Optimisation on Riemannian manifolds for uncertainty quantification

ETICS

Baalu KETEMA

IMT Advisors: Fabrice GAMBOA, Francesco COSTANTINO
EDF R&D Advisors: Roman SUEUR, Nicolas BOUSQUET, Bertrand IOOSS

October 9 to 13, 2023

I. Introduction

Evaluate the robustness of

$$G(X) = Y$$

w.r.t. distributions of vector X

- $\{P_\theta\}_{\theta\in\Theta}$ = possible distributions for X
- $\mathrm{QoI}(Y^\theta)$ = quantity of interest on $Y^\theta := G(X^\theta)$ where $X^\theta \sim P_\theta$

Define the following function (called PLI [Lemaître, 2015])

$$S_\theta = \frac{\mathrm{QoI}(Y^\theta) - \mathrm{QoI}(Y^{\theta_0})}{\mathrm{QoI}(Y^{\theta_0})},$$

where $\theta_0 \in \Theta$ is a fixed reference parameter

I. Introduction

Consider

$$\min_{\theta \in B_\delta(\theta_0)} S_\theta \quad \text{and} \quad \max_{\theta \in B_\delta(\theta_0)} S_\theta \quad (\star),$$

where $B_\delta(\theta_0) \subset \Theta$ is a closed ball centered at θ_0 with radius $\delta > 0$ for the Fisher-Rao distance d

I. Introduction

Consider

$$\min_{\theta \in B_\delta(\theta_0)} S_\theta \quad \text{and} \quad \max_{\theta \in B_\delta(\theta_0)} S_\theta \quad (\star),$$

where $B_\delta(\theta_0) \subset \Theta$ is a closed ball centered at θ_0 with radius $\delta > 0$ for the Fisher-Rao distance d

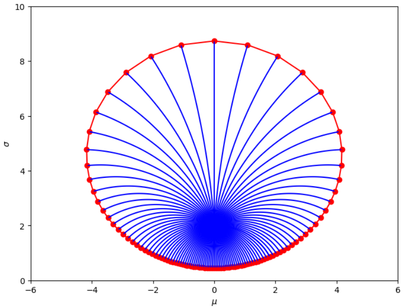

Figure: Fisher ball for $\{\mathcal{N}(\mu, \sigma)\}_{(\mu,\sigma)\in\Theta}$

Fisher geodesic distance

This distance is obtained from the information geometry of the family $\{P_\theta\}_{\theta\in\Theta}$ i.e. from

$$(I_\theta)_{ij} = \mathbb{E}_{X\sim P_\theta}\left[\partial_i \log p_\theta(X)\partial_j \log p_\theta(X)\right],$$

which is a Riemannian metric on Θ

Fisher geodesic distance

This distance is obtained from the information geometry of the family $\{P_\theta\}_{\theta\in\Theta}$ i.e. from

$$(I_\theta)_{ij} = \mathbb{E}_{X\sim P_\theta}\left[\partial_i \log p_\theta(X)\partial_j \log p_\theta(X)\right],$$

which is a Riemannian metric on Θ

In this setting, $(\star)$ is an optimization problem on a Riemannian manifold

$\rightarrow$ This leads us to consider Riemannian optimization algorithms

Starting point

Our work is in the continuation of the paper "An information geometry approach to robustness analysis for the uncertainty quantification of computer codes" [Gauchy et al., 2022]

The Fisher distance presents good properties (invariance under reparametrization, measures dissimilarity,...) and gives more interpretability than previously used robustness analysis methods

Our main goals are:

- in depth study of the induced geometry from the Fisher matrices,
- develop adapted optimization algorithms for problem $(\star)$

II. Why **Riemannian** optimization ?

The problem

$$\min_{x \in E} \; f(x)$$

is a **Riemannian optimization** problem when E is a Riemannian manifold and f is a differentiable function on E

