Robust optimization and inference on manifolds

Lizhen Lin\textsuperscript{1}, Drew Lazar\textsuperscript{2}, Bayan Saparpabayeva\textsuperscript{3} and David Dunson\textsuperscript{4}

\textit{The University of Notre Dame, Ball State University, The University of Rochester and Duke University}

Supplementary Material

This supplementary material contains the proof for the main theorems in the paper and the algorithms for computing various summary statistics on the sphere $S^d$.

S1 Proofs of Propositions and Theorems

\textbf{Proposition 1} (Proposition 2.1 of the main paper). Let $S^d = \{ p \in \mathbb{R}^{d+1} : \|p\| = 1 \}$ which is the $d$-dimensional sphere. The inverse exponential map, $\log_p$, on $S^d$ is $2$-Lipschitz continuous from $B(p, \pi/2)$ to $T_p S^d$ for all $p \in S^d$.

\textit{Proof.} The tangent space at $p$ is given as

$$T_p S^d = \{ v \in \mathbb{R}^{d+1} : v^T p = 0 \}. $$
Then for \( q \in S^d \) the inverse exponential map can be expressed as

\[
\log_p(q) = \frac{\arccos(p^T q)}{\sqrt{1 - (p^T q)^2}} (q - (p^T q)p).
\]

Hence, the distance between \( \log_p q_1 \) and \( \log_p q_2 \) is equal to

\[
\| \log_p q_1 - \log_p q_2 \| = \sqrt{\arccos(p^T q_1)^2 + \arccos(p^T q_2)^2 - 2 \arccos(p^T q_1) \arccos(p^T q_2) \cos \varphi}
\]

where \( \varphi \) is the angle between \( \log_p q_1 \) and \( \log_p q_2 \). The geodesic distance between \( q_1 \) and \( q_2 \) is then given by

\[
d_g(q_1, q_2) = \arccos(q_1^T q_2).
\]

One can easily obtain that

\[
q_1^T q_2 = (p^T q_1)(p^T q_2) + \sqrt{1 - (p^T q_1)^2} \sqrt{1 - (p^T q_2)^2} \cos \varphi.
\]

Then one can check directly that

\[
\| \log_p q_1 - \log_p q_2 \| \leq 2d_g(q_1, q_2).
\]

The following proposition shows that the log map in similarity-shape spaces [Kendall 1984] also satisfies the \( K- \) Lipschitz condition.

**Proposition 2.** *Proposition 2.2 of the main paper*

The similarity or planar shape space is given as

\[
\Sigma^k_2 = S^{2k-3}/S^1.
\]  \hspace{1cm} (S1.1)
The inverse exponential map, \( \log_p \), on \( \Sigma_2^k \) is 2-Lipschitz continuous from \( B(p, \pi/4) \) to \( T_p \Sigma_2^k \) for all \( p \in \Sigma_2^k \).

**Proof.** \( \Sigma_2^k \) is the quotient of the sphere \( S^{2k-3} \) under the following group of transformations

\[
G = \left\{ \begin{pmatrix} A & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & A \end{pmatrix} \in \text{M}(2k), \ A \in \text{SO}(2) \right\} \simeq S^1.
\]

For any \( B \in G \), we have that \( B = \cos t I + \sin t \dot{I} \), where

\[
I = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}, \quad \dot{I} = \begin{pmatrix} 0 & 1 & \ldots & 0 & 0 \\ -1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & -1 & 0 \end{pmatrix}.
\]

For each \( p \in \Sigma_2^k \) we define the tangent space

\[
T_p \Sigma_2^k = \{ v \in \mathbb{R}^{2k-2} : v^T p = 0, (Ip)^T v = 0 \}.
\]

The inverse exponential map can be expressed as

\[
\log_p(q) = \frac{\arccos(p^T q)}{\sqrt{1 - (p^T q)^2}}(q - (p^T q)p).
\]

Hence, the distance between \( \log_p q_1 \) and \( \log_p q_2 \) is equal to

\[
\| \log_p q_1 - \log_p q_2 \| = \sqrt{\arccos(p^T q_1)^2 + \arccos(p^T q_2)^2 - 2 \arccos(p^T q_1) \arccos(p^T q_2) \cos \varphi},
\]
where \( \varphi \) is an angle between \( \log p q_1 \) and \( \log p q_2 \). The geodesic distance between \( q_1 \) and \( q_2 \) is then given by

\[
d_g(q_1, q_2) = \inf_{t \in (-\pi, \pi]} \arccos (q_1^T (\cos t \hat{I} + \sin t \hat{I}) q_2)
\]

\[
= \arccos \sup_{t \in (-\pi, \pi]} (\cos t q_1^T q_2 + \sin t q_1^T \hat{I} q_2)
\]

\[
= \arccos \sqrt{(q_1^T q_2)^2 + (q_1^T \hat{I} q_2)^2}.
\]

