Simultaneous Optimization of Geodesics and Fréchet Means

Frederik Möbius Rygaard[1*], Søren Hauberg[1*] and Steen Markvorsen[1*]

[1*]DTU Compute, Technical University of Denmark (DTU), Anker Engelundsvej 1, Kongens Lyngby, 2800, Denmark.

*Corresponding author(s). E-mail(s): fmry@dtu.dk; sohau@dtu.dk; stema@dtu.dk;

Abstract

A central part of geometric statistics is to compute the Fréchet mean. This is a well-known intrinsic mean on a Riemannian manifold that minimizes the sum of squared Riemannian distances from the mean point to all other data points. The Fréchet mean is simple to define and generalizes the Euclidean mean, but for most manifolds even minimizing the Riemannian distance involves solving an optimization problem. Therefore, numerical computations of the Fréchet mean require solving an embedded optimization problem in each iteration. We introduce the *GEORCE-FM* algorithm to simultaneously compute the Fréchet mean and Riemannian distances in each iteration in a local chart, making it faster than previous methods. We extend the algorithm to Finsler manifolds and introduce an adaptive extension such that *GEORCE-FM* scales to a large number of data points. Theoretically, we show that *GEORCE-FM* has global convergence and local quadratic convergence and prove that the adaptive extension converges in expectation to the Fréchet mean. We further empirically demonstrate that *GEORCE-FM* outperforms existing baseline methods to estimate the Fréchet mean in terms of both accuracy and runtime.

Keywords: Riemannian Manifolds, Finsler Manifolds, Fréchet Mean, Control Problem , Adaptive Optimization

Mathematics Subject Classification 53C22, 49Q99, 65D15

1 Introduction

Computing means on Riemannian manifolds is hard. In Euclidean space, computing means is taken for granted, due to its simple closed-form expression, which is not the case for general Riemannian manifolds. At the same time, Riemannian manifolds play an increasing role in data analysis with application to, e.g., generative models [1–4], robotics [5–7], and protein modeling [8–10].

At the center of Riemannian statistics lies the concept of the *Fréchet mean* [11] as the building block for more elaborate statistical models [12–17]. The Fréchet mean naturally generalizes the Euclidean mean, which, in the discrete case, minimizes the sum of squared distances to all data points [11],

$$\mu = \underset{y \in \mathcal{M}}{\arg\min} \sum_{i=1}^{N} \mathrm{dist}^2(y, a_i), \tag{1}$$

where $\{a_i\} \subset \mathcal{M}$ are the given data points on a Riemannian manifold, $(\mathcal{M}, g)$, with metric tensor g and induced distance function, $\mathrm{dist} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$. Only for a few selected families of Riemannian manifolds is the distance known in closed form, and the Fréchet mean can be computed in such cases using, for instance, Riemannian gradient descent [14, 18]. For general Riemannian manifolds, the shortest curves, known as *geodesics*, have to be computed by either solving a system of (typically non-linear) ordinary differential equations (ODE) [19, Page 62] or by minimizing the energy functional with respect to curves γ [19, Page 194]

$$\mathcal{E}_\gamma(a, b) = \frac{1}{2} \int_0^1 \dot{\gamma}^\top(t) G(\gamma(t)) \dot{\gamma}(t) \, \mathrm{d}t, \quad \gamma(0) = a, \gamma(1) = b, \tag{2}$$

where $a, b \in \mathcal{U}$ are the fixed start and endpoint of the curve γ, respectively, where $\phi : U \to \mathcal{M}$ is a local chart for an open subset $U \subseteq \mathbb{R}^d$. Thus, for general Riemannian manifolds, the Fréchet mean is iteratively updated by constructing geodesics minimizing Eq. 2 with respect to a curve γ connecting each data point and the candidate of the Fréchet mean. This makes computations of the Fréchet mean time consuming and possibly intractable in practice.

Even computing geodesics inside a geodesically strongly convex set by minimizing Eq. 2 can in itself be complicated and time consuming. Many off-the-shelf algorithms, such as *ADAM* [20] and *BFGS* [21–24], either scale poorly or are not sufficiently accurate due to slow convergence [25]. Recently, the *GEORCE*-algorithm [25] has shown promising results in computing geodesics on general Riemannian manifolds with fast convergence and high accuracy.

In this paper, we show how to rewrite the distance-based notion of the Fréchet mean in Eq. 1 for both Riemannian and Finslerian manifolds to jointly minimize the Fréchet mean and geodesics. Using this, we introduce *GEORCE-FM* (*GEORCE for Fréchet Mean*) by extending *GEORCE* to simultaneously update the geodesics and the Fréchet mean in each iteration in a local chart, as illustrated in Fig. 1. Not only does this significantly reduce runtime; we also prove that the extended version converges to

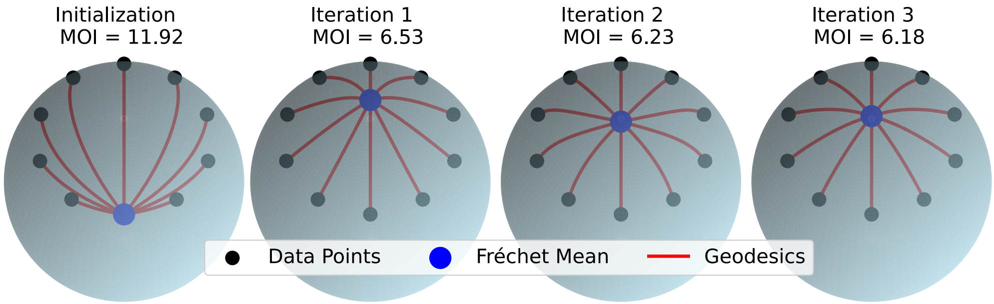

Fig. 1 Iterative computation for 10 data points on a circle within the convexity radius of the north pole, and hence guaranteeing the existence and uniqueness of the Fréchet mean of the data points around the north pole on a unit sphere, $\mathbb{S}^2$, using *GEORCE-FM*. After only 3 iterations the gradient of the discretized geodesic points and Fréchet mean is less than 10^{-4}. "MOI" refers to moment of inertia given by the sum of squared distances in Eq. 1.

a local minimum (global convergence) and local quadratic convergence similar to the original *GEORCE* algorithm. We further introduce an adaptive extension to estimate the Fréchet mean to make it scalable to a high number of data points. We prove that the adaptive extension converges in expectation to the Fréchet mean. Empirically, we show that *GEORCE-FM* outperforms baseline methods to compute the Fréchet mean in terms of accuracy and runtime for various manifolds and dimensions.

2 Background and related work

A Riemannian manifold is a differentiable manifold, $\mathcal{M}$, equipped with a Riemannian metric [19, Page 38]. The Riemannian metric, g, defines a smoothly varying inner product in the tangent bundle $T\mathcal{M}$. For each point $x \in \mathcal{M}$, the tangent space $T_x\mathcal{M}$ consists of tangent vectors to all smooth curves passing through $x \in \mathcal{M}$, and the tangent space itself is a vector space [19, Page 7 Definition 2.6]. The inner product induces the length of curves, $\gamma : [0,1] \to U$ connecting points.