A **manifold** M is a “curved” space that locally “looks” flat

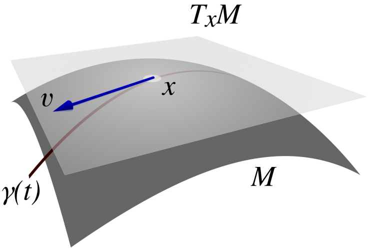

Figure: Manifold and tangent space

Source: https://en.wikipedia.org/wiki/Tangent_space

II. Why **Riemannian** optimization ?

Some simple optimization problems are naturally manifold optimization

1st example (N. Boumal, 2014) : Finding eigenvector v_1 with smallest eigenvalue λ_1 of a symmetric matrix A

Eigenvector v_1 minimizes the Rayleigh quotient

$$r : \mathbb{R}^d \backslash \{0\} \to \mathbb{R} : r(x) = \frac{\langle Ax, x \rangle}{\langle x, x \rangle}$$

r is invariant under scaling, v_1 (normalized) solves

$$\min_{x \in \mathbb{S}^{d-1}} \langle Ax, x \rangle$$

II. Why **Riemannian** optimization ?

2nd example (N. Boumal 2014) : PCA for $y_1, ..., y_n$ data points in $\mathbb{R}^d$

Define the **Grassmann manifold** $\mathrm{Gr}(k, d)$ as the set of k-dimensional subspaces of $\mathbb{R}^d$ and consider

$$\min_{L \in \mathrm{Gr(k,d)}} \sum_{i=1}^{n} \mathrm{dist}(L, y_i)^2$$

II. Why **Riemannian** optimization ?

2nd example (N. Boumal 2014) : PCA for $y_1, ..., y_n$ data points in $\mathbb{R}^d$

Define the **Grassmann manifold** $\mathrm{Gr}(k, d)$ as the set of k-dimensional subspaces of $\mathbb{R}^d$ and consider

$$\min_{L \in \mathrm{Gr(k,d)}} \sum_{i=1}^{n} \mathrm{dist}(L, y_i)^2$$

Gr(k,d) can be identified to the following **quotient manifold**

$$M = \{X \in \mathcal{M}_{d,k}(\mathbb{R}) \mid X^\intercal X = \mathrm{id}_k\}/_{O(k)}$$

where $O(k) = \{Q \in \mathcal{M}_k(\mathbb{R}) \mid Q^\intercal Q = \mathrm{id}_k\}$ is the orthogonal group and $L = \mathrm{span}(X)$

II. Why **Riemannian** optimization ?

2nd example (N. Boumal 2014) : PCA for $y_1, \ldots, y_n$ data points in $\mathbb{R}^d$

Define the **Grassmann manifold** $\mathrm{Gr}(k,d)$ as the set of k-dimensional subspaces of $\mathbb{R}^d$ and consider

$$\min_{L \in \mathrm{Gr(k,d)}} \sum_{i=1}^{n} \mathrm{dist}(L, y_i)^2$$

Gr(k,d) can be identified to the following **quotient manifold**

$$M = \{X \in \mathcal{M}_{d,k}(\mathbb{R}) \mid X^\intercal X = \mathrm{id}_k\} / _{O(k)}$$

where $O(k) = \{Q \in \mathcal{M}_k(\mathbb{R}) \mid Q^\intercal Q = \mathrm{id}_k\}$ is the orthogonal group and $L = \mathrm{span}(X)$

It can be endowed with a metric g (Frobenius inner product), PCA is an optimization problem on a Riemannian quotient manifold

III. Riemannian optimization algorithms

Examples of Riemannian optimization algorithms for

$$\min_{x \in E} \; f(x),$$

where E is a manifold and f is differentiable

1. Gradient descent : we choose a starting point x_0 and define

$$x_{n+1} := \exp_{x_n}\big(-\varepsilon_n \cdot \nabla_x f(x_n)\big),$$

where $\varepsilon_n > 0$ are the step sizes and ∇f is the Riemannian gradient

III. Riemannian optimization algorithms

2. Newton's method : if f is twice differentiable, then we can define

$$x_{n+1} := \exp_{x_n}\big(-(\mathrm{Hess}_{x_n} f)^{-1} \cdot \nabla f(x_n)\big),$$

