mathrm{Mat}(\mathbf{z})\mathbf{A}^{-1}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$.

The block-diagonal approximation of the FIM under the local parameterization $\boldsymbol{\eta}$ is given below. Note that we also numerically verify the following computation of FIM by Auto-Diff.

$$\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0) = \begin{bmatrix}\mathbf{I}_\Delta & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & 2d\mathbf{I}_M & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & 2p\mathbf{I}_N\end{bmatrix} \tag{48}$$

where the red terms are set to be zero due to the block-diagonal approximation while the black terms are obtained from the exact FIM.

Thanks to the block-diagonal approximation of the FIM, we can show that Assumption 2 is satisfied for each parameter block by holding the remaining blocks fixed.

Now, we discuss how to compute Euclidean gradients w.r.t. local parameterization $\boldsymbol{\eta}$. Since each matrix Gaussian $\mathcal{MN}(\mathbf{W}|\mathbf{E}, \mathbf{S}_U^{-1}, \mathbf{S}_V^{-1})$ can be re-expressed as a vector Gaussian $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{S}^{-1})$, The Euclidean gradients w.r.t. global parameter $\boldsymbol{\tau}_{\text{vec}} = \{\boldsymbol{\mu}, \mathbf{S}\}$ of the vector Gaussian are

$$\mathbf{g}_\mu = \alpha\boldsymbol{\mu} + \mathbb{E}_{\mathcal{N}(\mathbf{w}|\boldsymbol{\tau}_{\text{vec}})}\left[\nabla_w \ell(\mathbf{w})\right]$$
$$\mathbf{g}_\Sigma = \tfrac{1}{2}\big(\alpha\mathbf{I}_\Sigma + \mathbb{E}_{\mathcal{N}(\mathbf{w}|\boldsymbol{\tau}_{\text{vec}})}\left[\nabla_w^2 \ell(\mathbf{w})\right] - \gamma\mathbf{S}\big)$$

where $\mathbf{w} = \mathrm{vec}(\mathbf{W})$, $\boldsymbol{\mu} = \mathrm{vec}(\mathbf{E})$, $\boldsymbol{\Sigma} = \mathbf{S}^{-1} = \mathbf{S}_V^{-1} \otimes \mathbf{S}_U^{-1}$.

To avoid computing the Hessian $\nabla_w^2 \ell(\mathbf{w})$, we use the per-example Gauss-Newton approximation (Graves, 2011; Osawa et al., 2019a) as

$$\mathbf{g}_\Sigma \approx \tfrac{1}{2}\big(\alpha\mathbf{I}_\Sigma + \mathbb{E}_{\mathcal{N}(\mathbf{w}|\boldsymbol{\tau}_{\text{vec}})}\left[\nabla_w \ell(\mathbf{w})\nabla_w^T \ell(\mathbf{w})\right] - \gamma\mathbf{S}\big)$$

Recall that

$$\mathbf{E} = \mathbf{E}_t + \mathbf{B}_t^{-T}\boldsymbol{\Delta}\mathbf{A}_t^{-1}$$
$$\mathbf{S}_V = \mathbf{A}_t\mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T\mathbf{A}_t^T$$
$$\mathbf{S}_U = \mathbf{B}_t\mathbf{h}(\mathbf{N})\mathbf{h}(\mathbf{N})^T\mathbf{B}_t^T$$

Let's denote $\mathbf{g} = \nabla_w \ell(\mathbf{w})$ and $\mathbf{G} = \nabla_W \ell(\mathbf{W})$, where $\mathbf{w} = \mathrm{vec}(\mathbf{W})$ and $\mathbf{g} = \mathrm{vec}(\mathbf{G})$. By matrix calculus, we have

$$\mathbf{g}_\Delta\Big|_{\boldsymbol{\eta}=\mathbf{0}} = \mathbf{B}_t^{-1}\mathrm{Mat}(\mathbf{g}_\mu)\mathbf{A}_t^{-T} = \mathbf{B}_t^{-1}\big(\alpha\mathbf{E} + \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\tau})}\left[\nabla_W \ell(\mathbf{W})\right]\big)\mathbf{A}_t^{-T} = \mathbf{B}_t^{-1}\big(\alpha\mathbf{E} + \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\tau})}\left[\mathbf{G}\right]\big)\mathbf{A}_t^{-T}$$

Now, we discuss how to compute a Euclidean gradient w.r.t. $\mathbf{M}$. By the chain rule, we have

$$\mathbf{g}_{M_{ij}}\Big|_{\boldsymbol{\eta}=\mathbf{0}} = \mathrm{Tr}\big(\left[\nabla_{M_{ij}}\boldsymbol{\Sigma}\right]\mathbf{g}_\Sigma\big)$$
$$= -2\mathrm{Tr}\big(\left[(\mathbf{A}_t^{-T}\left[\nabla_{M_{ij}}\mathbf{M}\right]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\right]\mathbf{g}_\Sigma\big)$$

where M_{ij} is the entry of $\mathbf{M}$ at position (i, j).

By the Gauss-Newton approximation of the Hessian, we have

$$\mathbf{g}_{M_{ij}}\Big|_{\boldsymbol{\eta}=\mathbf{0}} \approx -\mathrm{Tr}\big(\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\big]\big(\alpha\mathbf{I}_\Sigma + \mathbb{E}_{\mathcal{N}(\mathbf{w}|\boldsymbol{\tau}_{\text{vec}})}\big[\nabla_w \ell(\mathbf{w})\nabla_w^T\ell(\mathbf{w})\big] - \gamma\mathbf{S}_t\big)\big)$$

Let's consider the first term in the approximated $\mathbf{g}_\Sigma$.

$$-\mathrm{Tr}\big(\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\big]\alpha\mathbf{I}_\Sigma\big) = -\alpha\mathrm{Tr}(\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\mathrm{Tr}(\mathbf{A}^{-1}\mathbf{A}^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big])$$

Now, we consider the second term in the approximated $\mathbf{g}_\Sigma$.

$$\begin{aligned}
&-\mathrm{Tr}\big(\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\big]\mathbb{E}_{\mathcal{N}(\mathbf{w}|\boldsymbol{\tau}_{\text{vec}})}\big[\nabla_w \ell(\mathbf{w})\nabla_w^T\ell(\mathbf{w})\big]\big)\\
=&-\mathbb{E}_{\mathcal{N}(\mathbf{w}|\boldsymbol{\tau}_{\text{vec}})}\Big[\mathrm{Tr}\big(\mathbf{g}^T\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\big]\mathbf{g}\big)\Big]\\
=&-\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\Big[\mathrm{Tr}\big(\mathrm{vec}(\mathbf{G})^T\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\big]\mathrm{vec}(\mathbf{G})\big)\Big]
\end{aligned}$$

Using the identity $(\mathbf{B}^T \otimes \mathbf{A})\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{AXB})$, we can simplify the above expression as

$$\begin{aligned}
&-\mathrm{Tr}\big(\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\big]\mathbb{E}_{\mathcal{N}(\mathbf{w}|\boldsymbol{\tau}_{\text{vec}})}\big[\nabla_w \ell(\mathbf{w})\nabla_w^T\ell(\mathbf{w})\big]\big)\\
=&-\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\Big[\mathrm{Tr}\big(\mathrm{vec}(\mathbf{G})^T\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\big]\mathrm{vec}(\mathbf{G})\big)\Big]\\
=&-\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\Big[\mathrm{Tr}\big(\mathrm{vec}(\mathbf{G})^T\mathrm{vec}\big[(\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\mathbf{G}(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}^T\big]\mathbf{A}_t^{-1})\big]\big)\Big]\\
=&-\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\Big[\mathrm{Tr}\big(\mathbf{G}^T(\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\mathbf{G}(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}^T\big]\mathbf{A}_t^{-1})\big)\Big]\\
=&-\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\Big[\mathrm{Tr}\big(\mathbf{A}_t^{-1}\mathbf{G}^T\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{G}\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}^T\big]\big)\Big]\\
=&-\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\Big[\mathrm{Tr}\big(\mathbf{A}_t^{-1}\mathbf{G}^T\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{G}\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\big)\Big] \quad (\text{ since } \mathrm{Tr}(\mathbf{CD}) = \mathrm{Tr}(\mathbf{C}^T\mathbf{D}^T))
\end{aligned}$$

where $\mathbf{C} := \mathbf{A}_t^{-1}\mathbf{G}^T\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{G}\mathbf{A}_t^{-T}$, $\mathbf{D} := \nabla_{M_{ij}}\mathbf{M}^T$ and $\mathbf{C}^T = \mathbf{C}$.

Finally, we consider the last term in the approximated $\mathbf{g}_\Sigma$.

$$\begin{aligned}
&-\mathrm{Tr}\big(\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\big](-\gamma\mathbf{S}_t)\big)\\
=&\gamma\mathrm{Tr}\big(\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\big]\mathbf{S}_t\big)\\
=&\gamma\mathrm{Tr}\big(\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\big]\big[(\mathbf{A}_t\mathbf{A}_t^T) \otimes (\mathbf{B}_t\mathbf{B}_t^T)\big]\big)\\
=&\gamma\mathrm{Tr}\big(\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^{-1}(\mathbf{A}_t\mathbf{A}_t^T)) \otimes (\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}(\mathbf{B}_t\mathbf{B}_t^T)\big]\big)\\
=&\gamma\mathrm{Tr}\big(\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^T) \otimes \mathbf{I}_B\big]\big)\\
=&\gamma\mathrm{Tr}\big(\big[(\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\mathbf{A}_t^T)\big)\mathrm{Tr}\big(\mathbf{I}_B\big)\\
=&\gamma d\mathrm{Tr}\big(\big[\nabla_{M_{ij}}\mathbf{M}\big]\big)
\end{aligned}$$

Therefore, we have the following expression due to the Gauss-Newton approximation.

$$\mathbf{g}_{M_{ij}}\Big|_{\boldsymbol{\eta}=\mathbf{0}} \approx -\alpha\mathrm{Tr}(\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\mathrm{Tr}(\mathbf{A}^{-1}\mathbf{A}^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]) - \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\Big[\mathrm{Tr}\big(\mathbf{A}_t^{-1}\mathbf{G}^T\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{G}\mathbf{A}_t^{-T}\big[\nabla_{M_{ij}}\mathbf{M}\big]\big)\Big] + \gamma d\mathrm{Tr}\big(\big[\nabla_{M_{ij}}\mathbf{M}\big]\big)$$

We can re-express it in a matrix form as

$$\mathbf{g}_M\Big|_{\boldsymbol{\eta}=\mathbf{0}} \approx -\alpha\mathrm{Tr}(\mathbf{B}_t^{-T}\mathbf{B}_t^{-1})\mathbf{A}_t^{-1}\mathbf{A}_t^{-T} - \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\Big[\mathbf{A}_t^{-1}\mathbf{G}^T\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{G}\mathbf{A}_t^{-T}\Big] + \gamma d\mathbf{I}_M$$

Similarly, we can show

$$\mathbf{g}_N\Big|_{\boldsymbol{\eta}=\mathbf{0}} \approx -\alpha\mathrm{Tr}(\mathbf{A}_t^{-T}\mathbf{A}_t^{-1})\mathbf{B}^{-1}\mathbf{B}^{-T} - \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau})}\left[\mathbf{B}_t^{-1}\mathbf{G}\mathbf{A}_t^{-T}\mathbf{A}_t^{-1}\mathbf{G}^T\mathbf{B}_t^{-T}\right] + \gamma p\mathbf{I}_N$$