$$\mathcal{L}_\gamma(a_i, a_j) = \int_0^1 \sqrt{\dot{\gamma}^\top(t) G(\gamma(t)) \dot{\gamma}(t)}\, \mathrm{d}t, \quad \gamma(0) = a_i, \gamma(1) = a_j,$$

where $U \subseteq \mathbb{R}^d$ denotes an open set in a local chart $\phi : U \to \mathcal{M}$. A geodesic is a curve, γ that minimizes the length between any points $x, y \in \mathcal{M}$ over all curves connecting x and y, thus giving rise to the distance function, $\mathrm{dist} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ [19, Page 146]. We assume that all manifolds are *geodesically complete* inside a strongly convex set containing the data points such that for any two datapoints $x, y \in \mathcal{M}$ there exists at least one connecting geodesic. Geodesics can be found by minimizing the energy functional in Eq. 2 or by solving a system of ODE [Lemma 2.3 in [19]]

$$\ddot{\gamma}^k(t) + \Gamma^k_{ij}\left(\gamma(t)\right) \dot{\gamma}^i(t) \dot{\gamma}^j(t) = 0, \quad \gamma(0) = x, \gamma(1) = y, \tag{3}$$

where $\{\Gamma^k_{ij}\}$ denote the Christoffel symbols and the multiplication of the upper and lower indices implies a sum following the Einstein convention. The *exponential map* is constructed explicitly with geodesics, $\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M}$ defined by $\mathrm{Exp}_x(v) = \gamma(1)$ with γ a geodesic such that $\gamma(0) = x$ and $\dot{\gamma}(0) = v$ [19, Page 70]. The *exponential map* is locally a diffeomorphism in a sufficiently small star-shaped neighborhood $U \subset T_x\mathcal{M}$ for $x \in \mathcal{X}$ and within this neighborhood, the *logarithmic map* is defined as $\mathrm{Log}_x : \mathcal{M} \to T_x\mathcal{M}$ as $\mathrm{Log}_x = \mathrm{Exp}_x^{-1}$ [19, Page 70].

A Finsler metric is a generalization of Riemannian manifolds, where the tangent spaces are equipped with a Finsler metric rather than an inner product. A Finsler metric is a Minkowski norm, $F : \mathcal{M} \times T\mathcal{M} \setminus \{0\} \to \mathbb{R}_+$, in the sense that it is a smooth function and satisfies the following requirements [26, Page 11]

- $F(x, cv) = cF(x, v)$ for all $v \in T_x\mathcal{M} \setminus \{0\}$ and $c > 0$ for $x \in \mathcal{M}$.
- The fundamental tensor, $G := \frac{\partial^2 F}{\partial v^i \partial v^j}(x, v)$, is symmetric and positive definite.

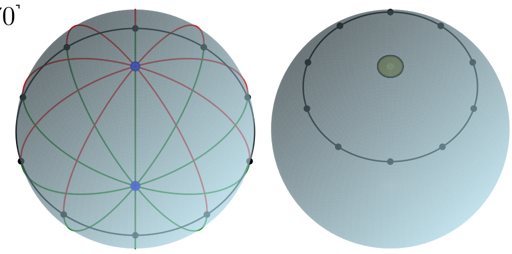

Fig. 2 The left figure shows a distribution of data points (black) on the equator, where both the north and south pole are Fréchet means (blue). The right figure shows a distribution of data points on a circle around the north pole, but where a closed set represented by the green area is removed from the unit sphere. In this case, the Fréchet mean does not exist.

Geodesics for a Finsler manifold are curves that minimize the integrated squared Finsler length over a set of curves γ,

$$\min_{\gamma} \frac{1}{2} \int_0^1 F^2\left(\gamma(t), \dot{\gamma}(t)\right) \mathrm{d}t. \tag{4}$$

which gives rise to the following ODE system generalizing Eq. 3

$$\ddot{\gamma}^k(t) + \eta_{ij}^k\left(\gamma(t), \dot{\gamma}(t)\right) \dot{\gamma}^i(t)\dot{\gamma}^j(t) = 0, \quad \gamma(0) = x, \gamma(1) = y, \tag{5}$$

where $\{\eta_{ij}^k\}$ generalizes the Riemannian Christoffel symbols to the Finslerian case and also depends on the velocity [26, Page 27].

The Fréchet Mean is defined in the continuous case as [11, 14]

$$\mu = \underset{y \in \mathcal{M}}{\arg\min}\, \mathbb{E}\left[\mathrm{dist}^2\left(y, X\right)\right] = \underset{y \in \mathcal{M}}{\arg\min} \int_{\mathcal{M}} \mathrm{dist}^2\left(y, z\right) p_X(z)\, \mathrm{d}\mathcal{M}(z), \tag{6}$$

for a random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathrm{d}\mathcal{M}(z)$ refers to the volume measure of the manifold $\mathcal{M}$ [14]. The discretized estimator of the Fréchet mean in Eq. 6 is given by Eq. 1. For symmetric manifolds, the Fréchet mean is a maximum likelihood estimator [12]. Despite the fact that the Fréchet mean naturally generalizes the Euclidean mean, it is not necessarily unique, nor does it exist, as illustrated in Fig. 2 for two simple data distributions on the unit sphere. If the data distribution for a complete Riemannian manifold is within a convex open ball of radius ρ around a point $m \in \mathcal{M}$, then the Fréchet mean is unique and exists [27, 28], while it has also been shown that the estimator of the Fréchet mean has strong consistency

[29–31]. In this paper, we will implicitly assume that there exists at least one Fréchet mean for any data distribution on a Riemannian manifold $\mathcal{M}$. The gradient of the sum of squared distances in Eq. 1 is the sum of logarithmic maps [14, 28]

$$\nabla_y \sum_{i=1}^{N} \mathrm{dist}^2(y, a_i) = \sum_{i=1}^{N} 2\mathrm{Log}(y, a_i). \tag{7}$$

If the logarithmic map is available in closed-form, the Fréchet mean can be computed using gradient descent [14, 18] or Karcher flows [28]. However, if the logarithmic map is not available, then using gradient descent will require estimating the logarithmic map numerically by either solving the ODE system in Eq. 3 or minimizing the energy in Eq. 2.

Computation of the Fréchet mean.

Different algorithms have been proposed to estimate the Fréchet mean such as Karcher-flows [28], which has been used for computing the Fréchet mean for the space of symmetric positive definite matrices [32], and Riemannian gradient descent [18]. However, both algorithms assume access to the logarithmic map, which requires computing geodesics. Different algorithms achieve better convergence than Karcher flows and Riemannian gradient descent, but are specifically designed for subclasses of Riemannian manifolds such as hyperbolic spaces [33] and n-dimensional spheres [34]. Without access to a closed-form solution to the logarithmic map, the computation of the Fréchet mean is based on geodesic computations. Various methods for computing geodesics have been proposed by solving an ODE system as a boundary value problem [35–38], as a fixed point problem [39, 40] or by directly minimizing discretized geodesics [25, 41, 42]. However, the Fréchet mean requires computing a potentially high number of geodesics in each iteration of the Fréchet mean making it intractable for a high number of data points. Although competing definitions of a mean value on a manifold have been proposed [43–45], the Fréchet mean remains the go-to standard for geometric computations on manifolds. Estimation of alternative definitions of the mean on a manifold is also computationally expensive, where, for example, the *diffusion mean* [43] involves Monte Carlo estimation of diffusion bridges [46] or gradient descent methods using neural networks [47].

3 GEORCE-FM

Rather than minimizing the sum of squared distance in Eq. 1, we will instead formulate the Fréchet mean as minimizing the sum of weighted energies in local coordinates.

$$\mu, \{\gamma_i\}_{i=1}^{N} = \operatorname*{arg\,min}_{\substack{y \in \mathcal{M} \\ \{\gamma_i\}_{i=1}^{N}}} \sum_{i=1}^{N} w_i \mathcal{E}_{\gamma_i}(a_i, y), \tag{8}$$

where $\mathcal{E}_\gamma(a, b)$ is the energy corresponding to Eq. 2 for a curve, γ, with starting point $a \in U$ and end point $b \in I$ for a local coordinate system $\phi : U \subseteq \mathbb{R}^d \to \mathcal{M}$, while

$\{w_i\}_{i=1}^N \subset \mathbb{R}_+$ are positive weights. We assume that $\{a_i\}_{i=1}^N \subset U$ are observed data in a local coordinate system within a geodesically convex set such that there exists at least one Fréchet mean. For the Fréchet mean in Eq. 1 the weights are one. We show in the following proposition that the minima of Eq. 8 are equivalent to the Fréchet means.

