

Riemannian Optimization and Approximate Joint Diagonalization for Blind Source Separation

Supplementary Materials

Florent Bouchard*, Jérôme Malick†, Marco Congedo*

*GIPSA-lab, CNRS, Univ. Grenoble Alpes, Grenoble Institute of Technology, Grenoble France

†LJK, CNRS, Univ. Grenoble Alpes, Grenoble France

Email: florent.bouchard@gipsa-lab.fr

I. RIEMANNIAN OPTIMIZATION FRAMEWORK: PROOFS

In this appendix, we provide the proofs of the statements of section II of the paper. The constraints defining \mathcal{SP}_n and \mathcal{OP}_n only involves the SPD component of the couple (S, U) . It is enough to study $\mathcal{M}_n^S = \{S \in \mathcal{S}_n^{++} : \det(S) = 1\}$ and $\mathcal{M}_n^O = \{S \in \mathcal{S}_n^{++} : \text{ddiag}(S^2) = I_n\}$ since the rest of the proofs follows from the properties of product of manifolds [1]. In the following, $\text{diag}(\cdot)$ denotes the vector of diagonal elements of its argument.

Proof of proposition 1: We have

$$Df(\mathcal{B})[\xi] = \text{tr}((\xi_S U + S \xi_U)^T \text{grad}_{\text{GL}} f(\pi(\mathcal{B}))).$$

The result for the gradient is then obtained by basic manipulations of the trace and identification. Concerning the Hessian, recalling that $\text{Hess}_{\mathcal{E}} f(\mathcal{B})[\xi] = D \text{grad}_{\mathcal{E}} f(\mathcal{B})[\xi]$ and basic calculations gives the result.

Proof of proposition 2: To show that \mathcal{M}_n^S , \mathcal{M}_n^O and \mathcal{IP}_n are submanifolds, consider functions $F_S : \mathcal{S}_n^{++} \rightarrow \mathbb{R}_*^+$ such that $F_S(S) = \det(S)$, $F_O : \mathcal{S}_n^{++} \rightarrow (\mathbb{R}_*^+)^n$ such that $F_O(S) = \text{diag}(S^2)$ and $F_{\mathcal{I}} : \mathcal{P}_n \rightarrow (\mathbb{R}_*^+)^n$ such that $F_{\mathcal{I}}(\mathcal{B}) = \sum_k \text{diag}(BC_k B^T)^2$. Clearly $\mathcal{M}_n^S = F_S^{-1}(\{1\})$, $\mathcal{M}_n^O = F_O^{-1}(\{\mathbf{1}_n\})$ and $\mathcal{IP}_n = F_{\mathcal{I}}^{-1}(\{\mathbf{1}_n\})$ where $\mathbf{1}_n$ is the vector of ones. Let $S_S \in \mathcal{M}_n^S$, $S_O \in \mathcal{M}_n^O$ and $\mathcal{B} \in \mathcal{IP}_n$, the directional derivatives of F_S , F_O and $F_{\mathcal{I}}$ are $D F_S(S_S)[Z] = \text{tr}(S_S^{-1} Z)$, $D F_O(S_O)[Z] = 2 \text{diag}(S_O Z)$ and $D F_{\mathcal{I}}(\mathcal{B})[Z] = 2 \text{diag}((Z_S U + Z_U S)Q)$. Let $z \in \mathbb{R}$, then $Z_S = \frac{z}{n} S \in \mathcal{S}_n$ and $D F_S(S_S)[Z_S] = z$. Let $z \in \mathbb{R}^p$, the diagonal matrix Z_O such that for all i , $(Z_O)_{ii} = z_{i/2(S_O)_{ii}}$ is in \mathcal{S}_n and $D F_O(S_O)[Z_O] = z$. Finally, let $\mathcal{Z} = (SP_S(UQ\Delta)S, P_U(S\Delta Q^T))$ which Δ diagonal matrix solution to (20) where the right side is replaced by z . We leave the proof of existence of such Δ to the proof of lemma 1. \mathcal{Z} is in $T_{\mathcal{B}}\mathcal{P}_n$ and $D F_{\mathcal{I}}(\mathcal{B})[\mathcal{Z}] = z$. This shows that F_S , F_O and $F_{\mathcal{I}}$ are submersions and proposition 3.3.3 in [1] completes the proof.