where $\mathrm{Hess}_x : T_x M \to T_x M$ is the Riemannian Hessian operator

III. Riemannian optimization algorithms

2. Newton's method : if f is twice differentiable, then we can define

$$x_{n+1} := \exp_{x_n}\big(-(\mathrm{Hess}_{x_n} f)^{-1} \cdot \nabla f(x_n)\big),$$

where $\mathrm{Hess}_x : T_x M \to T_x M$ is the Riemannian Hessian operator

3. Stochastic gradient descent : if f is given by

$$f(x) = \mathbb{E}_{Z \sim \mu}[h(x, Z)],$$

we can build the following algorithm

$$x_{n+1} = \exp_{x_n}\big(-\varepsilon_n \cdot \nabla_x h(x_n, Z_{n+1})\big),$$

where $Z_i \in \mathcal{Z}$ are iid samples from μ

IV. Riemannian barycenter estimation on $\mathbb{S}^2$

Example from [S. Bonnabel, 2013], given $y_1, \ldots, y_K$ in $\mathbb{S}^2$ we will solve

$$\min_{x \in \mathbb{S}^2} \frac{1}{2N} \sum_{i=1}^{N} d(x, y_i)^2$$

to compute the Riemannian Karcher (Fréchet) mean on $\mathbb{S}^2$

IV. Riemannian barycenter estimation on $\mathbb{S}^2$

Example from [S. Bonnabel, 2013], given $y_1, \ldots, y_K$ in $\mathbb{S}^2$ we will solve

$$\min_{x \in \mathbb{S}^2} \; \frac{1}{2N} \sum_{i=1}^{N} d(x, y_i)^2$$

to compute the Riemannian Karcher (Fréchet) mean on $\mathbb{S}^2$

Rewrite this problem as

$$\min_{x \in \mathbb{S}^2} \; \mathbb{E}_U \left[\frac{1}{2} d(x, y_U)^2 \right]$$

where U is uniform on $\{1, \ldots, K\}$ and apply the stochastic gradient descent algorithm

$$x_{n+1} = \exp_{x_n} \left(-\varepsilon_n \cdot \nabla_x \frac{1}{2} d(x_n, y_{U_{n+1}})^2 \right),$$

where $(U_i)_i \sim \mathcal{U}(\{1, \ldots, K\})$ and $\varepsilon_n = \frac{cst}{n}$