Lemma 1 *Let μ^{FM} denote the set of minima of Eq. 1, and let μ^{Energy} denote the set of minima of Eq. 8 with $w_i = 1$ for all $i \in \{1, \dots, N\}$. Then*

$$\mu^{\text{Energy}} = \mu^{\text{FM}}.$$

Proof See Appendix A.1. □

We can therefore compute geodesics and the Fréchet mean simultaneously by minimizing Eq. 8 rather than minimizing Eq. 1. We discretize the geodesics in Eq. 8 similarly to [41] in a local chart as

$$\begin{aligned} \min_{(x_{t,i},y)} \quad & \sum_{i=1}^{N} w_i \sum_{t=0}^{T-1} \left(x_{t,i+1} - x_{t,i}\right)^\top G(x_{t,i}) \left(x_{t,i+1} - x_{t,i}\right), \\ \text{s.t.} \quad & x_{0,i} = a_i, x_{T,i} = y, \quad \forall i \in \{1, \dots, N\}, \end{aligned} \tag{9}$$

where $\{x_{ti}\}_{t,i}$ corresponds to the discretized geodesics connecting the data points, $\{a_i\}_{i=1}^N$, and the candidate mean point, y. In this way, the estimation of the Fréchet mean corresponds to estimating N geodesics, where the Fréchet mean is the end point of the estimation.

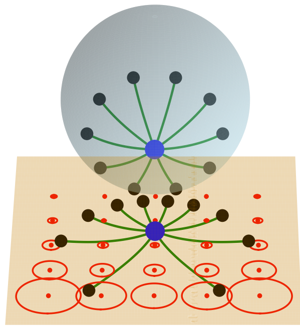

Fig. 3 We compute the geodesics and Fréchet mean simultaneously in a local chart of a Riemannian manifold. Green curves are the geodesics, black points are the datapoints, while red is the indicatrix field.

Similarly to [25], we formulate the above as a control problem in a local chart extending it to the computation of multiple geodesics, where all geodesics have the same non-fixed endpoint corresponding to the Fréchet mean

$$\begin{aligned} \min_{(x_{t,i},u_{t,i})} E(x) := \min_{(x_{t,i},u_{t,i})} & \left\{ \sum_{i=1}^{N} w_i \sum_{t=0}^{T-1} u_{t,i}^\top G(x_{t,i}) u_{t,i} \right\} \\ x_{t+1,i} = x_{t,i} + u_{t,i}, & \quad t = 0, \dots, T-1, \, i = 1, \dots, N, \\ x_{0,i} = a_i, x_{T,i} = y, & \quad i = 1, \dots, N. \end{aligned} \tag{10}$$

Unlike [25], we consider the end points of the geodesics as a variable to be estimated. We illustrate our approach in Fig. 3. The advantage of formulating the optimization problem in Eq. 9 as a control problem is that we are able to decompose the optimization problem in Eq. 10 into convex subproblems, resulting in the following necessary conditions for a minimum.

Proposition 2 *The necessary conditions for a minimum in Eq. 10 are*

$$\begin{aligned}
&2w_i G(x_{t,i})u_{t,i} + \mu_{t,i} = 0, \quad t = 0, \dots, T-1,\ i = 1, \dots, N, \\
&x_{t+1,i} = x_{t,i} + u_{t,i}, \quad t = 0, \dots, T-1,\ i = 1, \dots, N, \\
&\nabla_x \left[w_i u_{t,i}^\top G(x) u_{t,i} \right] \Big|_{x = x_{t,i}} + \mu_{t,i} = \mu_{t-1,i}, \quad t = 1, \dots, T-1,\ i = 1, \dots, N, \\
&0 = \sum_{i=1}^{N} \mu_{T-1,i}, \\
&x_{0,i} = a_i, x_{T,i} = y, \quad i = 1, \dots, N.
\end{aligned} \tag{11}$$

where $\mu_{t,i} \in \mathbb{R}^d$ *denotes the dual prices for the control problem for* $t = 0, \dots, T-1$ *and* $i = 1, \dots, N$.

Proof See Appendix A.4. □

If $u^\top G(x) u$ is convex in x, then the necessary conditions are also sufficient for a global minimum point, that is, there exists only one Fréchet mean, which is a global minimum point. If not, a local minimum point may be determined [28]. Note that compared to *GEORCE* [25], there is an additional condition related to the dual prices, $\{\mu_{t,i}\}_{t,i}$, since we consider the end point of the geodesics as a variable. The system of equations in Eq. 11 cannot be solved for general $G(\cdot)$. We aim to derive an iterative scheme to solve Eq. 11 and consider the variables $\{x_{t,i}^{(k)}\}_{t,i}$ and $\{u_{t,i}^{(k)}\}_{t,i}$ in iteration k. We apply a similar "trick" as [25] and fix the following variables.

$$\begin{aligned}
\nu_{t,i} &:= \nabla_x \left(w_i u_{t,i}^\top G(x) u_{t,i} \right) \Big|_{x = x_{t,i}^{(k)}, u_{t,i} = u_{t,i}^{(k)}}, \quad t = 1, \dots, T-1,\ i = 1, \dots, N, \\
G_{t,i} &:= G\left(x_{t,i}^{(k)} \right), \quad t = 0, \dots, T-1,\ i = 1, \dots, N.
\end{aligned}$$

Using this, Eq. 11 reduces to

$$\begin{aligned}
&\nu_{t,i} + \mu_{t,i} = \mu_{t-1,i}, \quad t = 1, \dots, T-1,\ i = 1, \dots, N, \\
&2w_i G_{t,i} u_{t,i} + \mu_{t,i} = 0, \quad t = 0, \dots, T-1,\ i = 1, \dots, N, \\
&\sum_{t=0}^{T-1} u_{t,i} = y - a_i, \quad i = 1, \dots, N, \\
&\sum_{i=1}^{N} \mu_{T-1,i} = 0.
\end{aligned} \tag{12}$$

The subproblems are tied together by the equation at time $T-1$ for the dual variables $\mu_{T-1,i}$, and the boundary conditions for the end point. In the following, this is circumvented by deriving a closed-form formula for y that allows a decomposition into N geodesic problems.

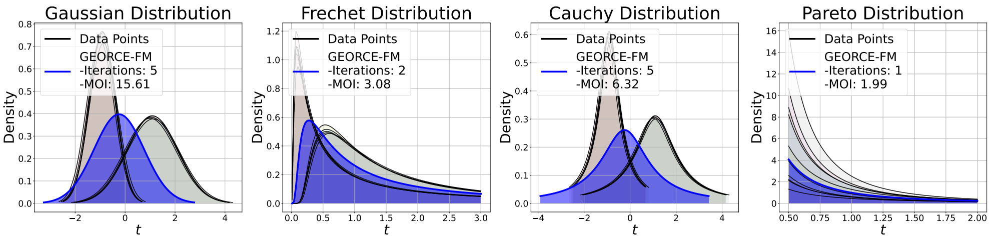

Fig. 4 The application of *GEORCE-FM* to Information geometry for 4 different distributions equipped with the Fisher-Rao metric [48] with synthetic data. From left to right we have: Gaussian distribution, Fréchet distribution, Cauchy distribution and Pareto distribution. The black outlined distributions are the data, where the blue outlined distribution is the estimated Fréchet mean using *GEORCE-FM* and "MOI" refers to moment of inertia given by the sum of squared distances in Eq. 1. The details of the manifolds and data points can be found in Appendix C.