Next, we show that \mathcal{M}_n^S , \mathcal{M}_n^O and \mathcal{IP}_n are connected. First consider the geodesic γ of \mathcal{S}_n^{++} defined for S_1, S_2 in \mathcal{S}_n^{++} and $t = [0, 1]$ as $\gamma(t) = S_1^{1/2} (S_1^{-1/2} S_2 S_1^{-1/2})^t S_1^{1/2}$ [2]. If S_1 and S_2 are in \mathcal{M}_n^S , then for all t , $\gamma(t)$ is also in \mathcal{M}_n^S . This is a consequence of $\det(MN) = \det(M)\det(N)$ and $\det(M^t) = \det(M)^t$ for M and N in \mathcal{S}_n^{++} . Thus, \mathcal{M}_n^S is

connected. Concerning \mathcal{M}_n^O , let $\gamma(t) = (tS_1^2 + (1-t)S_2^2)^{1/2}$ for S_1 and S_2 in \mathcal{M}_n^O . For all t , $\gamma(t)$ remains in \mathcal{M}_n^O , completing the proof. Finally, consider $P(t) = t\pi(\mathcal{B}_1) + (1-t)\pi(\mathcal{B}_2)$ for $\mathcal{B}_1, \mathcal{B}_2$ in \mathcal{IP}_n . Then, $\gamma(t) = \pi^{-1}(\Upsilon\pi(P(t)))$, where Υ is defined in proposition 4 with $R_{\mathcal{B}}(\xi)$ replaced by $P(t)$, is a smooth path in \mathcal{IP}_n . This is enough to conclude.

Finally, we find the tangent spaces of \mathcal{M}_n^S , \mathcal{M}_n^O and \mathcal{IP}_n . Let $F_S(S) = \det(S)$, $F_O(S) = \text{ddiag}(S^2)$ and $F_{\mathcal{I}}(\mathcal{B}) = \sum_k \text{ddiag}(BC_k B^T)^2$. The kernels of $D F_S(S)$, $D F_O(S)$ and $D F_{\mathcal{I}}(\mathcal{B})$ are the sets $\{\xi_S \in \mathcal{S}_n : \text{tr}(S^{-1}\xi_S) = 0\}$, $\{\xi_S \in \mathcal{S}_n : \text{ddiag}(S\xi_S) = 0\}$ and $\{\xi \in T_{\mathcal{B}}\mathcal{P}_n : \text{ddiag}(\dot{B}Q) = 0\}$. Equation (3.19) in [1] allows to conclude that these are $T_S\mathcal{M}_n^S$, $T_S\mathcal{M}_n^O$ and $T_{\mathcal{B}}\mathcal{IP}_n$.

Proof of lemma 1: Given $\mathcal{Z} = (Z_S, Z_U)$ in $T_{\mathcal{B}}\mathcal{P}_n$, basic calculations yield $P_S^S(Z_S)$ is in $T_S\mathcal{M}_n^S$, $P_S^O(Z_S)$ and $P_{\mathcal{B}}^{\mathcal{I}}(\mathcal{Z})$ are in $T_S\mathcal{M}_n^O$ and $T_{\mathcal{B}}\mathcal{IP}_n$ only if Δ_O and $\Delta_{\mathcal{I}}$ are solution to (19) and (20) respectively. Let us suppose that Δ_O and $\Delta_{\mathcal{I}}$ exist and are unique. One can check that $P_S^{S^2} = P_S^S$, $P_S^{O^2} = P_S^O$ and $P_{\mathcal{B}}^{\mathcal{I}^2} = P_{\mathcal{B}}^{\mathcal{I}}$ showing that these are projections. Finally, they are the orthogonal ones according to metric (11) since $\text{tr}(S^{-1}P_S^S(Z_S)S^{-1}\lambda S) = 0$ for any scalar λ , $\text{tr}(S^{-1}P_S^O(Z_S)S^{-1}P_S(S^2\Delta_O S)) = 0$ and $\langle P_{\mathcal{B}}^{\mathcal{I}}(\mathcal{Z}), \mathcal{Z} - P_{\mathcal{B}}^{\mathcal{I}}(\mathcal{Z}) \rangle_{\mathcal{B}} = 0$. It remains to show that equations (19) and (20) always admit a unique solution. After some manipulations, these equations can be vectorized as