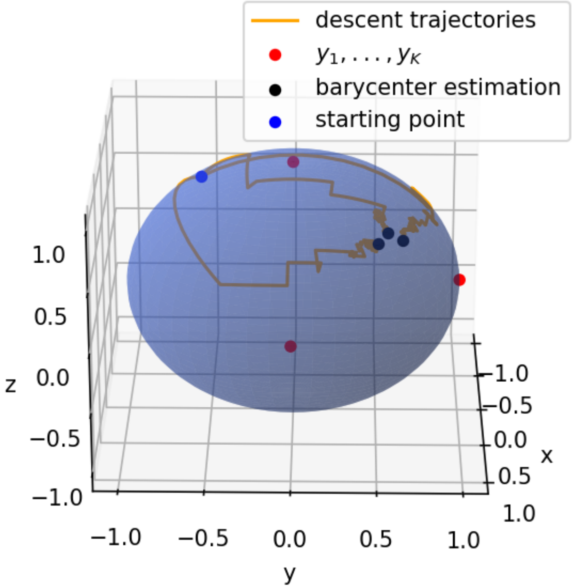

Figure: Barycenter estimation of 3 points on $\mathbb{S}^2$

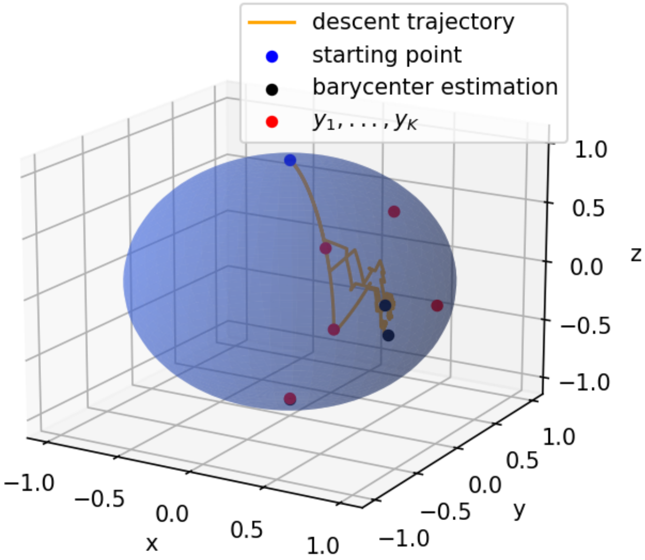

Figure: Barycenter estimation of 5 points on $\mathbb{S}^2$

Conclusion and works in progress

Our initial optimization problem was

$$\min_{\theta \in B_\delta(\theta_0)} S_\theta,$$

where

$$S_\theta = \frac{\mathrm{QoI}(Y^\theta) - \mathrm{QoI}(Y^{\theta_0})}{\mathrm{QoI}(Y^{\theta_0})}$$

It is indeed a Riemannian manifold optimization, but S_θ is difficult to compute and is estimated using importance sampling

Conclusion and works in progress

Our initial optimization problem was

$$\min_{\theta \in B_\delta(\theta_0)} S_\theta,$$

where

$$S_\theta = \frac{\mathrm{QoI}(Y^\theta) - \mathrm{QoI}(Y^{\theta_0})}{\mathrm{QoI}(Y^{\theta_0})}$$

It is indeed a Riemannian manifold optimization, but S_θ is difficult to compute and is estimated using importance sampling

1. Asymptotic/Non-asymptotic confidence intervals

We established a non-asymptotic confidence interval for q_θ^α : given $s > 0$ and $\theta \in \Theta$

$$\mathbb{P}\big(q_\theta^\alpha \in [q^-(\alpha), q^+(\alpha)]\big) \geq 1 - 2N^r \varepsilon_{s,\theta},$$

where q^- and q^+ depend on the sample $X_1, ..., X_N \sim P_{\theta_0}$

Conclusion and works in progress

2. Geometry of truncated distributions

Implement physical constraints on inputs on the Robustness Analysis method

For instance, for an input $X_i \sim \mathcal{N}(\mu, \sigma)$ with constraint $X_i \in [a, b]$, we studied the family of truncated Gaussian distributions

$$q_{(\mu,\sigma)}(x) = \frac{1}{P_{(\mu,\sigma)}([a,b])} p_{(\mu,\sigma)}(x) \mathbf{1}_{x \in [a,b]},$$

namely:

- Fisher matrices $\rightarrow$ defines a new geometry on $\mathbb{H}$,
- numerically compute geodesics and spheres

References

C. Gauchy, J. Stenger, R. Sueur, B. Iooss
An information geometry approach to robustness analysis for the uncertainty quantification of computer codes
Technometrics, 2022, vol. 64, no 1, p. 80-91.

P. Lemaître, E. Sergienko, A. Arnaud, N. Bousquet, F. Gamboa, B. Iooss
Density modification-based reliability sensitivity analysis
Journal of Statistical Computation and Simulation, 2015, vol. 85, no 6, p. 1200-1223

T. Labopin-Richard, F. Gamboa, A. Garivier, B. Iooss
Bregman superquantiles. Estimation methods and applications
Dependence Modeling, 2016, vol. 4, no 1

S. Chatterjee and P. Diaconis,
The sample size required in importance sampling
The Annals of Applied Probability, 2018, vol. 28, no 2, p. 1099-1135.

N. Boumal
An introduction to optimization on smooth manifolds
Cambridge University Press, 2023

S. Bonnabel
Stochastic gradient descent on Riemannian manifolds
IEEE Transactions on Automatic Control, 2013, vol. 58, no 9, p. 2217-2229.