Proposition 3 *The solution to Eq. 12 for u_t, μ_t and x_t is*

$$\begin{aligned}
&y = W^{-1}V, \\
&\mu_{T-1,i} = \left(\sum_{t=0}^{T-1} G_{t,i}^{-1}\right)^{-1} \left(2w_i(a_i - y) - \sum_{t=0}^{T-1} G_{t,i}^{-1} \sum_{t>j}^{T-1} \nu_{j,i}\right), \quad i = 1, \dots, N, \\
&u_{t,i} = -\frac{1}{2w_i} G_{t,i}^{-1} \left(\mu_{T-1,i} + \sum_{j>t}^{T-1} \nu_{j,i}\right), \quad t = 0, \dots, T-1,\, i = 1, \dots, N \\
&x_{t+1,i} = x_{t,i} + u_{t,i}, \quad t = 0, \dots, T-2,\, i = 1, \dots, N, \\
&x_{0,i} = a_i \quad i = 1, \dots, N,
\end{aligned} \tag{13}$$

where

$$\begin{aligned}
W &= \sum_{i=1}^{N} w_i \left(\sum_{t=0}^{T-1} G_{t,i}^{-1}\right)^{-1}, \\
V &= \sum_{i=1}^{N} w_i \left(\sum_{t=0}^{T-1} G_{t,i}^{-1}\right)^{-1} a_i - \frac{1}{2} \sum_{i=1}^{N} \left(\sum_{t=0}^{T-1} G_{t,i}^{-1}\right)^{-1} \sum_{t=0}^{T-1} G_{t,i}^{-1} \sum_{j>t}^{T-1} \nu_{j,i}.
\end{aligned} \tag{14}$$

Proof See Appendix A.5. □

Using the update scheme in Proposition 3, we show *GEORCE-FM* in pseudocode in Algorithm 1, where we apply line-search for the sum of discretized energy using backtracking with Armijo condition [49] with decay rate $\rho = 0.5$ [25]. We stop *GEORCE-FM* when the average gradient norm is below a threshold. Note that in Algorithm 1, then $x_{T,i} = y$ for all $i = 1, \dots, N$ in each iteration. Furthermore, when y has been determined, the remaining update formulas can be run in parallel for $i = 1, \dots, N$. Note that similar to *GEORCE* [25], all computations are within a local chart, and it is implicitly assumed that the initial curve and updates are defined within this chart. This can be ensured using line-search, or by changing the chart if the computations are close to the boundary of the domain. For a detailed discussion on this, we refer to [25]. For the examples in this paper, we will assume that the computations are well-defined within the chart, and we will therefore not change the chart doing computations, although this can easily be added to the algorithm.

Similarly to *GEORCE* [25], the algorithm, as will be proven below, will converge to a local minimum. To check whether the found solution is actually a local minimum or a global minimum, the algorithm can be run for different initialization curves and compare the corresponding sum of squared lengths in Eq. 9.

We estimate in Fig. 4 the Fréchet mean using *GEORCE-FM* for four different examples of distributions in information geometry: normal distribution, Fréchet distribution, Cauchy distribution and Pareto distribution for $T = 100$ grid points, which are equipped with the Fisher-Rao metric [48]. We see that *GEORCE-FM* converges after only a few iterations.

Algorithm 1 GEORCE-FM for Riemannian Manifolds

Input: tol, $a_{1:N}$, T

Output: Geodesic estimate $x_{0:T}$ Set $y^{(0)} \leftarrow a_0$, $x_{t,i}^{(0)} \leftarrow a_i + \frac{y^{(0)}-a_i}{T}t$ and $u_{t,i}^{(0)} \leftarrow \frac{y^{(0)}-a_i}{T}$ for $t = 0., \dots, T$ and $i = 1, \dots, N$.

while do $\frac{1}{N}\left\| \nabla_y E(y)\big|_{y=x_{t,i}^{(k)}} \right\|_2 > \mathrm{tol}$

  $G_{t,i} \leftarrow G\left(x_{t,i}^{(k)}\right)$ for $t = 0, \dots, T-1$ and $i = 1, \dots, N$.

  $\nu_{t,i} \leftarrow \nabla_x \left(u_{t,i}^{(k)} G(x)\, u_{t,i}^{(k)}\right)\Big|_{x=x_{t,i}^{(k)}}$ for $t = 1, \dots, T-1$ and $i = 1, \dots, N$.

  $y \leftarrow W^{-1}V$ with W, V given by Eq. 14.

  $\mu_{T-1,i} \leftarrow \left(\sum_{t=0}^{T-1} G_{t,i}^{-1}\right)^{-1}\left(2w_i(a_i - y) - \sum_{t=0}^{T-1} G_{t,i}^{-1} \sum_{t>j}^{T-1} \nu_{j,i}\right)$ for $i = 1, \dots, N$ and $t = 1, \dots, T-1$.

  $u_{t,i} \leftarrow -\frac{1}{2w_i} G_{t,i}^{-1}\left(\mu_{T-1,i} + \sum_{j>t}^{T-1} \nu_{j,i}\right)$ for $t = 0, \dots, T-1$ and $i = 1, \dots, N$.

  $x_{t+1,i} \leftarrow x_{t,i} + u_{t,i}$ for $t = 0, \dots, T-2$ and $i = 1, \dots, N$.

  Using line search find α^* for the following optimization problem for the discrete sum of energy E

$$
\begin{aligned}
\min_{\alpha} \quad & E\left(x_{0:T,1:N}\right) \quad \text{(exact line search)} \\
\text{s.t.} \quad & x_{t+1,i} = x_{t,i} + \alpha \tilde{u}_{t,i} + (1-\alpha) u_{t,i}^{(k)}, \quad t = 0, \dots, T-1,\ i = 1, \dots, N. \\
& \tilde{u}_{t,i} = \alpha u_{t,i} + (1-\alpha) u_{t,i}^{(k)}, \quad t = 0, \dots, T-1,\ i = 1, \dots, N. \\
& x_{0,i} = a_i.
\end{aligned}
$$

  Set $u_{t,i}^{(k+1)} \leftarrow \alpha^* u_{t,i} + (1-\alpha^*) u_{t,i}^{(k)}$ for $t = 0, \dots, T-1$ and $i = 1, \dots, N$.

  Set $x_{t+1,i}^{(k+1)} \leftarrow x_{t,i}^{(k+1)} + u_{t,i}^{(k+1)}$ for $t = 0, \dots, T-1$ and $i = 1, \dots, N$.

end while

return $x_{t,i}$ for $t = 0, \dots, T-1$ for $i = 1, \dots, N$.

Complexity.

GEORCE-FM in Algorithm 1 requires matrix inversions for N geodesics and therefore scales $\mathcal{O}\left(NTd^3\right)$ in each iteration, where d is the manifold dimension. The advantage of *GEORCE-FM* is that we can simultaneously compute geodesics and Fréchet mean rather than solving the geodesic subproblem in each iteration of the Fréchet mean.

Convergence results.

We prove that *GEORCE-FM* has global convergence and local quadratic convergence by generalizing the original proofs in the *GEORCE*-algorithm [25].

Proposition 4 (Convergence Properties) *Let $E^{(k)}$ be the value of the sum of discretized energy functional for the solution after iteration k (with line search) in GEORCE-FM. The GEORCE-FM algorithm has the following properties*

- *Global convergence: If the starting point $\left(x^{(0)}, u^{(0)}\right)$ is feasible in the sense that the geodesic candidates have the correct start and share the same end point, then the series $\left\{E^{(k)}\right\}_{i>0}$ will converge to a (local) minimum.*
- *Local quadratic convergence: Under sufficient regularity of the energy functional (see Appendix A.3) it holds that if the energy functional has a strongly unique (local) minimum point $z^* = (x^*, u^*, y^*)$ and locally the step-size $\alpha^* = 1$, i.e. no line search, then GEORCE-FM has locally quadratic convergence, i.e.*

$$\exists \epsilon > 0: \quad \exists c > 0: \quad \forall z^{(k)} \in B_\epsilon(z^*): \quad \left\| z^{(k+1)} - z^*) \right\| \leq c \left\| z^{(k)} - z^* \right\|^2,$$

  where $z^{(k)} = (x^{(k)}, u^{(k)}, y^{(k)})$ is the solution from GEORCE-FM in iteration k, and $B_\epsilon\left(z^*\right)$ is the ball with radius $\epsilon > 0$ and center z^*.