Our update in terms of the auxiliary parameterization is

$$\begin{aligned}
\mathbf{E}_{t+1} &\leftarrow \mathbf{E}_t - \beta \overbrace{\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}}^{\mathbf{S}_U^{-1}}\left[\alpha\mathbf{E}_t + \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\tau}_t)}\left[\mathbf{G}\right]\right]\overbrace{\mathbf{A}_t^{-T}\mathbf{A}_t^{-1}}^{\mathbf{S}_V^{-1}} \\
\mathbf{A}_{t+1} &\leftarrow \mathbf{A}_t\mathbf{h}\Big[\frac{\beta}{2d}\Big\{-d\gamma\mathbf{I}_A + \alpha\mathrm{Tr}((\mathbf{B}_t\mathbf{B}_t^T)^{-1})\mathbf{A}_t^{-1}\mathbf{A}_t^{-T} + \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau}_t)}\left[\mathbf{A}_t^{-1}\mathbf{G}^T(\mathbf{B}_t\mathbf{B}_t^T)^{-1}\mathbf{G}\mathbf{A}_t^{-T}\right]\Big\}\Big] \\
\mathbf{B}_{t+1} &\leftarrow \mathbf{B}_t\mathbf{h}\Big[\frac{\beta}{2p}\Big\{\underbrace{-p\gamma\mathbf{I}_B}_{\text{from the entropy}} + \underbrace{\alpha\mathrm{Tr}((\mathbf{A}_t\mathbf{A}_t^T)^{-1})\mathbf{B}_t^{-1}\mathbf{B}_t^{-T}}_{\text{from the regularization}} + \underbrace{\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau}_t)}\left[\mathbf{B}_t^{-1}\mathbf{G}(\mathbf{A}_t\mathbf{A}_t^T)^{-1}\mathbf{G}^T\mathbf{B}_t^{-T}\right]}_{\text{from the NN loss}}\Big\}\Big]
\end{aligned} \tag{49}$$

By adding a natural momentum term $\mathbf{Z}$ (Khan et al., 2018) and an exponential weighted step-size $\beta_t = \frac{1-c_2^t}{1-c_1^t}$, we can obtain the following update for DNN with the Gauss-Newton approximation.

$$\begin{aligned}
\mathbf{Z}_t &\leftarrow (1-c_1)\left[\alpha\mathbf{E}_t + \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau}_t)}\left[\mathbf{G}\right]\right] + c_1\mathbf{Z}_{t-1} \\
\mathbf{E}_{t+1} &\leftarrow \mathbf{E}_t - \beta_t\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{Z}_t\mathbf{A}_t^{-T}\mathbf{A}_t^{-1} \\
\mathbf{A}_{t+1} &\leftarrow \mathbf{A}_t\mathbf{h}\Big[\frac{\beta_t}{2d}\Big\{-d\gamma\mathbf{I}_A + \alpha\mathrm{Tr}((\mathbf{B}_t\mathbf{B}_t^T)^{-1})\mathbf{A}_t^{-1}\mathbf{A}_t^{-T} + \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau}_t)}\left[\mathbf{A}_t^{-1}\mathbf{G}^T(\mathbf{B}_t\mathbf{B}_t^T)^{-1}\mathbf{G}\mathbf{A}_t^{-T}\right]\Big\}\Big] \\
\mathbf{B}_{t+1} &\leftarrow \mathbf{B}_t\mathbf{h}\Big[\frac{\beta_t}{2p}\Big\{-p\gamma\mathbf{I}_B + \alpha\mathrm{Tr}((\mathbf{A}_t\mathbf{A}_t^T)^{-1})\mathbf{B}_t^{-1}\mathbf{B}_t^{-T} + \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\tau}_t)}\left[\mathbf{B}_t^{-1}\mathbf{G}(\mathbf{A}_t\mathbf{A}_t^T)^{-1}\mathbf{G}^T\mathbf{B}_t^{-T}\right]\Big\}\Big]
\end{aligned} \tag{50}$$

where $\mathbf{G} = \nabla_W \ell(\mathbf{W})$, c_1 and c_2 are fixed to 0.9 and 0.999, respectively, as the same used in the Adam optimizer.

The time complexity for our update above is $O(d^3 + p^3)$, which is the same as noisy-KFAC (Zhang et al., 2018). In our approach, the update for $\mathbf{A}$ ($\mathbf{S}_V$) and $\mathbf{B}$ ($\mathbf{S}_U$) blocks use the *exact* FIM block. It can be shown that the corresponding updates for $\mathbf{S}_V = \mathbf{A}\mathbf{A}^T$ and $\mathbf{S}_U = \mathbf{B}\mathbf{B}^T$ blocks also use the *exact* FIM block and our update ensures that $\mathbf{S}_V$ and $\mathbf{S}_U$ are always non-singular. Our approach is different from noisy-KFAC (Zhang et al., 2018). In noisy-KFAC, the FIM of $\mathbf{S}_V$ and $\mathbf{S}_U$ are approximated by KFAC. The authors have to use additional damping to ensure that $\mathbf{S}_V$ and $\mathbf{S}_U$ are non-singular.

I.1. Complexity Reduction

A nice property of our update in (50) is that we can easily incorporate extra structures to reduce the time and space complexity. As shown in Appendix J, we can further exploit group-structures both in $\mathbf{A}$ and $\mathbf{B}$ so that the precision $\mathbf{S} = \mathbf{S}_V \otimes \mathbf{S}_U = (\mathbf{A}\mathbf{A}^T) \otimes (\mathbf{B}\mathbf{B}^T) = (\mathbf{A} \otimes \mathbf{B})(\mathbf{A} \otimes \mathbf{B})^T$ has a *low-rank* Kronecker structure to further reduce the computational complexity. Note that the Kronecker product of two matrix groups such as $\mathbf{A} \otimes \mathbf{B}$ is also a matrix group closed under the matrix multiplication. Therefore, $\mathbf{A} \otimes \mathbf{B}$ is a *Kronecker product group* when $\mathbf{A}$ and $\mathbf{B}$ are matrix groups.

Recall that the time complexity of Adam for a matrix weight $\mathbf{W} \in \mathbb{R}^{d\times p}$ is linear $O(dp)$. If a block triangular group structure (see Appendix J.1) is exploited in both $\mathbf{A}$ and $\mathbf{B}$, the time complexity of our update reduces to $O(kdp)$ from $O(d^3 + p^3)$, where $0 < k < \min(d, p)$ is a sparsity parameter for the group defined in Appendix J. In this case, our update has a linear time complexity like Adam, which is much faster than noisy-KFAC. Although we present the update based on the Gauss-Newton approximation of the Hessian, our update with the triangular group structure can be easily applied to the case with Hessian information if each Hessian has a Kronecker form such as a example about layer-wise weight matrices in a NN discussed in the next section.

Notice that our update can be automatically parallelized by Auto-Diff since our update only use basic linear algebra operations (i.e., matrix multiplication, low-rank matrix solve, and the Einstein summation) , which is more efficient than Newton-CG type updates, where a sequential conjugate-gradient (CG) step is used at each iteration.

I.2. A Layer-wise Hessian and its Approximation

We consider the following loss function parameterized by a MLP/CNN evaluated at one data point. We will show that a layer-wise Hessian of matrix weights has a Kronecker form. This result has been exploited in Dangel et al. (2020); Chen

et al. (2019). For simplicity, we only consider the matrix weight $\mathbf{W}$ at the input layer of a MLP. It is easy to extend this computation to other layers and CNN.

$$\ell(\mathbf{W}) = c(f(\mathbf{W}\mathbf{x}))$$

where x is a single data point with shape $p \times 1$, $c(\cdot)$ is a function that returns a scalar output, and $\mathbf{W}$ is the matrix weight at the input layer with shape $d \times p$.

We assume $f(\mathbf{z})$ is an element-wise C^2-smooth activation function (e.g., the tanh function). Let $\mathbf{u} := \mathbf{W}\mathbf{x}$ and $\mathbf{v} := f(\mathbf{u}) = f(\mathbf{W}\mathbf{x})$

By the chain rule, it is easy to check that

$$\begin{aligned}\nabla_W \ell(\mathbf{W}) &= \big[\nabla_v \ell\big]\big[\nabla_W \mathbf{v}\big] \\ &= \big[\underbrace{\big[\nabla_v \ell\big]}_{d\times 1} \odot \underbrace{f'(\mathbf{u})}_{d\times 1}\big] \underbrace{\mathbf{x}^T}_{1\times p}\end{aligned}$$

where $\odot$ denotes the element-wise product.

Let $\mathbf{W}_{i,:}$ denotes the i-th row of the matrix $\mathbf{W}$. We know that the shape of $\mathbf{W}_{i,:}$ is $1 \times p$.

Now, we can show that the Hessian is a Kronecker product.

$$\begin{aligned}\nabla_{W_{i,:}} \nabla_{W_{k,:}} \ell(\mathbf{W}) &= \mathcal{I}(i == k)\big[\nabla_{v_i} \ell\big] f''(u_i)\mathbf{x}\mathbf{x}^T + \big[\nabla_{v_i} \nabla_{v_j} \ell\big] f'(u_k) f'(u_i)\mathbf{x}\mathbf{x}^T \\ &= \underbrace{\Big(\mathcal{I}(i == k)\big[\nabla_{v_i} \ell\big] f''(u_i) + \big[\nabla_{v_i} \nabla_{v_k} \ell\big] f'(u_k) f'(u_i)\Big)}_{\text{a scalar}} \mathbf{x}\mathbf{x}^T\end{aligned}$$

We assume vec uses the row-major order. Therefore, if we use $\mathbf{w} = \mathrm{vec}(\mathbf{W})$ to denote a vector representation of $\mathbf{W}$, the Hessian w.r.t. $\mathbf{w} = \mathrm{vec}(\mathbf{W})$ with shape $dp \times 1$ is

$$\nabla_w^2 \ell = \underbrace{\mathbf{A}}_{d\times d} \underbrace{\otimes}_{\text{Kronecker Product}} \underbrace{(\mathbf{x}\mathbf{x}^T)}_{p\times p}$$

where $\mathbf{A}$ is a symmetric matrix with entry $A_{ik} = \mathcal{I}(i == k)\big[\nabla_{v_i} \ell\big] f''(u_i) + \big[\nabla_{v_i} \nabla_{v_k} \ell\big] f'(u_k) f'(u_i)$.

Now, we discuss the Gauss-Newton approximation of the Hessian. Note that

$$\nabla_{W_{i,:}} \ell(\mathbf{W}) = \underbrace{\big[\big[\nabla_{v_i} \ell\big] f'(u_i)\big]}_{\text{a scalar}} \mathbf{x}^T$$

where $\odot$ denotes the element-wise product.

$$\nabla_{W_{k,:}}^T \ell(\mathbf{W}) \big[\nabla_{W_{i,:}} \ell(\mathbf{W})\big] = \underbrace{\big[\nabla_{v_i} \ell\big] f'(u_i) \big[\nabla_{v_k} \ell\big] f'(u_k)}_{\text{a scalar}} \mathbf{x}\mathbf{x}^T$$

Therefore, the Gauss-Newton approximation in term of $\mathbf{w}$ can be re-expressed as

$$\mathbf{B} \otimes \big(\mathbf{x}\mathbf{x}^T\big)$$

where $\mathbf{B}$ is a symmetric matrix with entry $B_{ik} = \big[\nabla_{v_i} \ell \nabla_{v_k} \ell\big] f'(u_k) f'(u_i)$.