S.T. Smith
Optimization techniques on Riemannian manifolds
arXiv preprint arXiv:1407.5965, 2014

S.-I. Amari, S. Douglas
Why natural gradient ?
Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998, vol. 2, p. 1213-1216

Appendix

Convergence theorem for the Riemannian version of **Newton's method**

$$x_{n+1} := \exp_{x_n}\big(-(\mathrm{Hess}_{x_n} f)^{-1} \cdot \nabla_x f(x_n)\big).$$

Theorem (S.T. Smith, 2014)

Assume that

- *(E, d) is a complete metric space (geodesically complete),*
- *there exists x_∞ nondegenerate critical point,*

then there exists a neighborhood U of x_∞ (domain of attraction) such that if $x_0 \in U$, then x_n converges quadratically to x_∞:

$$d(x_n, x_\infty) \underset{n\to\infty}{=} \mathcal{O}(n^{-2}).$$

Convergence theorem for **Riemannian stochastic gradient descent** algorithm i.e. when the function f is given by $f(x) = \mathbb{E}_{Z\sim\mu}[h(x, Z)]$. The iteration is given by

$$x_{n+1} = \exp_{x_n}\big(-\varepsilon_n \cdot \nabla_x h(x_n, Z_{n+1})\big),$$

where $(Z_i)_i$ are iid samples from μ.

Theorem (S. Bonnabel, 2013)

Assume that:

- *the manifold E is connected with injectivity radius $I > 0$,*
- *the step size ε_n verify $\sum_n \varepsilon_n = \infty$ and $\sum_n \varepsilon_n^2 < \infty$,*
- *we have $\nabla f(x) = \mathbb{E}_{Z\sim\mu}[\nabla_x h(x, Z)]$,*
- *there exists $K \subset E$ compact such that $x_n \in K$ for all n,*
- *$\nabla_x h$ is bounded on K i.e. $\sup_{x\in K, z\in\mathcal{Z}} |\nabla_x h(x, z)| < \infty$.*

Therefore, we have

$$\big(f(x_n)\big)_{n\geq 0} \ \textit{converges a.s. and} \quad \nabla f(x_n) \underset{n\to\infty}{\to} 0 \ \textit{a.s.}.$$

3rd example (S.-I. Amari 1998) : In our context, the manifold is given by $M = \{P_\theta\}_{\theta \in \Theta}$ endowed with the Fisher information metric

$$(I_\theta)_{ij} = \mathbb{E}_{X \sim P_\theta}\left[\partial_i \log p_\theta(X) \partial_j \log p_\theta(X)\right]$$

To estimate a parameter θ^*, minimize the KL divergence of P_{θ^*} from P_θ :

$$\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmin}} \; \mathbb{E}_{X \sim \theta^*}\left[\log\left(\frac{p_{\theta^*}(X)}{p_\theta(X)}\right)\right]$$

this is the same problem as

$$\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmax}} \; \mathbb{E}_{X \sim \theta^*}\left[\log p_\theta(X)\right]$$

$$\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmax}} \ \mathbb{E}_{X \sim \theta^*}\left[\log p_\theta(X)\right]$$

Given $X_1, ..., X_N \sim P_{\theta^*}$, estimate θ^* using **gradient descent**

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n + \frac{1}{n} \nabla_\theta \log p_{\tilde{\theta}_n}(X_{n+1})$$

which is consistent but not Fisher efficient in general

$$\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmax}} \ \mathbb{E}_{X \sim \theta^*}\left[\log p_\theta(X)\right]$$

Given $X_1, ..., X_N \sim P_{\theta^*}$, estimate θ^* using **gradient descent**

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n + \frac{1}{n} \nabla_\theta \log p_{\tilde{\theta}_n}(X_{n+1})$$

which is consistent but not Fisher efficient in general

But the following update called **natural gradient descent**

$$\widehat{\theta}_{n+1} = \widehat{\theta}_n + \frac{1}{n} \ I_{\widehat{\theta}_n}^{-1} \ \nabla_\theta \log p_{\widehat{\theta}_n}(X_{n+1}) \qquad (\star),$$

gives a Fisher efficient estimator [Amari, 1998] i.e.

$$\lim_{N \to \infty} N \ \mathbb{E}[(\widehat{\theta}_N - \theta_*)(\widehat{\theta}_N - \theta_*)^\intercal] = I_\theta^{-1}$$