Riemannian Multinomial Logistics Regression for SPD Neural Networks

Supplementary Material

A. Notations

For better understanding, we briefly summarize all the notations used in this paper in Tab. 7.

B. Brief review of Riemannian manifolds

Intuitively, manifolds are locally Euclidean spaces. Differentials are the generalization of classical derivatives. For more details on smooth manifolds, please refer to [34, 52]. Riemannian manifolds are the manifolds endowed with Riemannian metrics, which can be intuitively viewed as pointwise inner products. When manifolds are endowed with Riemannian metrics, various Euclidean operators can find their counterparts in manifolds. A plethora of discussions can be found in [16].

Definition B.1 (Riemannian Manifolds)**.** A Riemannian metric on $\mathcal{M}$ is a smooth symmetric covariant 2-tensor field on $\mathcal{M}$, which is positive definite at every point. A Riemannian manifold is a pair $\{\mathcal{M}, g\}$, where $\mathcal{M}$ is a smooth manifold and g is a Riemannian metric.

The main paper relies on pullback isometry to study SPD manifolds. This idea is a natural generalization of bijection from set theory.

Definition B.2 (Pullback Metrics)**.** Suppose $\mathcal{M}, \mathcal{N}$ are smooth manifolds, g is a Riemannian metric on $\mathcal{N}$, and $f : \mathcal{M} \to \mathcal{N}$ is smooth. Then the pullback of a tensor field g by f is defined point-wisely,

$$(f^* g)_p(V_1, V_2) = g_{f(p)}(f_{*,p}(V_1), f_{*,p}(V_2)), \tag{30}$$

where p is an arbitrary point in $\mathcal{M}$, $f_{*,p}(\cdot)$ is the differential map of f at p, and V_1, V_2 are tangent vectors in $T_p\mathcal{M}$. If f^*g is positive definite, it is a Riemannian metric on $\mathcal{M}$, called the pullback metric defined by f.

Definition B.3 (Isometries)**.** If $\{M, g\}$ and $\{\widetilde{M}, \widetilde{g}\}$ are both Riemannian manifolds, a smooth map $f : M \to \widetilde{M}$ is called a (Riemannian) isometry if it is a diffeomorphism that satisfies $f^*\tilde{g} = g$.

If two manifolds are isometric, they can be viewed as equivalent. Riemannian operators in these two manifolds are also closely related.

A Lie group is a manifold with a smooth group structure. It is a combination of algebra and geometry.

Definition B.4 (Lie Groups)**.** A manifold is a Lie group, if it forms a group with a group operation $\odot$ such that $m(x, y) \mapsto x \odot y$ and $i(x) \mapsto x_{\odot}^{-1}$ are both smooth, where $x_{\odot}^{-1}$ is the group inverse of x.

At last, we briefly review the Riemannian gradient. It is a natural generalization of the Euclidean gradient.

Definition B.5 (Riemannian gradient)**.** The Riemannian gradient $\tilde{\nabla} f$ of a smooth function $f \in C^\infty(\mathcal{M})$ is a smooth vector field over $\mathcal{M}$, satisfying

$$\langle \tilde{\nabla}_p f, V \rangle_p = V(f), \forall p \in \mathcal{M}, V \in T_p\mathcal{M} \tag{31}$$

C. Proofs for the lemmas, propositions, theorems, and corollaries stated in the paper

C.1. Proof of Prop. 3.3

This claim can be proven by either definition [52, Def. 9.1] or the constant rank level set theorem [52, Thm. 11.2]. We focus on the latter.

Proof. Consider any $P \in \mathcal{S}_{++}^n$ and $A \in T_P\mathcal{S}_{++}^n$. Define the function $f(S) = \langle \mathrm{Log}_P S, A \rangle_P : \mathcal{S}_{++}^n \to \mathbb{R}$. For the SPD hyperplane $\tilde{H}_{A,P}$, we have $\tilde{H}_{A,P} = f^{-1}(0)$. Due to geodesically completeness, Log_P is globally defined, and f is therefore well-defined. We can rewrite f as a composition, i.e., $f = h \circ \mathrm{Log}_P$, where $h(\cdot) = \langle \cdot, A \rangle_P$ is a linear map.