$$\begin{aligned} (S^3 \odot S + S^2 \odot S^2) \text{diag}(\Delta_O) &= 2 \text{diag}(Z_S S), \\ (S \odot S_1 + S^2 \odot S_2) \text{diag}(\Delta_{\mathcal{I}}) &= 2 \text{diag}((Z_S U + S Z_U)Q), \end{aligned}$$

where \odot is the Hadamard product, $S_1 = Q^T U^T S U Q$ and $S_2 = Q^T Q$ are SPD matrices. The Schur product theorem shows that $(S^3 \odot S + S^2 \odot S^2)$ and $(S \odot S_1 + S^2 \odot S_2)$ are SPD and can be inverted. This completes the proof.

Proof of lemmas 2, 3: To show that $\hat{\Delta}_O$ and $\hat{\Delta}_{\mathcal{I}}$ are solution to these equations, it suffices to derive (19) and (20). The proofs of existence and uniqueness are identical to the ones of (19) and (20) since only the right side is modified.

Proof of proposition 3: By definition, $\text{Hess}_{\mathcal{M}} f(\mathcal{B})[\xi] = \nabla_{\xi} \text{grad}_{\mathcal{M}} f(\mathcal{B})$, where ∇ is the Levi Civita connection on \mathcal{M} [1]. The Levi Civita connection ∇ at \mathcal{B} in \mathcal{P}_n is defined for the vector field $\eta = (\eta_S, \eta_U)$ and $\xi = (\xi_S, \xi_U)$ in $T_{\mathcal{B}}\mathcal{P}_n$

as [1], [3]

$$\nabla_{\eta}\xi = P_{\mathcal{B}}(D\xi_S[\eta] - \xi_S S^{-1}\eta_S, D\xi_U[\eta]).$$

The Levi Civita connections at \mathcal{B} in \mathcal{SP}_n , \mathcal{OP}_n and \mathcal{IP}_n are

$$\begin{array}{|c|c|c|} \hline \mathcal{SP}_n & \mathcal{OP}_n & \mathcal{IP}_n \\ \hline P_{\mathcal{B}}^S(\nabla_{\eta}\xi) & P_{\mathcal{B}}^O(\nabla_{\eta}\xi) & P_{\mathcal{B}}^I(\nabla_{\eta}\xi) \\ \hline \end{array}. \quad (\text{I-1})$$

For $\text{Hess}_{\mathcal{SP}} f(S)[\xi]$, basic calculations yield

$$\begin{aligned} \text{Hess}_{\mathcal{SP}} f(S)[\xi] &= P_S^S(\text{Hess}_{\mathcal{P}} f(S)[\xi]) \\ &+ P_S^S(DP_S^S[\xi](\text{grad}_{\mathcal{P}} f(S)) + \frac{1}{p} \text{tr}(S^{-1} \text{grad}_{\mathcal{P}} f(S) \xi_S) \end{aligned}$$

and $P_S^S(DP_S^S[\xi](Z_S) = -P_S^S(\frac{1}{p} \text{tr}(S^{-1} Z_S) \xi_S)$. The result follows. Concerning $\text{Hess}_{\mathcal{OP}} f(S)[\xi]$, calculations yield

$$\begin{aligned} \text{Hess}_{\mathcal{OP}} f(S)[\xi] &= P_S^O(\text{Hess}_{\mathcal{P}} f(S)[\xi]) \\ &+ P_S^O(DP_S^O[\xi](\text{grad}_{\mathcal{P}} f(S)) + P_S((S^2 \Delta_{\mathcal{O}} + S \Delta_{\mathcal{O}} S) \xi_S) \end{aligned}$$

and $D P_S^O[\xi](Z_S) = -P_S(S \dot{M}_{\mathcal{O}} S + (S \Delta_{\mathcal{O}} S + S^2 \Delta_{\mathcal{O}}) \xi_S)$. The result follows. Finally, for $\text{Hess}_{\mathcal{IP}} f(\mathcal{B})[\xi]$, its first component $\text{Hess}_{\mathcal{IP}} f(S)[\xi]$ is