From the above expression, we can clearly see that the Gauss-Newton approximation ignores diagonal terms involving $f''(u_i)$ and approximates $\big[\nabla_{v_i} \nabla_{v_k} \ell\big]$ by $\big[\nabla_{v_i} \ell \nabla_{v_k} \ell\big]$.

J. Group Structures

In this section, we use the Gaussian example with square-root precision form to illustrate group structures.

J.1. Block Triangular Group

J.1.1. Proof of Lemma 1

Proof Now, we show that $\mathcal{B}_{\text{up}}(k)$ is a matrix group.

$$\mathcal{B}_{\text{up}}(k) = \left\{ \begin{bmatrix} \mathbf{B}_A & \mathbf{B}_B \\ \mathbf{0} & \mathbf{B}_D \end{bmatrix} \middle| \mathbf{B}_A \in \mathcal{R}_{++}^{k\times k},\ \mathbf{B}_D \in \mathcal{D}_{++}^{d_0\times d_0} \right\}$$

(0) It is clear that matrix multiplication is an associate product.

(1) It is obvious that $\mathbf{I} = \begin{bmatrix} \mathbf{I}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_D \end{bmatrix} \in \mathcal{B}_{\text{up}}(k)$ since $\mathbf{I}_A \in \mathcal{R}_{++}^{k\times k}$ and $\mathbf{I}_D \in \mathcal{D}_{++}^{d_0\times d_0}$.

(2) For any $\mathbf{B} \in \mathcal{B}_{\text{up}}(k)$, we have

$$\mathbf{B}^{-1} = \begin{bmatrix} \mathbf{B}_A^{-1} & -\mathbf{B}_A^{-1}\mathbf{B}_B\mathbf{B}_D^{-1} \\ \mathbf{0} & \mathbf{B}_D^{-1} \end{bmatrix} \in \mathcal{B}_{\text{up}}(k)$$

since $\mathbf{B}_A^{-1} \in \mathcal{R}_{++}^{k\times k}$ and $\mathbf{B}_D^{-1} \in \mathcal{D}_{++}^{d_0\times d_0}$.

(3) For any $\mathbf{B}, \mathbf{C} \in \mathcal{B}_{\text{up}}(k)$, the matrix product is

$$\mathbf{BC} = \begin{bmatrix} \mathbf{B}_A & \mathbf{B}_B \\ \mathbf{0} & \mathbf{B}_D \end{bmatrix} \begin{bmatrix} \mathbf{C}_A & \mathbf{C}_B \\ \mathbf{0} & \mathbf{C}_D \end{bmatrix} = \begin{bmatrix} \mathbf{B}_A\mathbf{C}_A & \mathbf{B}_A\mathbf{C}_B + \mathbf{B}_B\mathbf{C}_D \\ \mathbf{0} & \mathbf{B}_D\mathbf{C}_D \end{bmatrix} \in \mathcal{B}_{\text{up}}(k)$$

since $\mathbf{B}_A\mathbf{C}_A \in \mathcal{R}_{++}^{k\times k}$ and $\mathbf{B}_D\mathbf{C}_D \in \mathcal{D}_{++}^{d_0\times d_0}$.

J.1.2. Proof of Lemma 2

Proof For any $\mathbf{M} \in \mathcal{M}_{\text{up}}(k)$, we have

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_A & \mathbf{M}_B \\ \mathbf{0} & \mathbf{M}_D \end{bmatrix},$$

where $\mathbf{M}_A$ is symmetric and $\mathbf{M}_D$ is diagonal. Therefore,

$$\begin{aligned}
\mathbf{h}(\mathbf{M}) &= \mathbf{I} + \mathbf{M} + \tfrac{1}{2}\mathbf{M}^2 \\
&= \begin{bmatrix} \mathbf{I}_A + \mathbf{M}_A & \mathbf{M}_B \\ \mathbf{0} & \mathbf{I}_D + \mathbf{M}_D \end{bmatrix} + \tfrac{1}{2}\begin{bmatrix} \mathbf{M}_A & \mathbf{M}_B \\ \mathbf{0} & \mathbf{M}_D \end{bmatrix}\begin{bmatrix} \mathbf{M}_A & \mathbf{M}_B \\ \mathbf{0} & \mathbf{M}_D \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I}_A + \mathbf{M}_A & \mathbf{M}_B \\ \mathbf{0} & \mathbf{I}_D + \mathbf{M}_D \end{bmatrix} + \tfrac{1}{2}\begin{bmatrix} \mathbf{M}_A^2 & \mathbf{M}_A\mathbf{M}_B + \mathbf{M}_B\mathbf{M}_D \\ \mathbf{0} & \mathbf{M}_D^2 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I}_A + \mathbf{M}_A + \frac{1}{2}\mathbf{M}_A^2 & \mathbf{M}_B + \frac{1}{2}(\mathbf{M}_A\mathbf{M}_B + \mathbf{M}_B\mathbf{M}_D) \\ \mathbf{0} & \mathbf{I}_D + \mathbf{M}_D + \frac{1}{2}\mathbf{M}_D^2 \end{bmatrix} \in \mathcal{B}_{\text{up}}(k)
\end{aligned}$$

Since $\mathbf{M}_A$ is symmetric, we have $\mathbf{I}_A + \mathbf{M}_A + \frac{1}{2}\mathbf{M}_A^2 = \frac{1}{2}\big(\mathbf{I}_A + (\mathbf{I}_A + \mathbf{M}_A)(\mathbf{I}_A + \mathbf{M}_A)^T\big) \succ \mathbf{0}$ is invertible and symmetric. Similarly, $\mathbf{I}_D + \mathbf{M}_D + \frac{1}{2}\mathbf{M}_D^2$ is diagonal and invertible.

Thus, $\mathbf{h}(\mathbf{M}) \in \mathcal{B}_{\text{up}}(k)$. Moreover, the determinant $|\mathbf{h}(\mathbf{M})| > 0$

J.1.3. Proof of Lemma 3

Proof we consider the following parametrization for the Gaussian $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{S}^{-1})$, where the precision $\mathbf{S}$ belongs to a sub-manifold of $\mathcal{S}_{++}^{p\times p}$, auxiliary parameter $\mathbf{B}$ belongs to $\mathcal{B}_{\text{up}}(k)$, and local parameter $\mathbf{M}$ belongs to $\mathcal{M}_{\text{up}}(k)$,

$$\begin{aligned}
\boldsymbol{\tau} &:= \left\{ \boldsymbol{\mu} \in \mathbb{R}^p,\ \mathbf{S} = \mathbf{B}\mathbf{B}^T \in \mathcal{S}_{++}^{p\times p} \ \middle|\ \mathbf{B} \in \mathcal{B}_{\text{up}}(k) \right\}, \\
\boldsymbol{\lambda} &:= \left\{ \boldsymbol{\mu} \in \mathbb{R}^p,\ \mathbf{B} \in \mathcal{B}_{\text{up}}(k) \right\}, \\
\boldsymbol{\eta} &:= \left\{ \boldsymbol{\delta} \in \mathbb{R}^p,\ \mathbf{M} \in \mathcal{M}_{\text{up}}(k) \right\}.
\end{aligned}$$

The map $\psi \circ \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta})$ at $\boldsymbol{\lambda}_t := \{\boldsymbol{\mu}_t, \mathbf{B}_t\}$ is chosen as below, which is the same as (23)

$$\begin{bmatrix}\boldsymbol{\mu}\\ \mathbf{S}\end{bmatrix} = \psi \circ \phi_{\boldsymbol{\lambda}_t}\Big(\begin{bmatrix}\boldsymbol{\delta}\\ \mathbf{M}\end{bmatrix}\Big) = \begin{bmatrix}\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta}\\ \mathbf{B}_t\mathbf{h}(\mathbf{M})\mathbf{h}(\mathbf{M})^T\mathbf{B}_t^T\end{bmatrix}$$

As shown in Appendix J.1.4, the FIM is non-singular. Therefore, Assumption 1 is satisfied.

In Appendix J.1.4, we show that $\mathbf{M}$ can be decomposed as

$$\mathbf{M} = \mathbf{M}_{\text{diag}} + \mathbf{M}_{\text{up}} + \mathbf{M}_{\text{up}}^T + \mathbf{M}_{\text{asym}}$$

Let $\mathcal{I}_{\text{up}}, \mathcal{I}_{\text{diag}}, \mathcal{I}_{\text{asym}}$ be the index set of the non-zero entries of $\mathbf{M}_{\text{up}}$, $\mathbf{M}_{\text{diag}}$, and $\mathbf{M}_{\text{asym}}$ respectively.

Now, we can show that Assumption 2 is also satisfied. This proof is similar to the one at (37). The key idea is to use an effective representation to represent $\boldsymbol{\tau}$ and $\boldsymbol{\eta}$.

Now, let's consider the global matrix parameter. Let $\mathcal{S}_1 = \{\mathbf{B}\mathbf{B}^T | \mathbf{B} \in \mathcal{B}_{\text{up}}(k)\}$, which represents the parameter space of the global matrix parameter. Consider another set

$$\mathcal{S}_2 = \{\mathbf{U}\mathbf{U}^T | \mathbf{U} = \begin{bmatrix}\mathbf{U}_A & \mathbf{U}_B\\ \mathbf{0} & \mathbf{U}_D\end{bmatrix}\}, \tag{51}$$

where $\mathbf{U}_A \in \mathbb{R}^{k\times k}$ is an *upper-triangular* and invertible matrix, $\mathbf{U}_D$ is an invertible and diagonal matrix and $\mathbf{U}$ has *positive* diagonal entries. We will first show that $\mathcal{S}_1 = \mathcal{S}_2$ and therefore, $\mathcal{S}_2$ represents the sub-manifold. The key reason is that $\mathbf{U}$ can be used as a global parameter while $\mathbf{B}$ does not. Recall that in $\mathbf{B}$ is used as an auxiliary parameter, which could be over-parameterized. Note that a global parameter should have the same degree of freedoms as a local parameter. It is easy to verify that $\mathcal{S}_2$ and $\mathcal{M}_{\text{up}}(k)$ both have $(k+1)k/2 + (p-k)k + (p-k) = (k+1)(p-k/2)$ degrees of freedom.

We will see that $\mathbf{U}$ is indeed the output of the upper-triangular version of the Cholesky method (Lin, 2021), denoted by CholUP. In other words, if $\mathbf{S} = \mathbf{U}_1\mathbf{U}_1^T \in \mathcal{S}_2$ and $\mathbf{U}_2 = \text{CholUP}(\mathbf{S})$, we will show $\mathbf{U}_1 = \mathbf{U}_2$. This Cholesky algorithm takes a positive-definite matrix $\mathbf{X}$ as an input and returns an upper-triangular matrix $\mathbf{W}$ with *positive diagonal entries* so that $\mathbf{X} = \mathbf{W}\mathbf{W}^T$ (e.g., $\mathbf{W} = \text{CholUP}(\mathbf{X})$). Like the original Cholesky method, this method gives a unique decomposition and is C^1-smooth w.r.t. its input $\mathbf{X}$ when $\mathbf{X}$ is positive-definite.