Proof See Appendix A.6 for the proof of global convergence and Appendix A.7 for the proof of local convergence. □

With Proposition 4 we see that not only does *GEORCE-FM* allow simultaneous optimization of the Fréchet mean geodesics, but it also has global convergence similar to gradient descent methods and local quadratic convergence similar to the Newton method.

4 Adaptive GEORCE-FM

In the previous section, we have seen that *GEORCE-FM* can compute geodesics simultaneously with the Fréchet mean. However, *GEORCE-FM* scales by $\mathcal{O}\left(NTd^3\right)$, making it computationally expensive to use for a large number of datapoints. In this section, we introduce an adaptive extension of *GEORCE-FM* by considering a stochastic estimator of the minimization problem in Eq. 8 with mini-batches applying subsets of data points.

$$\min_{y \in \mathcal{M}} \sum_{i \in \mathcal{I}} w_i \mathcal{E}\left(a_i, y\right), \tag{15}$$

where $\mathcal{I} = \{i_1, \dots, i_n\}$ is an index set with n distinct random integer variables of the form $1 \leq i_1 < i_2 < \dots i_n \leq N$. The idea is to adaptively update the estimator of the Fréchet mean using only the stochastic estimates from random mini-batches of data. Inspired by the update scheme in Proposition 3 and Eq. 14, we define the following stochastic values using only mini-batches of the data,

$$\begin{aligned} \tilde{W} &= \sum_{i \in \mathcal{I}} w_i \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1}, \\ \tilde{V} &= \sum_{i \in \mathcal{I}} w_i \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1} a_i - \frac{1}{2} \sum_{i \in \mathcal{I}} \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1} \sum_{t=0}^{T-1} G_{t,i}^{-1} \sum_{j>t}^{T-1} g_{j,i} \end{aligned} \tag{16}$$

such that the stochastic estimate, $\tilde{y}$, of the Fréchet mean in the *GEORCE-FM* algorithm is

$$\tilde{y} = \tilde{W}^{-1} \tilde{V}, \tag{17}$$

We propose to adaptively estimate $\tilde{W}$ and $\tilde{V}$ for different mini-batches of data and then update the Fréchet mean by Eq. 17. We show the algorithm in pseudo-code in Algorithm 2. Note that to have a proper estimate of V and W in Eq. 16, we run *GEORCE-FM* for a fixed number of iterations on the mini-batch of the dataset.

Algorithm 2 Adaptive GEORCE-FM for Riemannian Manifolds

Input: tol, $a_{1:N,i}$, T, sub_iters.
Output: Geodesic estimate $x_{0:T}$.
Compute random index set, $\mathcal{I}$.
Compute sub_iters iterations using $\left(\hat{W}^{(0)}, \hat{V}^{(0)}\right) \leftarrow$ *GEORCE_FM*$(\{a_i\}_{i\in\mathcal{I}})$ using algorithm 1 and Eq. 16.
$k \leftarrow 1$.
while do $\left\|y^{(k)} - y^{(k-1)}\right\|_2 > \text{tol}$
    Compute random index set, $\mathcal{I}$.
    Compute sub_iters iterations using $\left(\tilde{W}, \tilde{V}\right) \leftarrow$ *GEORCE_FM*$(\{a_i\}_{i\in\mathcal{I}})$ using algorithm 1 and Eq. 16.
    if convergence **then**
        $\alpha^* \leftarrow \frac{1}{k+1}$
    else
        $\alpha^* \leftarrow \lambda$
    end if
    $\hat{W}^{(k)} \leftarrow \alpha^* \tilde{W} + (1-\alpha^*)\hat{W}^{(k-1)}$.
    $\hat{V}^{(k)} \leftarrow \alpha^* \tilde{V} + (1-\alpha^*)\hat{V}^{(k-1)}$.
    $y^{(k)} \leftarrow \left(\hat{W}^{(k)}\right)^{-1} \hat{V}^{(k)}$.
    $k \leftarrow k+1$.
end while
return $y^{(k)}$.

Convergence results.

We prove in the following that the adaptive extension of *GEORCE-FM* converges in expectation assuming sufficient regularity of the index set size and adaptive scheme.

Proposition 5 (Adaptive Convergence) *Assume sufficient regularity of the index set and adaptive scheme (see Appendix A.8). Then the adaptive version of GEORCE-FM will converge to a local minimum point in expectation.*

Proof See Appendix A.8 for the proof and assumptions. □

In Proposition 18 we only assume sufficient regularity of the adaptive update scheme, and therefore the current update scheme in Algorithm 2 can easily be replaced by other adaptive updating schemes. Note that the "momentum" type of estimation for V and W is not assumed in the convergence proof, but is added in order to increase convergence speed similar to e.g. *ADAM* [20].

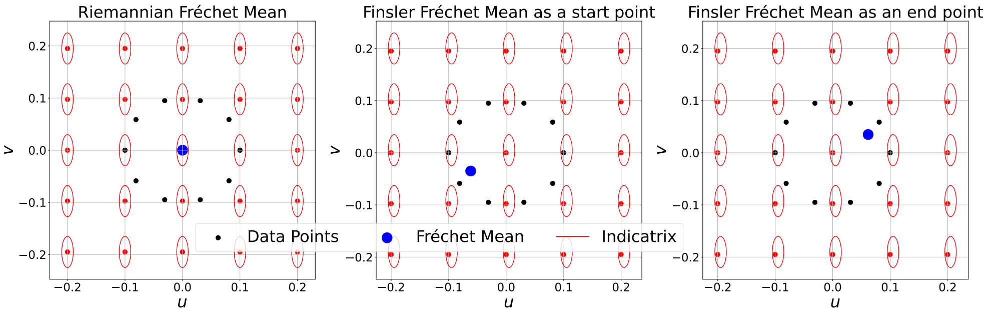

Fig. 5 We consider a constant indicatrix field of centered ellipsis in the left-most figure, and the corresponding Riemannian Fréchet mean. The center figures shows displaced ellipses such that they are non-centered and the corresponding Fréchet mean as a starting point. The right-most figure shows the Finslerian Fréchet mean as an end point.

5 Extension to Finsler manifolds

In this section, we generalize the previous results to the Finslerian manifolds. Let dist_F denote the distance function on a Finsler manifold and define the Fréchet mean similar to the Riemannian case as

$$\mu = \arg\min_{y \in \mathcal{M}} \sum_{i=1}^{N} \mathrm{dist}_F^2(y, a_i). \tag{18}$$

In Eq. 19 we define the Fréchet mean as the start point of the distance. Since the Finsler metric is not symmetric, this is not equivalent to having the Fréchet mean as an end point as illustrated in Fig. 5. Note that the Fréchet mean as a starting point or end point is also denoted the *forward p-mean* and *backward p-mean* [50]. Analogously to the Riemannian case we prove in Appendix A.2 that we can re-write the optimization problem in Eq. 18 to jointly minimize the energy with respect to the energy and candiate Fréchet mean, that is,

$$\mu^{\text{Energy}}, \{\gamma_i\}_{i=1}^N = \arg\min_{\substack{y \in \mathcal{M} \\ \{\gamma_i\}_{i=1}^N}} \sum_{i=1}^{N} \mathcal{E}_{\gamma_i}^{(F)}(y, a_i), \tag{19}$$

where $\mathcal{E}_{\gamma}^{(F)}(y, x) = \int_0^1 g\left(\gamma(t), \dot{\gamma}(t)\right) \mathrm{d}t$ denotes the energy of the curve γ with $\gamma(0) = y$ and $\gamma(1) = x$. In a local coordinate system, the fundamental tensor can be written as $g\left(\gamma(t), \dot{\gamma}(t)\right) = \dot{\gamma}(t)^\top G\left(\gamma(t), \dot{\gamma}(t)\right) \dot{\gamma}(t)$, and therefore the discretized version Eq. 19 becomes

$$\begin{aligned}
&\min_{(x_{t,i}, u_{t,i})} \left\{ \sum_{i=1}^{N} w_i \sum_{t=0}^{T-1} u_{t,i}^\top G(x_{t,i}, u_{t,i}) u_{t,i} \right\} \\
&x_{t+1,i} = x_{t,i} + u_{t,i}, \quad t = 0, \dots, T-1, \; i = 1, \dots, N, \\
&x_{0,i} = y, \; x_{T,i} = a_i, \quad i = 1, \dots, N,
\end{aligned} \tag{20}$$