Since Log_P is a diffeomorphism, and $h(\cdot)$ is a linear map, the rank of f is globally constant. So there exists a neighborhood (e.g., the whole SPD manifold) of $f^{-1}(0)$, where the rank of f is constant. According to the constant rank level set theorem [52, Thm. 11.2], we can obtain the claim. □

C.2. Proof of Lem. 3.5

Proof. By Thm. 2.2, we have the following,

$$\langle \mathrm{Log}_P Q, A \rangle_P = \langle \phi_{*,P} \phi_{*,\phi(P)}^{-1} (\phi(Q) - \phi(P)), \phi_{*,P} A \rangle \tag{32}$$

$$= \langle \phi(Q) - \phi(P)), \phi_{*,P} A \rangle \tag{33}$$

Therefore, the SPD hyperplane $\tilde{H}_{A_k,P_k}$ corresponds to the Euclidean hyperplane $H_{\phi_{*,P_k}(A_k),\phi(P_k)}$, due to the isometry of ϕ. Furthermore, the distances to margin hyperplanes are equivalent to the following,

$$\inf_{\phi(Q)} \|\phi(S) - \phi(Q)\|_{\mathrm{F}} \tag{34}$$

$$\text{s.t.} \langle \phi(Q) - \phi(P_k), \phi_{*,P_k} A_k \rangle = 0. \tag{35}$$

The problem above is the familiar Euclidean distance from a point to a hyperplane. By simple computation, one can obtain the results. □

Notation	Explanation
$\{\mathcal{M}, g\}$ or abbreviated as $\mathcal{M}$	A Riemannian manifold
$T_P\mathcal{M}$	The tangent space at $P \in \mathcal{M}$
$g_p(\cdot,\cdot)$ or $\langle\cdot,\cdot\rangle_P$	The Riemannian metric at $P \in \mathcal{M}$
$\|\cdot\|_P$	The norm induced by $\langle\cdot,\cdot\rangle_P$ on $T_P\mathcal{M}$
Log_P	The Riemannian logarithmic map at P
Exp_P	The Riemannian exponential map at P
$\Gamma_{P_1\to P_2}$	The Riemannian parallel transportation along the geodesic connecting P_1 and P_2
$H_{a,p}$	The Euclidean hyperplane
$\tilde{H}_{\tilde{A},P}$	The SPD hyperplane
$\odot$	A Lie group operation
$\{\mathcal{M}, \odot\}$	A Lie group
$P_{\odot}^{-1}$	The group inverse of P under $\odot$
L_P	The Lie group left translation by $P \in \mathcal{M}$
$f_{*,P}$	The differential map of the smooth map f at $P \in \mathcal{M}$
f^*g	The pullback metric by f from g
$\mathcal{S}_{++}^n$	The SPD manifold
$\mathcal{S}^n$	The Euclidean space of symmetric matrices
$\langle\cdot,\cdot\rangle$	The standard Frobenius inner product
$\|\cdot\|_{\mathrm{F}}$	The standard Frobenius norm
$\mathbf{ST}$	$\mathbf{ST} = \{(\alpha,\beta) \in \mathbb{R}^2 \mid \min(\alpha, \alpha + n\beta) > 0\}$
$\langle\cdot,\cdot\rangle^{(\alpha,\beta)}$	The $\mathrm{O}(n)$-invariant Euclidean inner product
mlog	Matrix logarithm
Chol	Cholesky decomposition
$\mathrm{Dlog}(\cdot)$	The diagonal element-wise logarithm
$\lfloor\cdot\rfloor$	The strictly lower triangular part of a square matrix
$\mathbb{D}(\cdot)$	A diagonal matrix with diagonal elements from a square matrix
Π_P	The tangential projection at P mapping a Euclidean gradient into a Riemannian one
$\nabla_P f$	The Euclidean gradient of f w.r.t. P

Table 7. Summary of notations.

C.3. Proof of Lem. 3.6

Proof. For simplicity, we abbreviate $\odot_\phi$ and g^ϕ as $\odot$ and g. By abuse of notation, we further denote $Q \odot P_{\odot}^{-1}$ as QP^{-1}, where $P_{\odot}^{-1}$ is the inversion of P under $\odot$. According to Thm. 2.2, $\{\mathcal{S}_{++}^n, \odot\}$ is an Abelian group, g is bi-invariant Riemannian metric. By Lin [35, Lem. 6], any parallel transportation can be expressed by a differential of left translation,

$$\Gamma_{P\to Q} = L_{QP^{-1}*,P}, \forall P, Q \in \mathcal{S}_{++}^n. \tag{36}$$

□

C.4. Proof of Lem. 3.7

Proof. Due to the geodesic completeness of $\mathcal{S}_{++}^n$, the existence interval of any geodesic is $\mathbb{R}$. Parallel transportation along geodesic thus exists for all $t \in \mathbb{R}$. Through Picard's uniqueness in ODE theories, one can obtain the results. □

C.5. Proof of Thm. 3.8

Proof.