Now, We show that $\mathcal{S}_1 = \mathcal{S}_2$. It is obvious that $\mathcal{S}_2 \subset \mathcal{S}_1$ since by construction $\mathbf{U} \in \mathcal{B}_{up}(k)$. Now, we show that $\mathcal{S}_1 \subset \mathcal{S}_2$. Consider any $\mathbf{S} \in \mathcal{S}_1$, it can be expressed as

$$\begin{aligned}\mathbf{S} &= \begin{bmatrix}\mathbf{B}_A & \mathbf{B}_B\\ \mathbf{0} & \mathbf{B}_D\end{bmatrix}\begin{bmatrix}\mathbf{B}_A & \mathbf{B}_B\\ \mathbf{0} & \mathbf{B}_D\end{bmatrix}^T\\ &= \begin{bmatrix}\mathbf{B}_A\mathbf{B}_A^T + \mathbf{B}_B\mathbf{B}_B^T & \mathbf{B}_B\mathbf{B}_D\\ \mathbf{B}_D\mathbf{B}_B^T & \mathbf{B}_D^2\end{bmatrix}\end{aligned}$$

Since $\mathbf{B}_D$ is an invertible and diagonal matrix, $\mathbf{d} := \text{abs}(\text{diag}(\mathbf{B}_D)) \odot \text{diag}^{-1}(\mathbf{B}_D)$ is a vector with entries whose value is either 1 or -1. Let $\mathbf{U}_A := \text{CholUP}(\mathbf{B}_A\mathbf{B}_A^T)$ be an upper-triangular matrix as an output by the upper-triangular version of the Cholesky method. Consider the following upper-triangular matrix $\mathbf{U}$

$$\mathbf{U} = \begin{bmatrix}\mathbf{U}_A & \mathbf{B}_B\text{Diag}(\mathbf{d}^{-1})\\ \mathbf{0} & \text{Diag}(\mathbf{d})\mathbf{B}_D\end{bmatrix}$$

We can show that this $\mathbf{U}$ has positive diagonal entries. Moreover, $\mathbf{U}\mathbf{U}^T \in \mathcal{S}_2$. Note that $\mathbf{B}_D$ is a diagonal matrix. We can show $\mathbf{U}\mathbf{U}^T = \mathbf{S}$ since

$$\mathbf{U}\mathbf{U}^T = \begin{bmatrix}\overbrace{\mathbf{U}_A\mathbf{U}_A^T}^{\mathbf{B}_A\mathbf{B}_A^T} + \mathbf{B}_B\overbrace{\text{Diag}(\mathbf{d}^{-2})}^{\mathbf{I}}\mathbf{B}_B^T & \mathbf{B}_B\mathbf{B}_D\\ \underbrace{\text{Diag}(\mathbf{d})\mathbf{B}_D\text{Diag}(\mathbf{d}^{-1})}_{\mathbf{B}_D}\mathbf{B}_B^T & \underbrace{\text{Diag}(\mathbf{d})\mathbf{B}_D\text{Diag}(\mathbf{d})\mathbf{B}_D}_{\mathbf{B}_D^2}\end{bmatrix} = \mathbf{S} \in \mathcal{S}_2$$

Therefore, $\mathcal{S}_1 = \mathcal{S}_2$ and we now show that $\mathbf{U}$ can be used as a global parameterization to represent the sub-manifold. Since $\mathcal{S}_1 = \mathcal{S}_2$, we can use $\mathcal{S}_2$ to denote the sub-manifold. Furthermore, $\mathbf{U}$ is indeed an upper-triangular and invertible matrix with positive diagonal entries, which implies that $\mathbf{U}$ is a (upper-triangular) Cholesky factor of $\mathbf{S} \in \mathcal{S}_2$. Note that the Cholesky decomposition gives a *unique* representation. Therefore, for any $\mathbf{S} = \mathbf{U}\mathbf{U}^T \in \mathcal{S}_2$, we have $\mathbf{U}_2 = \mathrm{CholUP}(\mathbf{S})$.

For the local parameter, since $\mathbf{M} \in \mathcal{M}_{\text{up}}(k)$, we have

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_A & \mathbf{M}_B \\ \mathbf{0} & \mathbf{M}_D \end{bmatrix}$$

Since $\mathbf{M}_A$ is symmetric, we can consider the upper-triangular part of $\mathbf{M}_A$, denoted by $\mathrm{triu}(\mathbf{M}_A)$. Therefore, the upper-triangular part of $\mathbf{M}$ is

$$\mathrm{triu}(\mathbf{M}) = \begin{bmatrix} \mathrm{triu}(\mathbf{M}_A) & \mathbf{M}_B \\ \mathbf{0} & \mathbf{M}_D \end{bmatrix}$$

Consider the vector representation of the non-zero entries of $\mathrm{triu}(\mathbf{M})$ denoted by $\mathbf{m}_{\text{vec}}$. Similarly, consider the vector representation of the non-zero entries of $\mathbf{U}$ denoted by $\mathrm{vec}(\mathbf{U})$. The length of $\mathbf{m}_{\text{vec}}$ is the same as the length of $\mathrm{vec}(\mathbf{U})$. Therefore, we can use these two vector representations to represent the global parameter and the local parameter in the structured spaces. Moreover, they have the same degree of freedoms. The remaining proof can be found at (37) by using the inverse function theorem and Assumption 1, where we need to use the result that if $\mathbf{S} = \mathbf{U}\mathbf{U}^T \in \mathcal{S}_2$ and $\mathbf{U}_2 = \mathrm{CholUP}(\mathbf{S})$, then $\mathbf{U} = \mathbf{U}_2$ and $\mathbf{S} \in \mathcal{S}_1$. Moreover, for any positive-definite matrix $\mathbf{X}$, $\mathrm{CholUP}(\mathbf{X})$ is C^1-smooth w.r.t. $\mathbf{X}$, which is as smooth as the original Cholesky method.

J.1.4. Natural Gradient Computation for Structured $\mathbf{M}$

we use a similar technique discussed in Appendix D.1.1 to deal with the FIM computation w.r.t. an asymmetric $\mathbf{M}$. The main idea is to decomposition $\mathbf{M}$ as a sum of special matrices so that the FIM computation is simple. We also numerically verify the following computation of FIM by Auto-Diff.

Since

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_A & \mathbf{M}_B \\ \mathbf{0} & \mathbf{M}_D \end{bmatrix} \in \mathcal{M}_{\text{up}}(k),$$

by Lemma 2, $\mathbf{h}(\mathbf{M})$ is invertible for any $\mathbf{M} \in \mathcal{M}_{\text{up}}(k)$. Moreover, by the structure of $\mathbf{M}$, $|\mathbf{h}(\mathbf{M})| > 0$.

Since $\mathbf{M}_A$ is symmetric, we can re-express the matrix $\mathbf{M}_A$ as follows. We use a similar decomposition in Appendix D.1.1.

$$\mathbf{M}_A = \mathbf{M}_{A_{\text{up}}} + \mathbf{M}_{A_{\text{up}}}^T + \mathbf{M}_{A_{\text{diag}}},$$

where $\mathbf{M}_{A_{\text{up}}}$ contains the upper-triangular half of $\mathbf{M}_A$ excluding the diagonal elements, and $\mathbf{M}_{A_{\text{diag}}}$ contains the diagonal entries of $\mathbf{M}_A$.

We will decompose the $\mathbf{M}$ as follows

$$\mathbf{M} = \mathbf{M}_{\text{diag}} + \mathbf{M}_{\text{up}} + \mathbf{M}_{\text{up}}^T + \mathbf{M}_{\text{asym}}$$

where $\mathbf{M}_{\text{diag}}$ is a diagonal matrix, $\mathbf{M}_{\text{asym}}$ is an asymmetric matrix, and $\mathbf{M}_{\text{low}}$ is a upper-triangular matrix with zero diagonal entries.

$$\mathbf{M}_{\text{diag}} = \begin{bmatrix} \mathbf{M}_{A_{\text{diag}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_D \end{bmatrix} \quad \mathbf{M}_{\text{asym}} = \begin{bmatrix} \mathbf{0} & \mathbf{M}_B \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \mathbf{M}_{\text{up}} = \begin{bmatrix} \mathbf{M}_{A_{\text{up}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Note that $\mathbf{M}_{\text{diag}}$, $\mathbf{M}_{\text{asym}}$, and $\mathbf{M}_{\text{low}}$ respectively contain the diagonal entries of $\mathbf{M}$, the asymmetric entries of $\mathbf{M}$, the upper-triangular half of the symmetric part of $\mathbf{M}$ excluding the diagonal entries.

Recall that the FIM $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0)$ is block-diagonal with two blocks—the $\boldsymbol{\delta}$ block and the $\mathbf{M}$ block. We will can show that the $\mathbf{M}$ block of the FIM is also block-diagonal with three blocks, where each block represents the non-zero entries in $\mathbf{M}_{\text{up}}$, $\mathbf{M}_{\text{diag}}$, and $\mathbf{M}_{\text{asym}}$, respectively.

Now, we will show that any cross term of the FIM between any two of these blocks is zero. We have three cases. Let $\mathcal{I}_{\text{up}}$, $\mathcal{I}_{\text{diag}}$, $\mathcal{I}_{\text{asym}}$ be the index set of the non-zero entries of $\mathbf{M}_{\text{up}}$, $\mathbf{M}_{\text{diag}}$, and $\mathbf{M}_{\text{asym}}$ respectively.

Case 1: For a cross term of the FIM between $\mathbf{M}_{\text{up}}$ and $\mathbf{M}_{\text{diag}}$, it is zero since this is the case shown in the symmetric case (see Lemma 13 in Appendix D.1.1 for details).

Case 2: For a cross term of the FIM between $\mathbf{M}_{\text{asym}}$ and $\mathbf{M}_{\text{diag}}$, we can compute it as follows.

By Eq. 28 and the chain rule, we have the following expressions, where $j > i$.

$$
\begin{aligned}
-\nabla_{M_{\text{asym}_{ij}}} \log q(\mathbf{w}|\boldsymbol{\eta}) &= -\mathrm{Tr}\big(\underbrace{\left[\nabla_{M_{\text{asym}_{ij}}} \mathbf{M}\right]}_{\mathbf{I}_{ij}} \left[\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big) \\
-\nabla_{M_{\text{diag}_{ii}}} \log q(\mathbf{w}|\boldsymbol{\eta}) &= -\mathrm{Tr}\big(\underbrace{\left[\nabla_{M_{\text{diag}_{ii}}} \mathbf{M}\right]}_{\mathbf{I}_{ii}} \left[\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big)
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
-\nabla_{M_{\text{asym}}} \log q(\mathbf{w}|\boldsymbol{\eta}) &= -\mathrm{Asym}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big) \\
-\nabla_{M_{\text{diag}}} \log q(\mathbf{w}|\boldsymbol{\eta}) &= -\mathrm{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)
\end{aligned}
$$

where we define the $\mathrm{Diag}(\cdot)$ function that returns a diagonal matrix with the same structure as $\mathbf{M}_{\text{diag}}$ and the $\mathrm{asym}(\cdot)$ function that returns a (upper) triangular matrix with the same structure as $\mathbf{M}_{\text{asym}}$.