$$\begin{aligned} \text{Hess}_{\mathcal{IP}} f(S)[\xi] &= P_S^I(\text{Hess}_{\mathcal{P}} f(S)[\xi]) \\ &+ P_S^I(DP_S^I[\xi](\text{grad}_{\mathcal{P}} f(S)) + P_S(\xi_S P_S(M_{\mathcal{I}}) S)) \end{aligned}$$

with $D P_S^I[\xi](Z_S) = -2P_S(\xi_S P_S(M_{\mathcal{I}}) S) - S P_S(\dot{M}_{\mathcal{I}}) S$. For its second component $\text{Hess}_{\mathcal{IP}} f(U)[\xi]$, calculations yield

$$\begin{aligned} \text{Hess}_{\mathcal{IP}} f(U)[\xi] &= P_U^I(\text{Hess}_{\mathcal{P}} f(U)[\xi]) \\ &- P_U^I(P_U(DP_U^I[\xi](\text{grad}_{\mathcal{P}} f(U)))) \end{aligned}$$

and $P_U(DP_U^I[\xi](Z_U)) = P_U(\dot{N}_{\mathcal{I}} - \xi_U \text{sym}(U^T N_{\mathcal{I}}))$. This is enough to conclude.

Proof of proposition 4: Given S in \mathcal{M}_n^S and ξ_S in $T_S \mathcal{M}_n^S$, one can show $\det(R_S(\xi_S)) = 1$. This arises from $\det(\exp(M)) = \exp(\text{tr}(M))$. This is enough to conclude that R_S defines a retraction on \mathcal{M}_n^S . Let F be the function defined in proposition 4. Since $R_S(0) = S$ and $F(S) = S$ for S in \mathcal{M}_n^O , $F(R_S(0)) = S$. We further have $R_S(\xi_S) = S + \xi_S + o(\xi_S)$. One can show that $\text{ddiag}((S + \xi_S)^2) = I_n + o(\xi_S)$ using $\text{ddiag}(S \xi_S) = 0$ and $\text{ddiag}(S^2) = I_n$. This leads to $F(S + \xi_S) = S + \xi_S + o(\xi_S)$. Hence, $DF(R_S(0))[\xi_S] = \xi_S$. This shows that R_S^O is a retraction on \mathcal{M}_n^O . Finally, since $R_{\mathcal{B}}(0) = \mathcal{B}$, the corresponding Υ defined in proposition 4 is equal to I_n for \mathcal{B} in \mathcal{IP}_n . Thus, we have $R_{\mathcal{B}}^I(0) = \mathcal{B}$ since π is bijective. Furthermore, $R_{\mathcal{B}}(\xi) = \mathcal{B} + \xi + o(\xi)$. One can show that $\Upsilon = I_n + o(\xi)$. It follows that $DR_{\mathcal{B}}^I(0)[\xi] = \xi$. This completes the proof.

II. APPROXIMATE JOINT DIAGONALIZATION

A. Gradients and Hessians

This appendix contains the Euclidean gradients and Hessians of the AJD cost functions of section III-B of the article. All gradients and Hessians are the weighted sum of the formulas given here in table (gradients on the left-hand side and Hessians on the right-hand side). We use notations: $M_k = BC_k B^T$, $\dot{M}_k = \xi C_k B^T + BC_k \xi^T$, $N_k = A \Lambda_k A^T$ and $\dot{N}_k = \xi \Lambda_k A^T + A \Lambda_k \xi^T$.

1) *Frobenius distance:* The Euclidean gradient and Hessian of $f_{\mathcal{F}}$ are classical results (see [4] for example), thus we skip them here.

grad _{GL}	Hess _{GL}
$\tilde{f}_{\mathcal{F}}$	
$4(N_k - M_k)A\Lambda_k$	$4(\dot{N}_k A + (N_k - M_k)\xi)\Lambda_k$
$\hat{f}_{\mathcal{F}}$	
$4(M_k - \Lambda_k)BC_k$	$4(\dot{M}_k B + (M_k - \Lambda_k)\xi)C_k$