One can easily obtain that

\[
(q_1^T q_2)^2 + (q_1^T \hat{I} q_2)^2 = (p^T q_1)(p^T q_2) + \sqrt{1 - (p^T q_1)^2} \sqrt{1 - (p^T q_2)^2} \cos \varphi
\]

\[
+ (1 - (p^T q_1)^2)(1 - (p^T q_2)^2)(\cos \psi)^2
\]

where \( \psi \) is the angle between \( \log p q_1 \) and \( \hat{I} \log p q_2 \). Note that

\[
\pi/2 - \varphi \leq \psi \leq \pi/2 + \varphi.
\]

Thus \( \cos \psi \geq \cos(\pi/2 - \varphi) = \sin \varphi \), and

\[
d_g(q_1, q_2) \geq 2 \arccos \left( (p^T q_1)(p^T q_2) + \sqrt{1 - (p^T q_1)^2} \sqrt{1 - (p^T q_2)^2} \cos \varphi \right)^2
\]

\[
+ (1 - (p^T q_1)^2)(1 - (p^T q_2)^2)(\sin \varphi)^2 \right)^{1/2}.
\]

Then it can be verified directly that

\[
\| \log p q_1 - \log p q_2 \| \leq 2 \arccos \left( (p^T q_1)(p^T q_2) + \sqrt{1 - (p^T q_1)^2} \sqrt{1 - (p^T q_2)^2} \cos \varphi \right)^2
\]

\[
+ (1 - (p^T q_1)^2)(1 - (p^T q_2)^2)(\sin \varphi)^2 \right)^{1/2}.
\]
S1. PROOFS OF PROPOSITIONS AND THEOREMS

Thus \( \| \log_p q_1 - \log_p q_2 \| \leq 2d_g(q_1, q_2) \).

**Proposition 3.** Proposition 2.3 of the main paper] The manifold of positive definite \( n \) by \( n \) matrices \( \text{PD}(n) \) has a 1-Lipchitz continuous inverse exponential map at any \( p \in \text{PD}(n) \).

*Proof.* We consider the Killing metric [Dubrovin, Fomenko, and Novikov (Dubrovin et al.)] in the manifold of invertible \( n \) by \( n \) matrices \( \text{GL}(n) \)

\[
ds^2(a) = \text{tr}(a^{-1}da)^2.
\]

In other words, in the Lie algebra \( \mathfrak{gl}(n) = T_I \text{GL}(n) = \text{M}(n) \), we have the symmetric inner product

\[
\langle A, B \rangle_I = \text{tr}(AB), \quad A, B \in \mathfrak{gl}(n),
\]

which generates the bilaterally invariant metric in the group \( \text{GL}(n) \). That is, for any \( A, B \in T_g \text{GL}(n) \) and \( a \in \text{GL}(n) \),

\[
\langle A, B \rangle_a = \langle a^{-1}A, a^{-1}B \rangle_I = \langle Aa^{-1}, Ba^{-1} \rangle_I = \text{tr}(a^{-1}Aa^{-1}B).
\]

Since vectors \( p^{-1}A, p^{-1}B \) do not always belong to the tangent space \( T_I \text{PD}(n) \), we instead take vectors \( p^{-1/2}Ap^{-1/2} \) and

\[
\langle A, B \rangle_p = \langle p^{-1/2}A, p^{-1/2}B \rangle_{p^{1/2}}
\]

\[
= \langle p^{-1/2}Ap^{-1/2}, p^{-1/2}Bp^{-1/2} \rangle_I
\]

\[
= \text{tr}(p^{-1/2}Ap^{-1/2}Bp^{-1/2}) = \text{tr}(p^{-1}Ap^{-1}B),
\]

\[
= \text{tr}(p^{-1}Ap^{-1}B).
\]
where $A, B \in T_p \text{PD}(n)$. Hence we have the metric in $\text{PD}(n)$ induced from the Killing metric in $\text{GL}(n)$. This metric is usually known as the Fisher-Rao metric.

For this metric we have the following exponential and logarithm mappings

\[ \exp_p A = p^{1/2} \exp \left( p^{-1/2} Ap^{-1/2} \right) p^{1/2}, \]

\[ \log_p q = p^{1/2} \log \left( p^{-1/2} qp^{-1/2} \right) p^{1/2}, \]

where

\[ \exp Y = I + \frac{Y}{1!} + \frac{Y^2}{2!} + \ldots + \frac{Y^n}{n!} + \ldots, \]

\[ \log x = (x - I) - \frac{(x - I)^2}{2} + \ldots + (-1)^{n-1} \frac{(x - I)^n}{n} + \ldots \]

for any $A, Y \in \text{Sym}(n)$ and $p, q, x \in \text{PD}(n)$.