Notice that we only consider non-zero entries in $\mathbf{M}_{\text{asym}}$, which implies that $j > i$ and $(i, j) \in \mathcal{I}_{\text{asym}}$ in the following expression. Therefore, any cross term can be expressed as below.

$$
\begin{aligned}
&-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{asym}_{ij}}} \nabla_{M_{\text{diag}}} \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = -\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{asym}_{ij}}} \mathrm{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\sum_{k,l}\left[\nabla_{M_{\text{asym}_{ij}}} M_{kl}\right] \nabla_{M_{kl}} \mathrm{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\underbrace{\left[\nabla_{M_{\text{asym}_{ij}}} M_{ij}\right]}_{=1} \nabla_{M_{ij}} \mathrm{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}} \mathrm{Diag}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&-\mathrm{Diag}\big(\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}} \nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big)\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&\mathrm{Diag}\big(\underbrace{\nabla_{M_{ij}}(\mathbf{M} + \mathbf{M}^T)}_{\mathbf{I}_{ij}+\mathbf{I}_{ji}}\big) = \mathbf{0}
\end{aligned}
$$

where we obtain the last step since $j > i$ and $\mathrm{Diag}(\mathbf{I}_{ij}) = \mathbf{0}$ since $(i, j) \in \mathcal{I}_{\text{asym}}$ and $(i, j) \notin \mathcal{I}_{\text{diag}}$.

Case 3: Now, we show that any cross term of the FIM between $\mathbf{M}_{\text{asym}}$ and $\mathbf{M}_{\text{up}}$ is zero. Let's denote a $\mathrm{Up}(\cdot)$ function that returns a upper-triangular part of an input matrix with the same (non-zero) structure as $\mathbf{M}_{\text{up}}$. Similarly, we can define a $\mathrm{Asym}(\cdot)$ function.

It is obvious see that the intersection between any two of these index sets are empty.

For any $i < j$, where $(i, j) \in \mathcal{I}_{\text{up}}$, we have $(i, j) \notin \mathcal{I}_{\text{asym}}$ and $\mathrm{Asym}\big(\mathbf{I}_{ij}\big) = \mathrm{Asym}\big(\mathbf{I}_{ji}\big) = \mathbf{0}$.

In this case, let $(i,j) \in \mathcal{I}_{\text{up}}$. The cross term can be computed as follows.

$$
\begin{aligned}
&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{up}_{ij}}} \nabla_{M_{\text{asym}}} \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = -\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{up}_{ij}}} \text{Asym}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\sum_{k,l}\left[\nabla_{M_{\text{up}_{ij}}} M_{kl}\right] \nabla_{M_{kl}} \text{Asym}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\underbrace{\left[\nabla_{M_{\text{up}_{ij}}} M_{ij}\right]}_{=1} \nabla_{M_{ij}} \text{Asym}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big) + \underbrace{\left[\nabla_{M_{\text{up}_{ij}}} M_{ji}\right]}_{=1} \nabla_{M_{ji}} \text{Asym}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}} \text{Asym}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big) + \nabla_{M_{ji}} \text{Asym}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \text{Asym}\big(\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}} \nabla_M \log q(\mathbf{w}|\boldsymbol{\eta}) + \nabla_{M_{ji}} \nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big)\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&\text{Asym}\big(\underbrace{\nabla_{M_{ij}}(\mathbf{M}+\mathbf{M}^T)}_{\mathbf{I}_{ij}+\mathbf{I}_{ji}} + \underbrace{\nabla_{M_{ji}}(\mathbf{M}+\mathbf{M}^T)}_{\mathbf{I}_{ij}+\mathbf{I}_{ji}}\big) = \mathbf{0}
\end{aligned}
$$

where we use $\mathbf{M} = \mathbf{M}_{\text{diag}} + \mathbf{M}_{\text{up}} + \mathbf{M}_{\text{up}}^T + \mathbf{M}_{\text{asym}}$ to move from step 2 to step 3, and obtain the last step since $\text{Asym}\big(\mathbf{I}_{ij}\big) = \text{Asym}\big(\mathbf{I}_{ji}\big) = \mathbf{0}$.

Now, we compute the FIM w.r.t. $\mathbf{M}_{\text{diag}}$, $\mathbf{M}_{\text{asym}}$ and $\mathbf{M}_{\text{up}}$ separately.

Like Eq (34) in Appendix D.1.1 , the FIM w.r.t. the upper-triangular block is

$$
-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{up}_{ij}}} \nabla_{M_{\text{up}}} \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = 4\mathbf{I}_{ij}
$$

.

Like Eq (33) in Appendix D.1.1 , the FIM w.r.t. the diagonal block is

$$
-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{diag}_{ij}}} \nabla_{M_{\text{diag}}} \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = 2\mathbf{I}_{ij}
$$

.

By the chain rule, the FIM w.r.t. $\mathbf{M}_{\text{asym}}$ can be computed as follows, where $(i,j) \in \mathcal{I}_{\text{asym}}$.

$$
\begin{aligned}
&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{asym}_{ij}}} \nabla_{M_{\text{asym}}} \log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{asym}_{ij}}} \text{Asym}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\sum_{k,l}\left[\nabla_{M_{\text{asym}_{ij}}} M_{kl}\right] \nabla_{M_{kl}} \text{Asym}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\underbrace{\left[\nabla_{M_{\text{asym}_{ij}}} M_{ij}\right]}_{=1} \nabla_{M_{ij}} \text{Asym}\big(\nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\big)\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&- \text{Asym}\big(\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{ij}} \nabla_M \log q(\mathbf{w}|\boldsymbol{\eta})\right]\big)\Big|_{\boldsymbol{\eta}=\mathbf{0}} \\
=&\text{Asym}\big(\underbrace{\nabla_{M_{ij}}\left[\mathbf{M}+\mathbf{M}^T\right]}_{=\mathbf{I}_{ij}+\mathbf{I}_{ji}}\big) \text{ (By Lemma 11)} \\
=&\mathbf{I}_{ij}
\end{aligned}
$$

where we obtain the last step since that $\text{Asym}(\mathbf{I}_{ji}) = \mathbf{0}$ when $i < j$ since $(i,j) \in \mathcal{I}_{\text{asym}}$ and $(j,i) \notin \mathcal{I}_{\text{asym}}$. Therefore, the

FIM w.r.t. the asymmetric block is

$$-\mathbb{E}_{q(w|\boldsymbol{\eta})}\left[\nabla_{M_{\text{asym}_{ij}}}\nabla_{M_{\text{asym}}}\log q(\mathbf{w}|\boldsymbol{\eta})\right]\Big|_{\boldsymbol{\eta}=\mathbf{0}} = \mathbf{I}_{ij}$$

.

Like the symmetric case (see Eq (35) Appendix D.1.1) when we evaluate gradients at $\boldsymbol{\eta}_0 = \{\boldsymbol{\delta}_0, \mathbf{M}_0\} = \mathbf{0}$, we have

$$\begin{aligned}\nabla_{\delta_i}\mathcal{L}\big|_{\boldsymbol{\eta}=0} &= \big[\nabla_{\delta_i}\boldsymbol{\delta}\big]^T\mathbf{B}_t^{-1}\nabla_\mu\mathcal{L}\\ \nabla_{M_{ij}}\mathcal{L}\big|_{\boldsymbol{\eta}=0} &= -\mathrm{Tr}\big(\big[\nabla_{M_{ij}}\big(\mathbf{M}+\mathbf{M}^T\big)\big]\mathbf{B}_t^{-1}\big[\nabla_\Sigma\mathcal{L}\big]\mathbf{B}_t^{-T}\big)\end{aligned}$$

Let's denote $\mathbf{G}_M = -2\mathbf{B}_t^{-1}\big[\nabla_\Sigma\mathcal{L}\big]\mathbf{B}_t^{-T}$. Therefore, we can show that Euclidean gradients are

$$\mathbf{G}_{M_{\text{diag}}} = \mathrm{Diag}(\mathbf{G}_M);\quad \mathbf{G}_{M_{\text{up}}} = \mathrm{Up}\big(\mathbf{G}_M+\mathbf{G}_M^T\big) = 2\mathrm{Up}(\mathbf{G}_M);\quad \mathbf{G}_{M_{\text{asym}}} = \mathrm{Asym}(\mathbf{G}_M);\quad \mathbf{g}_\delta = \mathbf{B}_t^{-1}\nabla_\mu\mathcal{L}$$

The natural gradients w.r.t. $\mathbf{M}_{\text{diag}}$, $\mathbf{M}_{\text{up}}$, and $\mathbf{M}_{\text{asym}}$ are $\frac{1}{2}\mathrm{Diag}(\mathbf{G})$, $\frac{1}{2}\mathrm{Up}(\mathbf{G})$, and $\mathrm{Asym}(\mathbf{G})$ respectively. The natural gradient w.r.t. $\boldsymbol{\delta}$ is $\mathbf{B}_t^{-1}\nabla_\mu\mathcal{L}$.

Natural gradients can be expressed as in the following compact form:

$$\begin{aligned}\hat{\boldsymbol{g}}_{\delta_0}^{(t)} &= \mathbf{B}_t^{-1}\nabla_\mu\mathcal{L}\\ \hat{\boldsymbol{g}}_{M_0}^{(t)} &= \mathbf{C}_{\text{up}}\odot\kappa_{\text{up}}\big(-2\mathbf{B}_t^{-1}\big[\nabla_\Sigma\mathcal{L}\big]\mathbf{B}_t^{-T}\big)\end{aligned}$$

where

$$\mathbf{C}_{\text{up}} = \begin{bmatrix}\frac{1}{2}\mathbf{J}_A & \mathbf{J}_B\\ \mathbf{0} & \frac{1}{2}\mathbf{I}_D\end{bmatrix} \in \mathcal{M}_{\text{up}}(k)$$

Therefore, our update is

$$\begin{aligned}\boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t - \beta\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{g}_{\mu_t}\\ \mathbf{B}_{t+1} &\leftarrow \mathbf{B}_t\mathbf{h}\left(\beta\mathbf{C}_{\text{up}}\odot\kappa_{\text{up}}\big(2\mathbf{B}_t^{-1}\mathbf{g}_{\Sigma_t}\mathbf{B}_t^{-T}\big)\right)\end{aligned} \tag{52}$$

J.1.5. Induced Structures

When $\mathbf{B}\in\mathcal{B}_{\text{up}}(k)$, we can show that the covariance matrix $\boldsymbol{\Sigma} = (\mathbf{B}\mathbf{B}^T)^{-1}$ has a low rank structure. The update is like the DFP update in the quasi-Newton family. This structure is useful for posterior approximation

Notice that the precision matrix $\mathbf{S} = \mathbf{B}\mathbf{B}^T$ is a block arrowhead matrix as shown below.

$$\begin{aligned}\mathbf{S} &= \mathbf{B}\mathbf{B}^T\\ &= \begin{bmatrix}\mathbf{B}_A\mathbf{B}_A^T + \mathbf{B}_B\mathbf{B}_B^T & \mathbf{B}_B\mathbf{B}_D\\ \mathbf{B}_D\mathbf{B}_B^T & \mathbf{B}_D^2\end{bmatrix}\end{aligned}$$

Now, we can show that the covariance matrix $\boldsymbol{\Sigma} = \mathbf{P}^{-1}$ admits a rank-k structure.