2) *Kullback-Leibler divergence:*

f_{IKL}	
$2(\Lambda_k^{-1} - M_k^{-1})BC_k$	$2\Lambda_k^{-1}(\xi - \Lambda_k^{-1}\dot{\Lambda}_k B)C_k + 2B^{-T}\xi^T B^{-T}$
$\Lambda_k = \text{ddiag}(M_k)$, $\dot{\Lambda}_k = \text{ddiag}(\dot{M}_k)$.	
f_{rKL}	
$2(I_n - M_k^{-1}\Lambda_k)B^{-T}$	$2M_k^{-1}(\dot{M}_k M_k^{-1}\Lambda_k - \dot{\Lambda}_k)B^{-T} - 2(I_n - M_k^{-1}\Lambda_k)B^{-T}\xi^T B^{-T}$
$\Lambda_k = \text{ddiag}(M_k^{-1})^{-1}$, $\dot{\Lambda}_k = \Lambda_k \text{ddiag}(M_k^{-1}\dot{M}_k M_k^{-1})\Lambda_k$.	
f_{sKL}	
$(\Lambda_k^{-1}M_k - M_k^{-1}\Lambda_k)B^{-T}$	$M_k^{-1}(\Lambda_k B^{-T}\xi^T + \dot{M}_k M_k^{-1}\Lambda_k)B^{-T} + \Lambda_k^{-1}(\xi - \Lambda_k^{-1}\dot{\Lambda}_k B)C_k$
$\Lambda_k = \text{ddiag}(M_k)^{1/2} \text{ddiag}(M_k^{-1})^{-1/2}$, $\dot{\Lambda}_k = \frac{1}{2}\Lambda_k(\text{ddiag}(\dot{M}_k) + \text{ddiag}(M_k^{-1})^{-1} \text{ddiag}(M_k^{-1}\dot{M}_k M_k^{-1}))$.	
\tilde{f}_{sKL}	
$(M_k^{-1}N_k - N_k^{-1}M_k)A^{-T}$	$N_k^{-1}(M_k A^{-T}\xi^T + \dot{N}_k N_k^{-1}M_k)A^{-T} + M_k^{-1}\xi\Lambda_k$
\hat{f}_{sKL}	
$(\Lambda_k^{-1}M_k - M_k^{-1}\Lambda_k)B^{-T}$	$M_k^{-1}(\Lambda_k B^{-T}\xi^T + \dot{M}_k M_k^{-1}\Lambda_k)B^{-T} + \Lambda_k^{-1}\xi C_k$

3) *Log-det α -divergence:*

$\tilde{f}_{\alpha\text{LD}}$	
$\frac{4}{1-\alpha}(Q_{k,\alpha}^{-1} - N_k^{-1})A\Lambda_k$	$\frac{4}{1-\alpha}Q_{k,\alpha}^{-1}(\xi - \dot{Q}_{k,\alpha}Q_{k,\alpha}^{-1}A)\Lambda_k + \frac{4}{1-\alpha}A^{-T}\xi^T A^{-T}$
$Q_{k,\alpha} = \frac{1-\alpha}{2}M_k + \frac{1+\alpha}{2}N_k$, $\dot{Q}_{k,\alpha} = \frac{1-\alpha}{2}\dot{N}_k$.	
$\hat{f}_{\alpha\text{LD}}$	
$\frac{4}{1+\alpha}(Q_{k,\alpha}^{-1} - M_k^{-1})BC_k$	$\frac{4}{1+\alpha}Q_{k,\alpha}^{-1}(\xi - \dot{Q}_{k,\alpha}Q_{k,\alpha}^{-1}B)C_k + \frac{4}{1+\alpha}B^{-T}\xi^T B^{-T}$
$Q_{k,\alpha} = \frac{1-\alpha}{2}M_k + \frac{1+\alpha}{2}\Lambda_k$, $\dot{Q}_{k,\alpha} = \frac{1-\alpha}{2}\dot{M}_k$.	