Let $a, q_1, q_2 \in \text{PD}(n)$. Then we have

\[ \| \log_a q_1 - \log_a q_2 \|_a = \| \log(a^{-1/2} q_1 a^{-1/2}) - \log(a^{-1/2} q_2 a^{-1/2}) \|_I \]

\[ \leq d_g(a^{-1/2} q_1 a^{-1/2}, a^{-1/2} q_2 a^{-1/2}) = d_g(q_1, q_2) \]

where the inequality follows from the exponential metric increasing property of the Fisher-Rao metric as in [Bhatia (2003)].

\[ \square \]

**Theorem 1** (Theorem 1 of the main paper). Let $\mu_1, \ldots, \mu_m$ be a collection of independent estimators of the parameter $\mu$, and let geometric median $\mu^* = \text{med}(\mu_1, \ldots, \mu_m)$. 


(a) Let \( \rho \) be the extrinsic distance on \( \mathcal{M} \) for some embedding \( J : \mathcal{M} \to \tilde{\mathcal{M}} \subset \mathbb{R}^D \). Assume for any \( \omega \in \mathcal{M} \) the angle between \( J(\omega) - J(\mu^*) \) and the tangent space \( T_{J(\mu^*)}\tilde{\mathcal{M}} \) is no bigger than \( \bar{\psi} \). For any \( \alpha \in (0, \cot \bar{\psi} \tan \frac{\bar{\psi}}{2}) \) set

\[
\bar{C}_\alpha = \frac{1 - \alpha}{\sqrt{1 - 2\alpha \cos \bar{\psi} - \alpha \sin \bar{\psi}}}
\]

(b) Let \( \rho \) be an intrinsic distance on \( \mathcal{M} \) with respect to some Riemannian structure. Assume \( \log_{\mu^*} \) is \( K \)-Lipschitz continuous from \( B(\mu^*, \epsilon) \) to \( T_{\mu^*}\mathcal{M} \). For any \( \alpha \in (0, \frac{1}{2}) \) set

\[
C_\alpha = K(1 - \alpha) \sqrt{\frac{1}{1 - 2\alpha}}.
\]

Under (a) or (b), if

\[
P(\rho(\mu_j, \mu) > \epsilon) \leq \eta \text{ for } i = 1, \ldots, n
\]

(S1.2)

where \( \eta < \alpha \) then

\[
P(\rho(\mu^*, \mu) > \bar{C}_\alpha \epsilon) \leq \exp(-m\phi(\alpha, \eta)),
\]

(S1.3)

where

\[
\phi(\alpha, \eta) = (1 - \alpha) \log \frac{1 - \alpha}{1 - \eta} + \alpha \log \frac{\alpha}{\eta}.
\]

Proof. Let \( \psi \) be the angle between \( J(\mu) - J(\mu^*) \) and the tangent space \( T_{\mu^*}\tilde{\mathcal{M}} \). Since \( \psi < \bar{\psi} \) we have \( C_\alpha \leq \bar{C}_\alpha \) and \( \cot \bar{\psi} \tan \frac{\bar{\psi}}{2} \leq \cot \psi \tan \frac{\psi}{2} \) where

\[
C_\alpha = \frac{1 - \alpha}{\sqrt{1 - 2\alpha \cos \psi - \alpha \sin \psi}}.
\]
Thus, when the event \( \{ \rho(\mu^*, \mu) > C_\alpha \epsilon \} \) occurs, the event \( \{ \rho(\mu^*, \mu) > C_\alpha \epsilon \} \) occurs. Then, by Lemma 1, when the event \( \{ \rho(\mu^*, \mu) > C_\alpha \epsilon \} \) occurs, there exists an \( \alpha \) portion of elements of \( \mu_1, \ldots, \mu_m \) which are at least \( \epsilon \) distance away from \( \mu \). Therefore,

\[
P(\rho(\mu^*, \mu) > \bar{C}_\alpha \epsilon) \leq P(\rho(\mu^*, \mu) > C_\alpha \epsilon) \leq P\left( \sum_{j=1}^{m} I_{\rho(\mu_j, \mu) > \epsilon} > \alpha m \right).
\]

(S1.4)