$$\begin{aligned}\boldsymbol{\Sigma} &= \begin{bmatrix}\mathbf{B}_A^{-T}\mathbf{B}_A^{-1} & -\mathbf{B}_A^{-T}\mathbf{B}_A^{-1}\mathbf{B}_B\mathbf{B}_D^{-1}\\ -\mathbf{B}_D^{-1}\mathbf{B}_B^T\mathbf{B}_A^{-T}\mathbf{B}_A^{-1} & \mathbf{B}_D^{-1}\mathbf{B}_B^T\mathbf{B}_A^{-T}\mathbf{B}_A^{-1}\mathbf{B}_B\mathbf{B}_D^{-1} + \mathbf{B}_D^{-2}\end{bmatrix}\\ &= \mathbf{U}_k\mathbf{U}_k^T + \begin{bmatrix}\mathbf{0} & \\ & \mathbf{B}_D^{-2}\end{bmatrix}\end{aligned}$$

where $\mathbf{U}_k$ is a p-by-k matrix as shown below and $\mathbf{U}_k$ is a rank-k matrix since $\mathbf{B}_A^{-T}$ is full k rank (invertible).

$$\mathbf{U}_k = \begin{bmatrix}-\mathbf{B}_A^{-T}\\ \mathbf{B}_D^{-1}\mathbf{B}_B^T\mathbf{B}_A^{-T}\end{bmatrix}$$

J.1.6. A SINGULARITY ISSUE OF THE FIM

In Appendix J.1.5, we know that when $\mathbf{B} \in \mathcal{B}_{\text{up}}(k)$ takes the block upper triangular structure, the covariance is a low-rank matrix.

$$\begin{aligned}\boldsymbol{\Sigma} &= (\mathbf{B}\mathbf{B}^T)^{-1} \\ &= \mathbf{U}_k\mathbf{U}_k^T + \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_D^{-2}\end{bmatrix}\end{aligned}$$

As shown in Appendix J.1.4, the FIM $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0)$ is non-singular. Equivalently, we can use auxiliary parameterization $\mathbf{A} \in \mathcal{B}_{\text{low}}(k)$ for the covariance $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$ if we choose to use the covariance as a global parameterization $\boldsymbol{\tau} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$.

In fact, the zero block (the k-by-k matrix) highlighted in red ensures the FIM $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0)$ is non-singular when $k > 0$. The group structure contains such a zero block so that the FIM is non-singular. It is tempting to use a non-zero block to replace the zero block in the above expression to get a more flexible structure. Unfortunately, the FIM $\mathbf{F}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_0)$ may become singular by doing so.

Th singularity issue also appears even when we use a common (global) parameterization $\boldsymbol{\tau}$ for a low-rank (e.g., rank-one) Gaussian (Tran et al., 2020; Mishkin et al., 2018; Sun et al., 2013) such as $\boldsymbol{\Sigma} = \mathbf{v}\mathbf{v}^T + \text{Diag}(\mathbf{d}^2)$, where $\mathbf{v}, \mathbf{d} \in \mathbb{R}^p$ are both learnable parameters. For illustration, let's consider a rank-one structure in the covariance matrix $\boldsymbol{\Sigma} \in \mathcal{S}_{++}^{p\times p}$ of Gaussians, which is a case considered in Tran et al. (2020), where the global parameterization is chosen to be $\boldsymbol{\tau} = \{\boldsymbol{\mu}, \mathbf{v}, \mathbf{d}\}$ so that the covariance $\boldsymbol{\Sigma} = \mathbf{v}\mathbf{v}^T + \text{Diag}(\mathbf{d}^2)$ has a rank-one structure. We will give two examples to show that the FIM $\mathbf{F}_{\boldsymbol{\tau}}$ is singular when $\boldsymbol{\tau} = \{\boldsymbol{\mu}, \mathbf{v}, \mathbf{d}\}$, where $\boldsymbol{\mu}, \mathbf{v}, \mathbf{d} \in \mathbb{R}^p$ are all learnable vectors. To avoid the singularity issue, Tran et al. (2020) have to use a block approximation of the FIM $\mathbf{F}_{\boldsymbol{\tau}}$. Mishkin et al. (2018) also consider a rank-one matrix in the precision matrix $\mathbf{S}$ of Gaussians, where an additional approximation is used to fix this singularity issue. Sun et al. (2013) reduce the degree of freedom in a p-dimensional low-rank Gaussians such as $\boldsymbol{\Sigma} = \mathbf{v}\mathbf{v}^T + d^2\mathbf{I}$ to avoid this issue[18], where d is chosen to be a learnable *scalar* instead of a vector. However, the covariance used in Sun et al. (2013) is less flexible than the covariance induced by our group structures since the degree of freedom for the covariance used in Sun et al. (2013) is $p + 1$ while the degree of freedom for the covariance induced by the block triangular group with $k = 1$ is $2p - 1$.

Now, we give two examples to illustrate the singularity issue in a rank-one p-dimensional Gaussian with *constant mean* and the covariance structure $\boldsymbol{\Sigma} = \mathbf{v}\mathbf{v}^T + \text{Diag}(\mathbf{d}^2)$, where $\boldsymbol{\tau} = \{\mathbf{v}, \mathbf{d}\}$ and $\mathbf{v}, \mathbf{d} \in \mathbb{R}^p$ are all learnable vectors.

Example (1): First of all, in 2-dimensional ($p = 2$) Gaussian cases with constant mean, we know that the degree of freedom of the full covariance $\boldsymbol{\Sigma}$ is 3 since $\boldsymbol{\Sigma} \in \mathcal{S}_{++}^{2\times 2}$ is symmetric. It is easy to see when $\boldsymbol{\tau} = \{\mathbf{v}, \mathbf{d}\}$, the degree of freedom in the rank-one Gaussian case with constant mean is 4, which implies the FIM is singular since the maximum degree of freedom is 3 obtained in the full Gaussian case.

Example (2): This issue also appears in higher dimensional cases. We consider an example in a 3-dimensional ($p = 3$) rank-one Gaussian with *constant zero* mean. Let's consider the following case where $\mathbf{v} = \begin{bmatrix}1\\0\\0\end{bmatrix}$, and $\mathbf{d} = \begin{bmatrix}1\\1\\1\end{bmatrix}$ so that $\boldsymbol{\Sigma} := \mathbf{v}\mathbf{v}^T + \text{Diag}(\mathbf{d}^2)$. Let $\boldsymbol{\alpha} = \begin{bmatrix}\mathbf{d}\\\mathbf{v}\end{bmatrix} \in \mathbb{R}^6$. The FIM in this case is denoted by $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\alpha})$, where the global parameter is $\boldsymbol{\tau} = \{\mathbf{v}, \mathbf{d}\}$. In this case, $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\alpha})$ computed by Auto-Diff is given below.

$$\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}) = \begin{bmatrix}0.5 & 0 & 0 & 0.5 & 0 & 0\\ 0 & 2 & 0 & 0 & 0 & 0\\ 0 & 0 & 2 & 0 & 0 & 0\\ 0.5 & 0 & 0 & 0.5 & 0 & 0\\ 0 & 0 & 0 & 0 & 0.5 & 0\\ 0 & 0 & 0 & 0 & 0 & 0.5\end{bmatrix}$$

where $\boldsymbol{\alpha} = \begin{bmatrix}1 & 1 & 1 & 1 & 0 & 0\end{bmatrix}^T$ when $\mathbf{d} = \begin{bmatrix}1 & 1 & 1\end{bmatrix}^T$ and $\mathbf{v} = \begin{bmatrix}1 & 0 & 0\end{bmatrix}^T$.

It is easy to see that $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\alpha})$ is singular. Therefore, the FIM $\mathbf{F}_{\boldsymbol{\tau}}$ under the global parameterization $\boldsymbol{\tau} = \{\mathbf{v}, \mathbf{d}\}$ for the rank-one Gaussian can be singular.

[18] When $p = 1$, the FIM of the low-rank Gaussian considered by Sun et al. (2013) is still singular.

Even when we allow to learn the mean $\boldsymbol{\mu}$ in the rank-one Gaussian cases, the FIM $\mathbf{F}_{\boldsymbol{\tau}}$ is still singular where $\boldsymbol{\tau} = \{\boldsymbol{\mu}, \underbrace{\mathbf{v}, \mathbf{d}}_{\boldsymbol{\alpha}}\}$ since $\mathbf{F}_{\boldsymbol{\tau}} = \begin{bmatrix} \mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\mu}) & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}) \end{bmatrix}$ is block-diagonal and $\mathbf{F}_{\boldsymbol{\tau}}(\boldsymbol{\alpha})$ is singular at $\boldsymbol{\mu} = \mathbf{0}$.

J.1.7. Complexity analysis and Efficient Computation

When $\mathbf{B} \in \mathcal{B}_{\text{up}}(k)$ is a p-by-p invertible matrix, it can be written as

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_A & \mathbf{B}_B \\ \mathbf{0} & \mathbf{B}_D \end{bmatrix}$$

where $\mathbf{B}_A$ is a k-by-k invertible matrix and $\mathbf{B}_D$ is a diagonal and invertible matrix.

To generate samples, we first compute the following matrix.

$$\mathbf{B}^{-T} = \begin{bmatrix} \mathbf{B}_A^{-T} & \mathbf{0} \\ -\mathbf{B}_D^{-T}\mathbf{B}_B^T\mathbf{B}_A^{-T} & \mathbf{B}_D^{-T} \end{bmatrix}$$

Given $\mathbf{B}^{-T}$ is known, for variational inference, we can easily generate a sample in $O(k^2p)$ as $\mathbf{w} = \boldsymbol{\mu} + \mathbf{B}^{-T}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Similarly, $\mathbf{S}^{-1}\mathbf{g}_\mu = \mathbf{B}^{-T}\mathbf{B}^{-1}\mathbf{g}_\mu$ can be computed in $O(k^2p)$.

Since $\mathbf{M} \in \mathcal{M}_{\text{up}}(k)$, it can be written as

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_A & \mathbf{M}_B \\ \mathbf{0} & \mathbf{M}_D \end{bmatrix}$$

where $\mathbf{M}_A$ is a k-by-k symmetric matrix and $\mathbf{M}_D$ is a diagonal matrix.

We can compute $\mathbf{h}(\mathbf{M})$ in $O(k^2p)$ when $\mathbf{M} \in \mathcal{M}_{\text{up}}(k)$

$$\mathbf{h}(\mathbf{M}) := \mathbf{I} + \mathbf{M} + \tfrac{1}{2}\mathbf{M}^2 = \begin{bmatrix} \mathbf{I}_A + \mathbf{M}_A + \frac{1}{2}\mathbf{M}_A^2 & \mathbf{M}_B + \frac{1}{2}\big(\mathbf{M}_A\mathbf{M}_B + \mathbf{M}_B\mathbf{M}_D\big) \\ \mathbf{0} & \mathbf{I}_D + \mathbf{M}_D + \frac{1}{2}\mathbf{M}_D^2 \end{bmatrix}$$

Similarly, we can compute the matrix product $\mathbf{B}\mathbf{h}(\mathbf{M})$ in $O(k^2p)$.