The only difference compared to the Riemannian case is that G now also depends on the velocity and position. However, we cannot directly apply the same approach as in the Riemannian case. In the Riemannian case we had implicitly changed the order of start and end point such that the Fréchet mean was the end point of the connecting geodesics. Unlike Riemannian manifolds, the distance is not symmetric on Finsler manifolds and we can therefore not apply a similar "trick" in the Finslerian case. To circumvent this, consider the following transformations.

$$\begin{aligned}
u_{t,i} &= -\tilde{u}_{T-t-1,i}, \quad t = 0, \dots, T-1, \, i = 1, \dots, N, \\
x_{t,i} &= \tilde{x}_{T-t,i}, \quad t = 0, \dots, T-1, \, i = 1, \dots, N.
\end{aligned}$$

With the change of variable, Eq. 20 can be formulated as

$$\begin{aligned}
&\min_{(\tilde{x}_{t,i}, \tilde{u}_{t,i})} \left\{ \sum_{i=1}^{N} w_i \sum_{t=0}^{T-1} \tilde{u}_{T-t-1,i}^{\top} G\left(\tilde{x}_{T-t,i}, -\tilde{u}_{T-t-1,i}\right) \tilde{u}_{T-t-1,i} \right\} \\
&\tilde{x}_{T-t-1,i} = \tilde{x}_{T-t,i} - \tilde{u}_{T-t-1,i}, \quad t = 0, \dots, T-1, \, i = 1, \dots, N, \\
&\tilde{x}_{0,i} = a_i, \, \tilde{x}_{T,i} = y, \quad i = 1, \dots, N.
\end{aligned}$$

Consider the change of index by $s := T - 1 - t$ for $t = 0, \dots, T-1$. When T is sufficiently large, $G\left(\tilde{x}_{T-t,i}, \tilde{u}_{T-1-1,i}\right)$ can be approximated by $G\left(\tilde{x}_{T-t-1,i}, \tilde{u}_{T-1-1,i}\right)$. With this modification, we get the following.

$$\begin{aligned}
\min_{(\tilde{x}_{s,i}, \tilde{u}_{s,i})} E^{(F)}(x) := \min_{(x_{t,i}, u_{t,i})} &\left\{ \sum_{i=1}^{N} w_i \sum_{s=0}^{T-1} \tilde{u}_{s,i}^{\top} \tilde{G}\left(\tilde{x}_{s,i}, \tilde{u}_{s,i}\right) \tilde{u}_{s,i} \right\} \\
\tilde{x}_{s+1,i} &= \tilde{x}_{s,i} + \tilde{u}_{s,i}, \quad s = 0, \dots, T-1, \, i = 1, \dots, N, \\
\tilde{x}_{0,i} &= a_i, \, \tilde{x}_{T,i} = y, \quad i = 1, \dots, N,
\end{aligned} \tag{21}$$

where $\tilde{G}(x, u) := G(x, -u)$. In Eq. 21 the Fréchet mean can be treated as the end point. By following a completely similar approach as in the Riemannian case with some small modifications, we get the following iterative scheme to estimate the geodesics and Fréchet mean for Finsler manifolds.

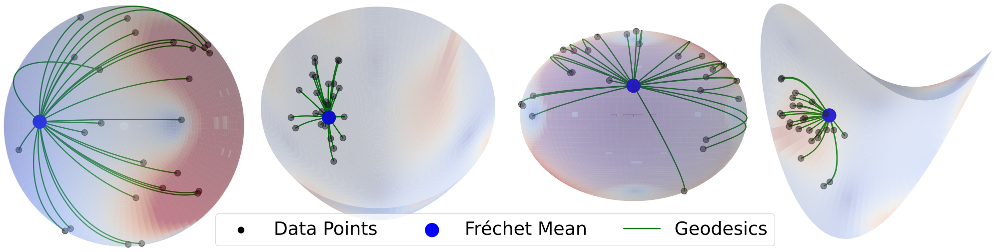

Fig. 6 The application of *GEORCE-FM* to Finsler manifolds, where we show the estimated Fréchet mean and geodesics using *GEORCE-FM* for (from left to right) a sphere, paraboloid, ellipsoid and hyperbolic paraboloid all equipped with a wind field. The details of the manifolds and data points can be found in Appendix C.

Proposition 6 *The update scheme for u_t, μ_t and x_t is*

$$
\begin{aligned}
&y = W^{-1}V, \\
&\mu_{T-1,i} = \left(\sum_{t=0}^{T-1} \tilde{G}_{t,i}^{-1}\right)^{-1} \left(2w_i(a_i - y) - \sum_{t=0}^{T-1} \tilde{G}_{t,i}^{-1}\left(\zeta_{t,i} + \sum_{t>j}^{T-1} \nu_{j,i}\right)\right), \quad i = 1, \dots, N, \\
&u_{t,i} = -\frac{1}{2w_i}\tilde{G}_{t,i}^{-1}\left(\mu_{T-1,i} + \zeta_{t,i} + \sum_{j>t}^{T-1} \nu_{j,i}\right), \quad t = 0, \dots, T-1,\ i = 1, \dots, N \\
&x_{t+1,i} = x_{t,i} + u_{t,i}, \quad t = 0, \dots, T-2,\ i = 1, \dots, N, \\
&x_{0,i} = a_i \quad i = 1, \dots, N,
\end{aligned}
\tag{22}
$$

where

$$
\begin{aligned}
W &= \sum_{i=1}^{N} w_i \left(\sum_{t=0}^{T-1} \tilde{G}_{t,i}^{-1}\right)^{-1}, \\
V &= \sum_{i=1}^{N} w_i \left(\sum_{t=0}^{T-1} \tilde{G}_{ti}^{-1}\right)^{-1} a_i - \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{t=0}^{T-1} \tilde{G}_{t,i}^{-1}\right)^{-1} \sum_{t=0}^{T-1} \tilde{G}_{ti}{}^{-1}\left(\zeta_{t,i} + \sum_{j>t}^{T-1} \nu_{j,i}\right).
\end{aligned}
$$

Here $\nu_{t,i} := \nabla_y u_{t,i}^\top \tilde{G}(y, u_{t,i}) u_{t,i}\Big|_{y=x_{t,i}}$ *and* $\zeta_{t,i} := \nabla_v u_{t,i}^\top \tilde{G}(x_{t,i}, v) u_{t,i}\Big|_{v=u_{t,i}}$.

Proof See Appendix A.9. □

In Fig. 6 we show the estimated Fréchet mean as well as the estimated connecting geodesics for four different Finsler manifolds. We also see that the adaptive extension of *GEORCE-FM* in algorithm 2 can be directly extended to Finsler manifolds, where

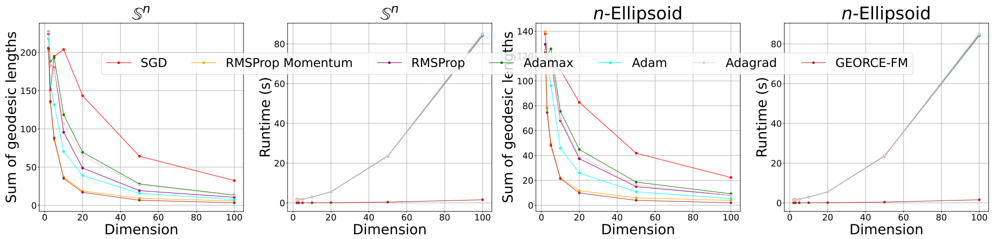

Fig. 7 Estimated squared geodesic length and runtime for the sphere and ellipsoid for different dimensions and different optimization schemes.