$$A_k = \Gamma_{I\to P_k}(\tilde{A}_k) \tag{37}$$
$$= \phi^{-1}_{*,\phi(P_k)} \circ \phi_{*,I}(A_k) \tag{38}$$

One can obtain the results by putting Eq. (38) into Eq. (18). □

C.6. Proof of Cor. 4.1

Proof. Denoting the matrix power as $\mathrm{Pow}_\theta : \mathcal{S}_{++}^n \to \mathcal{S}_{++}^n$, then we have:

$$\mathrm{Pow}_\theta(I) = I, \tag{39}$$
$$\mathrm{Pow}_{\theta *,I}(A) = \theta A, \forall A \in T_I\mathcal{S}_{++}^n. \tag{40}$$

Next, we begin to prove the case one by one.

(α,β)**-LEM:** We define the following map

$$\psi^{\mathrm{LEM}} = f \circ \mathrm{mlog} \tag{41}$$

where $f : \mathcal{S}^n \to \mathcal{S}^n$ is the linear isometry between the standard Frobenius inner product and the $\mathrm{O}(n)$-invariant inner product $\langle \cdot, \cdot \rangle^{(\alpha,\beta)}$. Then ψ^{LEM} pulls back the standard Euclidean metric on $\mathcal{S}^n$ to (α, β)-LEM on $\mathcal{S}^n_{++}$. Putting Eqs. (40) and (41) into Eq. (20), we have

$$\begin{aligned} &\exp(\langle \psi^{\mathrm{LEM}}(S) - \psi^{\mathrm{LEM}}(P), \psi^{\mathrm{LEM}}_{*,I}(\tilde{A}_k) \rangle) \\ &= \exp\left[\langle f\left(\mathrm{mlog}(S) - \mathrm{mlog}(P_k)\right), f(\tilde{A}_k) \rangle \right] \\ &= \exp\left[\langle \mathrm{mlog}(S) - \mathrm{mlog}(P_k), \tilde{A}_k \rangle^{(\alpha,\beta)} \right], \end{aligned} \tag{42}$$

where the last equation comes from the fact that $f = f_*$.

(θ)-LCM: We denote

$$\psi^{\mathrm{LCM}} = \mathrm{Dlog} \circ \mathrm{Chol} \circ \mathrm{Pow}_\theta, \tag{43}$$

then ψ^{LCM} pulls back the Euclidean metric $\frac{1}{\theta^2} g^{\mathrm{E}}$ on the Euclidean space $\mathcal{L}^n$ of lower triangular matrices to the (θ)-LCM on $\mathcal{S}^n_{++}$. The differential of Cholesky decomposition is presented in Lin [35, Prop. 4], while the differential of Dlog can be found in [13]. Then, simple computations show that

$$\psi^{\mathrm{LCM}}_{*,I}(A) = \theta \left(\lfloor A \rfloor + \frac{1}{2} \mathbb{D}(A) \right), \forall A \in T_I \mathcal{S}^n_{++}. \tag{44}$$

Putting Eqs. (43) and (44) into Eq. (20), we can obtain the results. □

C.7. Proof of Prop. 5.1

To prove Prop. 5.1, we first present two lemmas about the general cases under PEMs.

One can observe that Eq. (20) and Eq. (21) are very similar to a Euclidean MLR. However, since ϕ is normally nonlinear and P_k is an SPD parameter, Eq. (20) cannot hastily be identified with a Euclidean MLR. However, under some special circumstances, SPD MLR can be reduced to the familiar Euclidean MLR. To show this result, we first present the Riemannian Stochastic Gradient Descent (RSGD) under PEMs. General RSGD [4] is formulated as

$$W_{t+1} = \mathrm{Exp}_{W_t}(-\gamma_t \Pi_{W_t}(\nabla_W f|_{W_t})) \tag{45}$$

where Π_{W_t} denotes the projection mapping Euclidean gradient $\nabla_W f|_{W_t}$ to Riemannian gradient, and γ_t denotes learning rate. We have already obtained the formula for the Riemannian exponential map as shown in Eq. (7). We proceed to formulate Π.

Lemma C.1. *For a smooth function $f : \mathcal{S}^n_{++} \to \mathbb{R}$ on $\mathcal{S}^n_{++}$ endowed with any kind of PEMs, the projection map $\Pi_P : \mathcal{S}^n \to T_P \mathcal{S}^n_{++}$ at $P \in \mathcal{S}^n_{++}$ is*

$$\Pi_P(\nabla_P f) = \phi^{-1}_{*,P}(\phi^{-*}_{*,P})(\nabla_P f), \tag{46}$$

where $\phi^{-*}_{*,P}$ is the adjoint operator of $\phi^{-1}_{*,P}$, i.e. $\langle V_1, \phi^{-1}_{*,P} V_2 \rangle_P = \langle \phi^{-*}_{*,P} V_1, V_2 \rangle_P$, for all $V_i \in T_P \mathcal{S}^n_{++}$.