Now, we discuss how to compute $\kappa_{\text{up}}\big(2\mathbf{B}_t^{-1}\mathbf{g}_\Sigma\mathbf{B}_t^{-T}\big)$

We assume $\mathbf{g}_\Sigma$ can be expressed as the following form.

$$\mathbf{g}_\Sigma = \tfrac{1}{2}\begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix}$$

where $\mathbf{H}_{21} = \mathbf{H}_{12}^T$.

$$2\mathbf{B}^{-1}\mathbf{g}_\Sigma\mathbf{B}^{-T} = \begin{bmatrix} \mathbf{E} - \mathbf{F}^T\mathbf{B}_B^T\mathbf{B}_A^{-T} - \mathbf{B}_A^{-1}\mathbf{B}_B\mathbf{F} + \mathbf{B}_A^{-1}\mathbf{B}_B\mathbf{B}_D^{-1}\mathbf{H}_{22}\mathbf{B}_D^{-T}\mathbf{B}_B^T\mathbf{B}_A^{-T} & \mathbf{F}^T - \mathbf{B}_A^{-1}\mathbf{B}_B\mathbf{B}_D^{-1}\mathbf{H}_{22}\mathbf{B}_D^{-T} \\ \mathbf{F} - \mathbf{B}_D^{-1}\mathbf{H}_{22}\mathbf{B}_D^{-T}\mathbf{B}_B^T\mathbf{B}_A^{-T} & \mathbf{B}_D^{-1}\mathbf{H}_{22}\mathbf{B}_D^{-T} \end{bmatrix}$$

where $\mathbf{E} = \mathbf{B}_A^{-1}\mathbf{H}_{11}\mathbf{B}_A^{-T}$ and $\mathbf{F} = \mathbf{B}_D^{-1}\mathbf{H}_{21}\mathbf{B}_A^{-T}$

Therefore, we have

$$\kappa_{\text{up}}\big(2\mathbf{B}_t^{-1}\mathbf{g}_\Sigma\mathbf{B}_t^{-T}\big) = \begin{bmatrix} \mathbf{E} - \mathbf{F}^T\mathbf{B}_B^T\mathbf{B}_A^{-T} - \mathbf{B}_A^{-1}\mathbf{B}_B\mathbf{F} + \mathbf{B}_A^{-1}\mathbf{B}_B\mathbf{B}_D^{-1}\mathbf{H}_{22}\mathbf{B}_D^{-T}\mathbf{B}_B^T\mathbf{B}_A^{-T} & \mathbf{F}^T - \mathbf{B}_A^{-1}\mathbf{B}_B\mathbf{B}_D^{-1}\mathbf{H}_{22}\mathbf{B}_D^{-T} \\ \mathbf{0} & \text{Diag}\big(\mathbf{B}_D^{-1}\mathbf{H}_{22}\mathbf{B}_D^{-T}\big) \end{bmatrix}$$

Notice that by Stein's identity , we have

$$\mathbf{g}_\Sigma = \tfrac{1}{2}\mathbb{E}_{q(w)}\big[\nabla_w^2 f(\mathbf{w})\big]$$

where $\mathbf{w} = \boldsymbol{\mu} + \mathbf{B}^{-T}\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

For a k-rank approximation, if we can compute $O(k)$ Hessian-vector products, let's consider the following product.

$$\begin{aligned}\begin{bmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{bmatrix} &= \begin{bmatrix}\mathbf{H}_{11}\mathbf{B}_A^{-T} - \mathbf{H}_{12}\mathbf{B}_D^{-T}\mathbf{B}_B^T\mathbf{B}_A^{-T}\\ \mathbf{H}_{21}\mathbf{B}_A^{-T} - \mathbf{H}_{22}\mathbf{B}_D^{-T}\mathbf{B}_B^T\mathbf{B}_A^{-T}\end{bmatrix}\\ &= \mathbb{E}_{q(w)}\Big[\begin{bmatrix}\nabla^2_{w_1} f(\mathbf{w}_1,\mathbf{w}_2) & \nabla_{w_1}\nabla_{w_2} f(\mathbf{w}_1,\mathbf{w}_2)\\ \nabla_{w_2}\nabla_{w_1} f(\mathbf{w}_1,\mathbf{w}_2) & \nabla^2_{w_2} f(\mathbf{w}_1,\mathbf{w}_2)\end{bmatrix}\begin{bmatrix}\mathbf{B}_A^{-T}\\ -\mathbf{B}_D^{-T}\mathbf{B}_B^T\mathbf{B}_A^{-T}\end{bmatrix}\Big]\end{aligned}$$

Therefore, we have

$$\kappa_{\text{up}}\big(2\mathbf{B}_t^{-1}\mathbf{g}_\Sigma\mathbf{B}_t^{-T}\big) = \begin{bmatrix}\big(\mathbf{B}_A^{-1}\mathbf{v}_1 - \mathbf{B}_A^{-1}\mathbf{B}_B\mathbf{B}_D^{-1}\mathbf{v}_2\big) & \big(\mathbf{B}_D^{-1}\mathbf{v}_2\big)^T\\ \mathbf{0} & \mathbf{B}_D^{-1}\text{Diag}\big(\mathbf{H}_{22}\big)\mathbf{B}_D^{-T}\end{bmatrix} \tag{53}$$

We can compute this in $O(k^2 p)$ since $\mathbf{B}_D$ is diagonal, where we assume we can efficiently compute $O(k)$ Hessian-vector products and compute/approximate diagonal entries of the Hessian $\text{Diag}\big(\mathbf{H}_{22}\big)$.

J.1.8. Block Lower-triangular Group

Similarly, we can define a block lower-triangular group $\mathcal{B}_{\text{low}}(k)$ and a local parameter space $\mathcal{M}_{\text{low}}(k)$.

$$\mathcal{B}_{\text{low}}(k) = \Big\{\begin{bmatrix}\mathbf{B}_A & \mathbf{0}\\ \mathbf{B}_C & \mathbf{B}_D\end{bmatrix}\Big|\mathbf{B}_A \in \mathcal{R}^{k\times k}_{++},\ \mathbf{B}_D \in \mathcal{D}^{d_0\times d_0}_{++}\Big\};\quad \mathcal{M}_{\text{low}}(k) = \Big\{\begin{bmatrix}\mathbf{M}_A & \mathbf{0}\\ \mathbf{M}_C & \mathbf{M}_D\end{bmatrix}\Big|\quad \mathbf{M}_A \in \mathcal{S}^{k\times k},\ \mathbf{M}_D \in \mathcal{D}^{d_0\times d_0}\Big\}$$

we consider the following parametrization for the Gaussian $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{S}^{-1})$, where the precision $\mathbf{S}$ belongs to a sub-manifold of $\mathcal{S}^{p\times p}_{++}$, auxiliary parameter $\mathbf{B}$ belongs to $\mathcal{B}_{\text{low}}(k)$, and local parameter $\mathbf{M}$ belongs to $\mathcal{M}_{\text{low}}(k)$,

$$\begin{aligned}\boldsymbol{\tau} &:= \Big\{\boldsymbol{\mu}\in\mathbb{R}^p,\ \ \mathbf{S} = \mathbf{B}\mathbf{B}^T \in \mathcal{S}^{p\times p}_{++}\ \mid\ \mathbf{B}\in\mathcal{B}_{\text{low}}(k)\Big\},\\ \boldsymbol{\lambda} &:= \{\boldsymbol{\mu}\in\mathbb{R}^p,\ \ \mathbf{B}\in\mathcal{B}_{\text{low}}(k)\},\\ \boldsymbol{\eta} &:= \{\boldsymbol{\delta}\in\mathbb{R}^p,\ \ \mathbf{M}\in\mathcal{M}_{\text{low}}(k)\}.\end{aligned}$$

The map $\psi\circ\phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta})$ at $\boldsymbol{\lambda}_t := \{\boldsymbol{\mu}_t, \mathbf{B}_t\}$ is chosen as below, which is the same as (23)

$$\begin{aligned}\begin{Bmatrix}\boldsymbol{\mu}\\ \mathbf{S}\end{Bmatrix} = \psi(\boldsymbol{\lambda}) &:= \begin{Bmatrix}\boldsymbol{\mu}\\ \mathbf{B}\mathbf{B}^\top\end{Bmatrix}\\ \begin{Bmatrix}\boldsymbol{\mu}\\ \mathbf{B}\end{Bmatrix} = \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) &:= \begin{Bmatrix}\boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta}\\ \mathbf{B}_t\mathbf{h}(\mathbf{M})\end{Bmatrix}.\end{aligned}$$

We can show Assumption 1 and 2 are satisfied similar to Appendix J.1.3.

Our update over the auxiliary parameters is

$$\begin{aligned}\boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t - \beta\mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{g}_{\mu_t}\\ \mathbf{B}_{t+1} &\leftarrow \mathbf{B}_t\mathbf{h}\Big(\beta\mathbf{C}_{\text{low}}\odot\kappa_{\text{low}}\big(2\mathbf{B}_t^{-1}\mathbf{g}_{\Sigma_t}\mathbf{B}_t^{-T}\big)\Big)\end{aligned} \tag{54}$$

where

$$\mathbf{C}_{\text{low}} = \begin{bmatrix}\frac{1}{2}\mathbf{J}_A & \mathbf{0}\\ \mathbf{J}_C & \frac{1}{2}\mathbf{I}_D\end{bmatrix} \in \mathcal{M}_{\text{low}}(k)$$

where $\mathbf{J}$ denotes a matrix of ones and factor $\frac{1}{2}$ appears in the symmetric part of $\mathbf{C}_{\text{low}}$. $\odot$ denotes the element-wise product, $\kappa_{\text{low}}(\mathbf{X})$ extracts non-zero entries of $\mathcal{M}_{\text{low}}(k)$ from $\mathbf{X}$ so that $\kappa_{\text{low}}(\mathbf{X})\in\mathcal{M}_{\text{low}}(k)$. We can compute this update in $O(k^2 p)$.

When $\mathbf{B}\in\mathcal{B}_{\text{low}}(k)$, we show that the precision matrix $\mathbf{S} = \mathbf{B}\mathbf{B}^T$ has a low rank structure. This update is like the BFGS update in the quasi-Newton family. This structure is useful for optimization.

The precision matrix $\mathbf{S}$ admits a rank-k structure as shown below.

$$\mathbf{S} = \mathbf{B}\mathbf{B}^T = \begin{bmatrix} \mathbf{B}_A\mathbf{B}_A^T & \mathbf{B}_A\mathbf{B}_C^T \\ \mathbf{B}_C\mathbf{B}_A^T & \mathbf{B}_C\mathbf{B}_C^T + \mathbf{B}_D^2 \end{bmatrix} = \mathbf{V}_k\mathbf{V}_k^T + \begin{bmatrix} \mathbf{0} & \\ & \mathbf{B}_D \end{bmatrix}; \quad \mathbf{V}_k = \begin{bmatrix} \mathbf{B}_A \\ \\ \mathbf{B}_C \end{bmatrix}$$

where $\mathbf{V}_k$ is a d-by-k matrix and $\mathbf{V}_k$ is a rank-k matrix since $\mathbf{B}_A$ is full k rank.

Similarly, we can show that the covariance matrix $\mathbf{\Sigma} = \mathbf{S}^{-1}$ is a block arrowhead matrix.

$$\begin{aligned} \mathbf{\Sigma} &= \begin{bmatrix} \mathbf{B}_A^{-T} & -\mathbf{B}_A^{-T}\mathbf{B}_C^T\mathbf{B}_D^{-1} \\ \mathbf{0} & \mathbf{B}_D^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{B}_A^{-1} & \mathbf{0} \\ -\mathbf{B}_D^{-1}\mathbf{B}_C\mathbf{B}_A^{-1} & \mathbf{B}_D^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{B}_A^{-T}\mathbf{B}_A^{-1} + \mathbf{B}_A^{-T}\mathbf{B}_C^T\mathbf{B}_D^{-2}\mathbf{B}_C\mathbf{B}_A^{-1} & -\mathbf{B}_A^{-T}\mathbf{B}_C^T\mathbf{B}_D^{-2} \\ -\mathbf{B}_D^{-2}\mathbf{B}_C\mathbf{B}_A^{-1} & \mathbf{B}_D^{-2} \end{bmatrix} \end{aligned}$$

Now, we discuss how to compute $\kappa_{\text{low}}\big(2\mathbf{B}_t^{-1}\mathbf{g}_\Sigma\mathbf{B}_t^{-T}\big)$.