$\tilde{V}$ and $\tilde{W}$ are given by

$$\begin{aligned}
\tilde{W} &= \sum_{i \in \mathcal{I}} w_i \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1}, \\
\tilde{V} &= \sum_{i \in \mathcal{I}} w_i \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1} a_i - \frac{1}{2} \sum_{i \in \mathcal{I}} \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1} \sum_{t=0}^{T-1} G_{t,i}^{-1} \left(\zeta_{t,i} + \sum_{j>t}^{T-1} g_{j,i} \right)
\end{aligned} \tag{23}$$

In the above, we have shown an iterative scheme to estimate the Fréchet mean as a starting point on a Finsler manifold. The Fréchet mean can also trivially be found as an end point by not applying the transformation of the variables x and u.

6 Experiments

Riemannian.

To compare *GEORCE-FM* to alternative similar methods, we use different optimizers for minimizing Eq. 9 directly. Since multiple geodesics have to be computed, we use methods that are designed to handle large number of parameters. [25] observed that methods such as *BFGS* and trust-region methods were slow, and therefore we will instead apply stochastic methods that are designed to handle a large number of parameters. Table 1 shows the runtime estimates for *ADAM* [20], *RMSprop Momentum* [51] and *GEORCE-FM*. To estimate the solution, we use *GEORCE* [25] with 10 iterations to estimate the geodesics between the data points and the estimated Fréchet mean to compute the sum of squared geodesics distances in Eq. 1. We summarize the results for the n-sphere and n-ellipsoids in Fig. 7. We see that *GEORCE-FM* is significantly faster than the alternative methods and generally achieves a better solution. For details on the manifolds and data, we refer to Appendix C and Appendix D, respectively. We also provide additional experiments in Appendix E.2 for other methods.

To further show the application of *GEORCE-FM* to real-world data, we consider a manifold learned using a Variational-Autoencoder (VAE) equipped with the pull-back metric [41, 42]

$$G(z) = J_{f_\theta}(z)^\top J_{f_\theta}(z),$$

where J_{f_θ} is the Jacobian of the decoder, f_θ, which is a neural network with parameters θ. We provide additional details on the VAE in Appendix C. In Fig. 8 we show

	ADAM (T=100)		RMSprop Momentum (T=100)		GEORCE-FM (T=100)	
Finsler Manifold	Length	Runtime	Length	Runtime	Length	Runtime
$\mathbb{S}^2$	217.60	1.6748 ± 0.0010	**202.28**	1.6695 ± 0.0013	204.99	$\mathbf{0.0611 \pm 0.0017}$
$\mathbb{S}^3$	158.24	1.7086 ± 0.0007	**133.92**	1.7062 ± 0.0012	135.98	$\mathbf{0.0534 \pm 0.0002}$
$\mathbb{S}^5$	131.77	1.8670 ± 0.0009	**85.48**	1.8372 ± 0.0016	88.01	$\mathbf{0.0537 \pm 0.0020}$
$\mathbb{S}^{10}$	70.37	2.9435 ± 0.0015	36.79	2.9340 ± 0.0011	**35.20**	$\mathbf{0.0808 \pm 0.0002}$
$\mathbb{S}^{20}$	39.04	5.5679 ± 0.0009	18.83	5.5143 ± 0.0010	**16.97**	$\mathbf{0.1250 \pm 0.0002}$
$\mathbb{S}^{50}$	15.70	25.1048 ± 0.6575	9.14	24.1128 ± 0.3806	**6.67**	$\mathbf{0.3963 \pm 0.0002}$
$\mathbb{S}^{100}$	7.64	85.3785 ± 0.4844	5.80	84.8019 ± 0.0659	**3.27**	$\mathbf{1.5691 \pm 0.0003}$
E (2)	123.22	1.6598 ± 0.0020	139.62	1.6661 ± 0.0008	**122.68**	$\mathbf{0.0357 \pm 0.0002}$
E (3)	117.67	1.7159 ± 0.0008	78.37	1.7002 ± 0.0008	**74.58**	$\mathbf{0.0653 \pm 0.0022}$
E (5)	96.16	1.8809 ± 0.0010	49.45	1.8445 ± 0.0017	**48.14**	$\mathbf{0.0529 \pm 0.0017}$
E (10)	45.91	2.9558 ± 0.0017	22.22	2.9349 ± 0.0012	**21.42**	$\mathbf{0.0715 \pm 0.0002}$
E (20)	25.82	5.6296 ± 0.0029	11.68	5.5802 ± 0.0024	**9.89**	$\mathbf{0.1300 \pm 0.0002}$
E (50)	10.82	23.7450 ± 0.0019	5.91	23.6992 ± 0.0069	**3.94**	$\mathbf{0.3960 \pm 0.0001}$
E (100)	5.39	84.4791 ± 0.1039	3.90	84.3799 ± 0.0973	**1.94**	$\mathbf{1.5685 \pm 0.0002}$
$\mathbb{T}^2$	1234.12	1.6483 ± 0.0011	1227.92	1.6429 ± 0.0015	**1225.65**	$\mathbf{0.0743 \pm 0.0028}$
$\mathbb{H}^2$	**174.60**	1.6616 ± 0.0017	174.61	1.6564 ± 0.0025	174.60	$\mathbf{0.1594 \pm 0.0015}$
Paraboloid	**103.95**	1.6080 ± 0.0012	123.29	1.6086 ± 0.0013	103.95	$\mathbf{0.0182 \pm 0.0002}$
Hyperbolic Paraboloid	**114.39**	1.6085 ± 0.0012	129.02	1.6069 ± 0.0012	114.40	$\mathbf{0.0082 \pm 0.0001}$
Gaussian Distribution	148.97	0.2134 ± 0.0028	154.36	0.2120 ± 0.0126	**148.52**	$\mathbf{0.0154 \pm 0.0017}$
Fréchet Distribution	31.22	0.2060 ± 0.0123	36.74	0.2116 ± 0.0136	**31.15**	$\mathbf{0.0065 \pm 0.0002}$
Cauchy Distribution	60.60	0.2388 ± 0.0003	$-/-$	$-$	**60.27**	$\mathbf{0.0138 \pm 0.0002}$
Pareto Distribution	24.03	0.2054 ± 0.0083	1690.17	0.2002 ± 0.0009	**24.02**	$\mathbf{0.0062 \pm 0.0002}$
VAE MNIST	858.06	39.0780 ± 0.0026	813.19	39.2018 ± 0.0039	**771.59**	$\mathbf{7.1326 \pm 0.0027}$
VAE CelebA	9420.09	832.8971 ± 0.0761	43617996.00	862.4910 ± 0.2627	**9414.38**	$\mathbf{79.2793 \pm 0.0061}$

Table 1 The table shows the sum of squared geodesic length for the estimated Fréchet mean for *ADAM*, *RMSprop Momentum* and *GEORCE-FM* on a GPU for Riemannian manifolds. The methods were terminated if the ℓ^2-norm of the average gradient across the data was less than 10^{-4} or $1,000$ iterations have been reached. $\mathrm{E}(n)$ denotes an Ellipsoid of dimension n, while $\mathcal{P}(n)$ denotes the space of $n \times n$ symmetric positive definite matrices. When the computational time was longer than 24 hours, or if the method returns nan, the value is set to $-$. Since the VAE's are learned manifolds based on data, and therefore can be unstable, the methods are terminated if the average gradient across the data was less than 10^{-3}. Further, to avoid memory complexity for the learned manifolds in storing the neural network, the geodesics are updated sequentially for VAE MNIST and VAE CelebA, while for the other manifolds the geodesics are updated in parallel. The VAE's use 10 datapoints, while the remaining manifolds use 100 synthetically generated datapoints. The data is described in Appendix D.