Proof. Given any smooth function $f : \mathcal{S}^n_{++} \to \mathbb{R}$, denote its Riemannian gradient at P as $\tilde{\nabla}_P f \in T_P \mathcal{S}^n_{++}$. Then we have the following,

$$\langle \tilde{\nabla}_P f, V \rangle_P = V(f), \forall V \in T_P \mathcal{S}^n_{++}. \tag{47}$$

By Eq. (4) and canonical chart, we have

$$\langle \phi_{*,P} \tilde{\nabla}_P f, \phi_{*,P} V \rangle = \langle \nabla_P f, V \rangle, \forall V \in T_P \mathcal{S}^n_{++} \cong \mathcal{S}^n, \tag{48}$$

where $\nabla_P f$ is the Euclidean gradient. By the arbitrary of V, we have

$$\phi^*_{*,P} \phi_{*,P} \tilde{\nabla}_P f = \nabla_P f, \tag{49}$$

where $\phi^*_{*,P}$ is the adjoint operator of the linear homomorphism $\phi_{*,P}$ w.r.t. $\langle , \rangle$. □

We can describe the special case we mentioned with the above lemma.

Lemma C.2. *Supposing the differential map $\phi_{*,I}$ is the identity map, and P_k in Eq. (20) is optimized by PEM-based RSGD, then Eq. (20) can be reduced to a Euclidean MLR in the codomain of ϕ updated by Euclidean SGD.*

Proof. Define a Euclidean MLR in the codomain of ϕ as

$$p(y = k \mid S) \propto \exp(\langle \phi(S) - \bar{P}_k, \bar{A}_k \rangle), \tag{50}$$

where $\bar{P}_k, \bar{A}_k \in \mathcal{S}^n$. We call this classifier ϕ-EMLR.

Define the SPD MLR under the PEM induced by ϕ is

$$p(y = k \mid S) \propto \exp(\langle \phi(S) - \phi(P_k), \tilde{A}_k \rangle), \tag{51}$$

where $P_k \in \mathcal{S}^n_{++}, \tilde{A}_k \in \mathcal{S}^n$.

Supposing the SPD MLR and ϕ-EMLR satisfying $\bar{P}_k = \phi(P_k)$. Other settings of the network are all the same, indicating the Euclidean gradients satisfying

$$\frac{\partial L}{\partial \bar{P}_k} = \frac{\partial L}{\partial \phi(P_k)}. \tag{52}$$

The updates of $\bar{P}_k$ in the ϕ-EMLR is

$$\bar{P}'_k = \bar{P}_k - \gamma \frac{\partial L}{\partial \bar{P}_k}. \tag{53}$$

The updates of P_k in the SPD MLR is

$$\begin{aligned} P'_k &= \mathrm{Exp}_{P_k}(-\gamma \Pi_{P_k}(\nabla_{P_k} f)) && (54) \\ &= \phi^{-1}(\phi(P_k) - \gamma \phi^{-*}_{*,P_k} \frac{\partial L}{\partial P_k}) && (55) \end{aligned}$$

Therefore $\phi(P'_k)$ satisfies

$$\begin{aligned} \phi(P'_k) &= \phi(P_k) - \gamma \phi^{-*}_{*,P_k} \frac{\partial L}{\partial P_k} && (56) \\ &= \phi(P_k) - \gamma \phi^{-*}_{*,P_k} \phi^*_{*,P_k} \frac{\partial L}{\partial \phi(P_k)} && (57) \\ &= \phi(P_k) - \gamma \frac{\partial L}{\partial \phi(P_k)} && (58) \\ &= \bar{P}'_k && (59) \end{aligned}$$

Eq. (57) comes from the Euclidean chain rule of differential. Let $Y = \phi(X)$, then we have

$$\frac{\partial L}{\partial Y} : \mathrm{d}\, Y = \frac{\partial L}{\partial Y} : \phi_{*,X}\, \mathrm{d}\, X = \phi^*_{*,X} \frac{\partial L}{\partial Y} : \mathrm{d}\, X, \quad (60)$$

where : means Frobenius inner product.

The equivalence of $\bar{A}_k$ and $\tilde{A}_k$ is obvious. By natural induction, the claim can be proven. □

Now, We can directly prove Prop. 5.1 by Lem. C.2.