4) *Riemannian distance:*

$\tilde{f}_{\mathcal{R}}$	
$4\log(M_k^{-1}N_k)A^{-T}$	$-4\log(M_k^{-1}N_k)A^{-T}\xi^T A^{-T} + 4D\log(M_k^{-1}N_k)[\xi]A^{-T}$
$\hat{f}_{\mathcal{R}}$	
$4\log(\Lambda_k^{-1}M_k)B^{-T}$	$-4\log(\Lambda_k^{-1}M_k)B^{-T}\xi^T B^{-T} + 4D\log(\Lambda_k^{-1}M_k)[\xi]B^{-T}$

5) *Log-Euclidean distance:*

f_{LE}	
$4\mathfrak{D}\log(M_k)[O_k]BC_k$	$4D\mathfrak{D}\log(M_k)[O_k][\xi]BC_k + 4\mathfrak{D}\log(M_k)[O_k]\xi C_k$
$O_k = \log(M_k) - \text{ddiag}(\log(M_k))$.	
\tilde{f}_{LE}	
$4\mathfrak{D}\log(N_k)[O_k]A\Lambda_k$	$4D\mathfrak{D}\log(N_k)[O_k][\xi]A\Lambda_k + 4\mathfrak{D}\log(N_k)[O_k]\xi\Lambda_k$
$O_k = \log(N_k) - \log(M_k)$.	
\hat{f}_{LE}	
$4\mathfrak{D}\log(M_k)[O_k]BC_k$	$4D\mathfrak{D}\log(M_k)[O_k][\xi]BC_k + 4\mathfrak{D}\log(M_k)[O_k]\xi C_k$
$O_k = \log(M_k) - \log(\Lambda_k)$.	

$\mathfrak{D} \log(M_k)[O_k]$ is the derivative of the logarithm at M_k in the direction O_k . We have $\text{D} \log(M_k)[\xi] = \mathfrak{D} \log(M_k)[\dot{M}_k]$ and $\text{D} \mathfrak{D} \log(M_k)[O_k][\xi] = \mathfrak{D}^{(2)} \log(M_k)[O_k][\dot{M}_k]$, where $\mathfrak{D}^{(2)}$ denotes the second derivative.

For the Riemannian and log-Euclidean distances, the first and second derivatives of the logarithm are computed as per [5] and [6], respectively.

6) *Wasserstein distance:*

\widehat{f}_w	$(I_n - M_k^{1/2} Q_k M_k^{1/2}) A \Lambda_k$	$(I_n - M_k^{1/2} Q_k M_k^{1/2}) \xi \Lambda_k$ $- M_k^{1/2} \dot{Q}_k M_k^{1/2} A \Lambda_k$
$Q_k = (M_k^{1/2} N_k M_k^{1/2})^{-1/2}$, \dot{Q}_k solution to the Sylvester equation $Q_k \dot{Q}_k + \dot{Q}_k Q_k = -M_k^{-1/2} N_k^{-1} \dot{N}_k N_k^{-1} M_k^{-1/2}$		
\widehat{f}_k	$(I_n - \Lambda_k^{1/2} Q_k \Lambda_k^{1/2}) B C_k$	$(I_n - \Lambda_k^{1/2} Q_k \Lambda_k^{1/2}) \xi C_k$ $- \Lambda_k^{1/2} \dot{Q}_k \Lambda_k^{1/2} B C_k$
$Q_k = (\Lambda_k^{1/2} M_k \Lambda_k^{1/2})^{-1/2}$, \dot{Q}_k solution to the Sylvester equation $Q_k \dot{Q}_k + \dot{Q}_k Q_k = -\Lambda_k^{-1/2} M_k^{-1} \dot{M}_k M_k^{-1} \Lambda_k^{-1/2}$		

B. *Closest diagonal matrix*

The closest diagonal matrix Λ to a SPD matrix M according to the distance or divergence $d(\cdot, \cdot)$ is defined as

$$\operatorname{argmin}_{\Lambda \in \mathcal{D}_n^{++}} d(M, \Lambda), \quad (\text{II-B.1})$$

where \mathcal{D}_n^{++} denotes the set of diagonal matrices with strictly positive elements. For some divergences, (II-B.1) does not admit a closed form solution and we approximate it again using Riemannian optimization.