Similarly, we assume $\mathbf{g}_\Sigma$ can be expressed as the following form.

$$\mathbf{g}_\Sigma = \tfrac{1}{2}\begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix}$$

where $\mathbf{H}_{21} = \mathbf{H}_{12}^T$.

Therefore, we have

$$\kappa_{\text{low}}\big(2\mathbf{B}^{-1}\mathbf{g}_\Sigma\mathbf{B}^{-T}\big) = \begin{bmatrix} \mathbf{F} & \mathbf{0} \\ \\ -\mathbf{B}_D^{-1}\mathbf{B}_C\mathbf{F} + \mathbf{B}_D^{-1}\mathbf{E}_2 & \mathbf{B}_D^{-1}\text{Diag}\big[\mathbf{B}_C\mathbf{F}\mathbf{B}_C^T + \mathbf{H}_{22} - \mathbf{B}_C\mathbf{E}_2^T - \mathbf{E}_2\mathbf{B}_C^T\big]\mathbf{B}_D^{-1} \end{bmatrix}$$

where

$$\begin{bmatrix} \mathbf{E}_1 \\ \\ \mathbf{E}_2 \end{bmatrix} := \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{21}^T \\ \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{B}_A^{-T} \\ \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{11}\mathbf{B}_A^{-T} \\ \\ \mathbf{H}_{21}\mathbf{B}_A^{-T} \end{bmatrix}$$
$$\mathbf{F} := \mathbf{B}_A^{-1}\mathbf{E}_1 = \mathbf{B}_A^{-1}\mathbf{H}_{11}\mathbf{B}_A^{-T}$$

Note that we have the following identity.

$$\text{Diag}(\mathbf{A}\mathbf{B}) = \text{Diag}(\mathbf{B}^T\mathbf{A}^T) = \text{Sum}(\mathbf{A} \odot \mathbf{B}^T, \text{column})$$

where $\text{Sum}(\mathbf{X}, \text{column})$ returns a column vector by summing $\mathbf{X}$ over its columns.

Using this identity, we can further simplify the term as

$$\kappa_{\text{low}}\big(2\mathbf{B}^{-1}\mathbf{g}_\Sigma\mathbf{B}^{-T}\big) = \begin{bmatrix} \mathbf{F} & \mathbf{0} \\ \\ -\mathbf{B}_D^{-1}\mathbf{B}_C\mathbf{F} + \mathbf{B}_D^{-1}\mathbf{E}_2 & \mathbf{B}_D^{-1}\big[\text{Diag}(\mathbf{H}_{22}) + \text{Sum}(\mathbf{B}_C \odot (\mathbf{B}_C\mathbf{F} - 2\mathbf{E}_2), \text{column})\big]\mathbf{B}_D^{-1} \end{bmatrix}$$

J.2. Alternative Structures Inspired by the Heisenberg Group

First of all, the Heisenberg group is defined as follows.

$$\mathbf{B} = \begin{bmatrix} 1 & \mathbf{a}^T & c \\ \mathbf{0} & \mathbf{I} & \mathbf{b} \\ 0 & \mathbf{0} & 1 \end{bmatrix}$$

where $\mathbf{a}$ and $\mathbf{b}$ are column vectors while c is a scalar.

We construct the following set inspired by the Heisenberg group, where $1 < k_1 + k_2 < p$ and $d_0 = p - k_1 - k_2$.

$$\mathcal{B}_{\text{up}}(k_1, k_2) = \{ \begin{bmatrix} \overbrace{\mathbf{B}_A}^{k_1\text{-by-}k_1} & \overbrace{\begin{matrix} \mathbf{B}_{B_1} & \mathbf{B}_{B_2} \\ \mathbf{B}_{D_1} & \mathbf{B}_{D_2} \end{matrix}}^{\mathbf{B}_B} \\ \mathbf{0} & \underbrace{\begin{matrix} \mathbf{0} & \mathbf{B}_{D_4} \end{matrix}}_{k_2\text{-by-}k_2} \end{bmatrix} | \mathbf{B}_A \in \mathcal{R}_{++}^{k_1 \times k_1}, \mathbf{B}_{D_1} \in \mathcal{D}_{++}^{d_0 \times d_0}, \mathbf{B}_{D_4} \in \mathcal{R}_{++}^{k_2 \times k_2} \}$$

We can re-express the structure as follows

$$\mathcal{B}_{\text{up}}(k_1, k_2) = \left\{ \begin{bmatrix} \mathbf{B}_A & \mathbf{B}_B \\ \mathbf{0} & \mathbf{B}_D \end{bmatrix} \middle| \mathbf{B}_D = \begin{bmatrix} \mathbf{B}_{D_1} & \mathbf{B}_{D_2} \\ \mathbf{0} & \mathbf{B}_{D_4} \end{bmatrix} \right\}$$

where $\mathbf{B}_A \in \mathcal{R}_{++}^{k_1 \times k_1}, \mathbf{B}_{D_1} \in \mathcal{D}_{++}^{d_0 \times d_0}, \mathbf{B}_{D_4} \in \mathcal{R}_{++}^{k_2 \times k_2}$.

We can show that $\mathcal{B}_{\text{up}}(k_1, k_2)$ is a matrix group, which is more flexible than the block triangular group.

Similarly, we define a local parameter space $\mathcal{M}_{\text{up}}(k_1, k_2)$ as

$$\mathcal{M}_{\text{up}}(k_1, k_2) = \left\{ \begin{bmatrix} \mathbf{M}_A & \mathbf{M}_{B_1} & \mathbf{M}_{B_2} \\ \mathbf{0} & \mathbf{M}_{D_1} & \mathbf{M}_{D_2} \\ \mathbf{0} & \mathbf{0} & \mathbf{M}_{D_4} \end{bmatrix} \middle| \mathbf{M}_A \in \mathcal{S}^{k_1 \times k_1},\ \mathbf{M}_{D_1} \in \mathcal{D}^{d_0 \times d_0},\ \mathbf{M}_{D_4} \in \mathcal{S}^{k_2 \times k_2} \right\}$$

Likewise, we consider the following parametrization for the Gaussian $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{S}^{-1})$, where the precision $\mathbf{S}$ belongs to a sub-manifold of $\mathcal{S}_{++}^{p \times p}$, auxiliary parameter $\mathbf{B}$ belongs to $\mathcal{B}_{\text{up}}(k_1, k_2)$, and local parameter $\mathbf{M}$ belongs to $\mathcal{M}_{\text{up}}(k_1, k_2)$,

$$\begin{aligned} \boldsymbol{\tau} &:= \left\{ \boldsymbol{\mu} \in \mathbb{R}^p,\ \mathbf{S} = \mathbf{B}\mathbf{B}^T \in \mathcal{S}_{++}^{p \times p} \ \middle|\ \mathbf{B} \in \mathcal{B}_{\text{up}}(k_1, k_2) \right\}, \\ \boldsymbol{\lambda} &:= \{ \boldsymbol{\mu} \in \mathbb{R}^p,\ \mathbf{B} \in \mathcal{B}_{\text{up}}(k_1, k_2) \}, \\ \boldsymbol{\eta} &:= \{ \boldsymbol{\delta} \in \mathbb{R}^p,\ \mathbf{M} \in \mathcal{M}_{\text{up}}(k_1, k_2) \}. \end{aligned}$$

The map $\boldsymbol{\psi} \circ \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta})$ at $\boldsymbol{\lambda}_t := \{\boldsymbol{\mu}_t, \mathbf{B}_t\}$ is chosen as below, which is the same as (23)

$$\begin{aligned} \begin{Bmatrix} \boldsymbol{\mu} \\ \mathbf{S} \end{Bmatrix} = \boldsymbol{\psi}(\boldsymbol{\lambda}) &:= \begin{Bmatrix} \boldsymbol{\mu} \\ \mathbf{B}\mathbf{B}^\top \end{Bmatrix} \\ \begin{Bmatrix} \boldsymbol{\mu} \\ \mathbf{B} \end{Bmatrix} = \boldsymbol{\phi}_{\boldsymbol{\lambda}_t}(\boldsymbol{\eta}) &:= \begin{Bmatrix} \boldsymbol{\mu}_t + \mathbf{B}_t^{-T}\boldsymbol{\delta} \\ \mathbf{B}_t \mathbf{h}(\mathbf{M}) \end{Bmatrix}. \end{aligned}$$

We can show Assumption 1 and 2 are satisfied similar to Appendix J.1.3. Our update over the auxiliary parameters is

$$\begin{aligned} \boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t - \beta \mathbf{B}_t^{-T}\mathbf{B}_t^{-1}\mathbf{g}_{\mu_t} \\ \mathbf{B}_{t+1} &\leftarrow \mathbf{B}_t \mathbf{h}\left(\beta \mathbf{C}_{\text{up}} \odot \kappa_{\text{up}}\big(2\mathbf{B}_t^{-1}\mathbf{g}_{\Sigma_t}\mathbf{B}_t^{-T}\big)\right) \end{aligned} \tag{55}$$

where $\odot$ denotes the element-wise product, $\kappa_{\text{up}}(\mathbf{X})$ extracts non-zero entries of $\mathcal{M}_{\text{up}}(k_1, k_2)$ from $\mathbf{X}$ so that $\kappa_{\text{up}}(\mathbf{X}) \in \mathcal{M}_{\text{up}}(k_1, k_2)$, $\mathbf{C}_{\text{up}}$ is a constant matrix defined below, $\mathbf{J}$ denotes a matrix of ones and factor $\frac{1}{2}$ appears in the symmetric part of $\mathbf{C}_{\text{up}}$.

$$\mathbf{C}_{\text{up}} = \begin{bmatrix} \frac{1}{2}\mathbf{J}_A & \mathbf{J}_{B_1} & \mathbf{J}_{B_2} \\ \mathbf{0} & \frac{1}{2}\mathbf{I}_{D_1} & \mathbf{J}_{D_2} \\ \mathbf{0} & \mathbf{0} & \frac{1}{2}\mathbf{J}_{D_4} \end{bmatrix} \in \mathcal{M}_{\text{up}}(k_1, k_2)$$

We can also efficiently implement this update by using Hessian-vector products.

Similarly, we can define a lower version of this group denoted by $\mathcal{B}_{\text{low}}(k_1, k_2)$ and derive our update for this structure.

$$\mathcal{B}_{\text{low}}(k_1, k_2) = \left\{ \begin{bmatrix} \mathbf{B}_A & \mathbf{0} \\ \mathbf{B}_C & \mathbf{B}_D \end{bmatrix} \middle| \mathbf{B}_D = \begin{bmatrix} \mathbf{B}_{D_1} & \mathbf{0} \\ \mathbf{B}_{D_3} & \mathbf{B}_{D_4} \end{bmatrix} \right\}$$

where $\mathbf{B}_A \in \mathcal{R}_{++}^{k_1 \times k_1}, \mathbf{B}_{D_1} \in \mathcal{D}_{++}^{d_0 \times d_0}, \mathbf{B}_{D_4} \in \mathcal{R}_{++}^{k_2 \times k_2}$.