Let \( A = |\{ j = 1, \ldots, m : \rho(\mu_j, \mu) > \epsilon \}| \) and let \( B \) be a random variable with a binomial distribution, \( B \sim b(m, \eta) \). Then with (S1.2) and by Lemma 23 in [Lerasle and Oliveira (2011)] there exists a coupling \( C = (\tilde{A}, \tilde{B}) \) such that \( \tilde{A} \) has the same distribution as \( A \) and \( \tilde{B} \) has the same distribution as \( B \) such that \( \tilde{A} \leq \tilde{B} \). Hence

\[
P(A > \alpha m) \leq P(B > \alpha m) \leq \exp(-m\phi(\alpha, \eta))
\]

where the second inequality follows from Chernoff’s bound. Then with (S1.4) we have

\[
P(\rho(\mu^*, \mu) > \bar{C}_\alpha \epsilon) \leq \exp(-m\phi(\alpha, \eta)).
\]

For the intrinsic case (b) we have a similar proof. \( \square \)

## S2 Computation of Sample Statistics on \( S^d \)

Given \( \{p_1, \ldots, p_n\} \subset S^d \) we compute sample statistics as follows:
S2. COMPUTATION OF SAMPLE STATISTICS ON $S^D$

1. **Intrinsic mean.** With objective function $L_n(x) = \frac{1}{n} \sum_{i=1}^{n} \arccos^2(\langle x, p_i \rangle)$ and constraint function $g(x) = \langle x, x \rangle$ let

$$\gamma_i(x) = \frac{\arccos(\langle x, p_i \rangle)}{\sqrt{1 - \langle x, p_i \rangle^2}}.$$ 

Then the sample mean $\hat{\mu}$ satisfies Lagrange multiplier condition

$$\sum_{i=1}^{n} \gamma_i(\hat{\mu}) p_i = \lambda \hat{\mu}$$

with $\langle \hat{\mu}, \hat{\mu} \rangle = 1$ and $\lambda = \sum_{i=1}^{n} \gamma_i(\hat{\mu}) \langle p_i, \hat{\mu} \rangle$.

As in [Huckemann and Ziezold (2006)], letting $\Psi(x) = \sum_{i=1}^{n} \gamma_i(x) p_i$, we use the fixed-point algorithm

$$\mu_k \mapsto \mu_{k+1}$$

$$\mu_{k+1} = \frac{\Psi(\mu_k)}{\|\Psi(\hat{\mu}_k)\|}.$$ 

Then $\mu_k \to \hat{\mu}$.

2. **Intrinsic median.** We use a generalization of Ostresh’s modification of Weiszfeld’s algorithm as introduced in [Fletcher et al. (2008)]. Let

$$\Psi(x) = \sum_i \frac{\log(p_i)}{\arccos(\langle x, p_i \rangle)} \left( \sum_i \frac{1}{\arccos(\langle x, p_i \rangle)} \right)^{-1}$$

$$m_k \mapsto m_{k+1}$$

$$m_{k+1} = \text{Exp}_{x_k}(\Psi(m_k)).$$

Then $m_k \to \hat{m}$ where $\hat{m}$ is the intrinsic sample median.
3. **Extrinsic mean.** As in Bhattacharya and Bhattacharya (2012), the extrinsic sample mean is the projection of the sample mean under the embedding. That is,

\[ \hat{\mu} = P \left( \frac{1}{n} \sum_{i=1}^{n} J(p_i) \right) \]

where \( J \) is our embedding map. In the case of \( S^d \), where \( J \) is the identity map and projection is done by normalizing in \( \mathbb{R}^{d+1} \), \( \mu = \bar{x}/\|\bar{x}\| \) where \( \bar{x} \) is the Euclidean sample mean.

4. **Extrinsic median.** Let

\[ L_n(x) = \frac{1}{n} \sum_{i=1}^{n} ||x - p_i|| \] for \( x \in \mathbb{R}^{d+1} \) and \( g(x) = L_n|_{S^d} \).

With \( S^d \) as a submanifold of \( \mathbb{R}^{d+1} \), for \( p \in S^d \) the gradient of \( g \) is the orthogonal projection of \( \nabla_p L_n \) onto \( T_p S^d \), that is,

\[ \nabla_p g = \text{proj}_{T_p S^d}(\nabla_p L_n). \]

We take \( \nabla_p L_n \) as in Weiszfeld’s algorithm Weiszfeld (2009) and compute the sample geometric median \( \hat{m} \) by gradient descent as follows:

\[ \Psi(x) = \sum_i \frac{p_i - x}{\|p_i - x\|} \left( \sum_i \frac{1}{\|p_i - x\|} \right)^{-1} \]
S2. COMPUTATION OF SAMPLE STATISTICS ON $S^D$

\[ m_k \mapsto m_{k+1} \]

\[ m_{k+1} = \text{Exp}_{m_k} \left( \text{proj}_{T_{m_k}S^d}(\Psi(m_k)) \right) \]

Then $m_k \to \hat{m}$, the extrinsic sample median.
References