Recall that \mathcal{D}_n^{++} is a closed Riemannian submanifold of \mathcal{S}_n^{++} . From the Riemannian gradient and Hessian of $\Lambda \mapsto d(M, \Lambda)$, one can obtain a descent direction ξ in \mathcal{D}_n (set of diagonal matrices) from a generic Riemannian algorithm. The next iterate is then defined through the canonical retraction (exponential map) of ξ at Λ in \mathcal{D}_n^{++} , which is given by

$$\exp_{\Lambda}(\xi) = \Lambda \exp(\Lambda^{-1} \xi). \quad (\text{II-B.2})$$

1) *Log-det α -divergence:* The Riemannian gradient and Hessian of $g_{\alpha\text{LD}} : \Lambda \mapsto d_{\alpha\text{LD}}(M, \Lambda)$ are

$$\begin{aligned} \operatorname{grad}_{\mathcal{D}_n^{++}} g_{\alpha\text{LD}}(\Lambda) &= \frac{2}{1-\alpha} \Lambda (\Sigma \Lambda - I_n) \\ \operatorname{Hess}_{\mathcal{D}_n^{++}} g_{\alpha\text{LD}}(\Lambda)[\xi] &= \frac{4}{1-\alpha} \Lambda \Sigma \xi + \frac{2}{1-\alpha} \Lambda \dot{\Sigma} \Lambda, \end{aligned} \quad (\text{II-B.3})$$

where we have $\Sigma = \operatorname{ddiag} \left(\left(\frac{1-\alpha}{2} M + \frac{1+\alpha}{2} \Lambda \right)^{-1} \right)$ and $\dot{\Sigma} = -\frac{1+\alpha}{2} \operatorname{ddiag} \left(\left(\frac{1-\alpha}{2} M + \frac{1+\alpha}{2} \Lambda \right)^{-1} \xi \left(\frac{1-\alpha}{2} M + \frac{1+\alpha}{2} \Lambda \right)^{-1} \right)$.

2) *Riemannian distance:* The Riemannian gradient and Hessian of $g_{\text{R}} : \Lambda \mapsto \delta_{\text{R}}^2(M, \Lambda)$ are

$$\begin{aligned} \operatorname{grad}_{\mathcal{D}_n^{++}} g_{\text{R}}(\Lambda) &= 2\Lambda \operatorname{ddiag}(\log(M^{-1}\Lambda)) \\ \operatorname{Hess}_{\mathcal{D}_n^{++}} g_{\text{R}}(\Lambda)[\xi] &= 2\Lambda \operatorname{ddiag}(\text{D} \log(M^{-1}\Lambda)[\xi]). \end{aligned} \quad (\text{II-B.4})$$

3) *Wasserstein distance:* The Riemannian gradient and Hessian of $g_{\text{W}} : \Lambda \mapsto \delta_{\text{W}}^2(M, \Lambda)$ are

$$\begin{aligned} \operatorname{grad}_{\mathcal{D}_n^{++}} g_{\text{W}}(\Lambda) &= \frac{1}{2} \Lambda (\Lambda - \operatorname{ddiag}(Q)) \\ \operatorname{Hess}_{\mathcal{D}_n^{++}} g_{\text{W}}(\Lambda)[\xi] &= \frac{1}{2} \Lambda (\xi - \operatorname{ddiag}(\dot{Q})), \end{aligned} \quad (\text{II-B.5})$$

where $Q = (\Lambda^{1/2} M \Lambda^{1/2})^{1/2}$ and \dot{Q} solution to the Sylvester equation

$$\dot{Q}Q + Q\dot{Q} = \operatorname{sym}(\xi \Lambda^{-1/2} M \Lambda^{1/2}). \quad (\text{II-B.6})$$

Finally, the closest diagonal matrix to M according to the Wasserstein distance is Λ defined such that the gradient of g_{W} is canceled. Thus, it is the solution to (36).

REFERENCES

- [1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, USA, 2008.
- [2] R. Bhatia. *Positive definite matrices*. Princeton University Press, 2009.
- [3] G. Meyer. *Geometric optimization algorithms for linear regression on fixed-rank matrices*. PhD thesis, University of Liège, 2011.
- [4] P.-A. Absil and K. A. Gallivan. Joint diagonalization on the oblique manifold for independent component analysis. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages 945–948, May 2006.
- [5] A. H. Al-Mohy, N. J. Higham, and S. D. Relton. Computing the Fréchet derivative of the matrix logarithm and estimating the condition number. *SIAM Journal on Scientific Computing*, 35(4):C394–C410, 2013.
- [6] N. J. Higham and S. D. Relton. Higher order Fréchet derivatives of matrix functions and the level-2 condition number. *SIAM Journal on Matrix Analysis and Applications*, 35(3):1019–1037, 2014.