Fig. 8 The application of *GEORCE-FM* to manifold learning for a VAE estimating the Fréchet mean for 10 images of the CelebA dataset [52] reconstructed using a VAE similar to [41] Each row shows 10 points on the geodesic, while the left most image is the data and the right most image is the estimated Fréchet mean using *GEORCE-FM*. The length of the estimated geodesic using *GEORCE-FM* is shown to the left, while the grid point number is shown at the bottom.

the estimated geodesics and Fréchet mean using $T = 100$ grid points for 10 images reconstructed using the VAE corresponding to the data used in Table 1.

Finslerian.

To illustrate our method across different Finsler manifolds, we consider a Riemannian background metric, which is affected by a wind field similar to the construction in [53]. We consider the same generic wind field as in [25]

$$f(x) = \frac{\sin x \odot \cos x}{(\cos x)^\top G(x) \cos x}, \tag{24}$$

We provide additional details on the construction in Appendix C. Table 2 shows the runtime and sum of squared geodesic length for different Finsler manifolds constructed using the wind field in Eq. 24 on a Riemannian background metric. The sum of squared geodesic lengths is computed using *GEORCE* [25] between the data points and the estimated Fréchet mean for the different methods. We see that *GEORCE-FM* generally is faster and achieves a better estimate of the Fréchet mean. Additional experiments can be found in Appendix E.2.

Adaptive Estimation

To compare the adaptive estimation of the Fréchet mean with alternative methods, we consider alternative stochastic methods estimating the Fréchet using mini-batches of data as in Eq. 15. For alternative methods, we estimate stochastically the Fréchet mean by minimizing

$$\min_{(x_{t,i},y)} \sum_{i\in\mathcal{I}} w_i \sum_{t=0}^{T-1} \left(x_{t,i+1} - x_{t,i}\right)^\top G(x_{t,i}) \left(x_{t,i+1} - x_{t,i}\right), \quad x_{0,i} = a_i, x_{T,i} = y, \tag{25}$$

where $\mathcal{I} = \{i_1, i_2, \dots, i_n\}$ is an index set consisting of n distinct random integers $i_1 < i_2 < \cdots < i_n$. Thus, for standard stochastic optimization solvers, we estimate only mini-batches of the geodesics as well as the Fréchet mean in each iteration.

We show the result in Table 3, where we consider datasets of $1,000$ data points and use 10% of the data in the estimation. For the adaptive estimation of *GEORCE-FM* we use 5 sub-iterations and terminate the algorithm if the change in the estimated Fréchet mean is less than 10^{-4}. For the other methods, we terminate the algorithm if the mean stochastic gradient of Eq. 25 over the number of data points is less than 10^{-4}. Since it is difficult to have similar stopping criteria for *GEORCE-FM* and the alternative methods, we do not report the runtime in Table 3, but only the sum of the squared geodesic length for all data using *GEORCE* [25] for the estimated Fréchet mean for each algorithm. In general, we see that the adaptive extension of *GEORCE-FM* is more accurate than the alternative methods.

Application to geometric statistics

In this section, we illustrate how geometric statistics can be easily computed using *GEORCE-FM*. To estimate the Fréchet mean, we compute both the Fréchet mean and geodesics as well as computing $u_{0,i}$ for each data point a_i for $i = 1, \dots, N$, which can be interpreted as the logarithmic map modulo scaling. These three "ingredients" often provide the basis for more elaborate statistics. To illustrate this, consider principal geodesic analysis [13], where data is projected onto geodesics that describe the most of the variance in the data. In this case, the principal geodesics are computed as the eigenvectors of

$$S = \frac{1}{N} \sum_{i=1}^{N} \mathrm{Log}_\mu(a_i) \mathrm{Log}_\mu(a_i)^\top,$$

where μ is the Fréchet mean for the data $\{a_i\}_{i=1}^N$. By computing the Fréchet mean using *GEORCE-FM*, the eigenvectors of S can be directly computed approximating

	ADAM (T=100)		**RMSprop Momentum (T=100)**		**GEORCE-FM (T=100)**	
Finsler Manifold	Length	Runtime	Length	Runtime	Length	Runtime
$\mathbb{S}^2$	83.64	2.3388 ± 0.0020	85.91	2.3085 ± 0.0029	**83.63**	$\mathbf{0.0230 \pm 0.0002}$
$\mathbb{S}^3$	76.33	2.4284 ± 0.0013	75.97	2.4215 ± 0.0016	**75.16**	$\mathbf{0.0330 \pm 0.0001}$
$\mathbb{S}^5$	70.40	3.1670 ± 0.0022	46.81	3.1471 ± 0.0021	**46.79**	$\mathbf{0.0763 \pm 0.0028}$
$\mathbb{S}^{10}$	37.68	5.5112 ± 0.0047	19.95	5.4865 ± 0.0047	**18.99**	$\mathbf{0.0938 \pm 0.0002}$
$\mathbb{S}^{20}$	20.44	12.4084 ± 0.0008	10.45	12.3972 ± 0.0021	**9.44**	$\mathbf{0.2542 \pm 0.0001}$
$\mathbb{S}^{50}$	8.82	55.1251 ± 0.0010	8.31	55.0981 ± 0.0017	**4.32**	$\mathbf{0.7822 \pm 0.0002}$
$\mathbb{S}^{100}$	7.30	180.1550 ± 0.0409	6.57	180.1470 ± 0.0476	**2.95**	$\mathbf{1.2838 \pm 0.0002}$
E (2)	45.17	2.3167 ± 0.0006	56.48	2.3075 ± 0.0002	**44.75**	$\mathbf{0.0508 \pm 0.0002}$
E (3)	**48.35**	2.4337 ± 0.0007	50.36	2.4048 ± 0.0020	48.68	$\mathbf{0.0403 \pm 0.0003}$
E (5)	48.36	3.1432 ± 0.0015	27.05	3.1245 ± 0.0025	**24.85**	$\mathbf{0.0530 \pm 0.0003}$
E (10)	23.57	5.1367 ± 0.0019	11.86	5.0912 ± 0.0015	**10.86**	$\mathbf{0.0617 \pm 0.0002}$
E (20)	13.59	12.5121 ± 0.0026	7.85	12.4686 ± 0.0023	**5.60**	$\mathbf{0.2530 \pm 0.0002}$
E (50)	6.46	54.7758 ± 0.0019	6.87	54.7446 ± 0.0015	**2.52**	$\mathbf{0.7813 \pm 0.0001}$
E (100)	5.29	177.0563 ± 0.1030	4.87	177.0098 ± 0.0526	**1.95**	$\mathbf{1.2813 \pm 0.0002}$
$\mathbb{T}^2$	**437.49**	2.4922 ± 0.0021	467.28	2.4912 ± 0.0021	467.19	$\mathbf{0.1128 \pm 0.0008}$
$\mathbb{H}^2$	74.70	2.5148 ± 0.0015	**74.69**	2.5139 ± 0.0022	74.72	$\mathbf{0.0831 \pm 0.0001}$
Paraboloid	58.80	2.2218 ± 0.0008	56.74	2.2359 ± 0.0009	**56.68**	$\mathbf{0.1344 \pm 0.0001}$
Hyperbolic Paraboloid	54.21	2.2172 ± 0.0010	58.32	2.2114 ± 0.0010	**46.69**	$\mathbf{0.0730 \pm 0.0004}$
Gaussian Distribution	62.73	0.7512 ± 0.0021	75.61	0.7504 ± 0.0016	**62.43**	$\mathbf{0.0285 \pm 0.0006}$
Fréchet Distribution	11.58	0.6400 ± 0.0021	50.27	0.6389 ± 0.0056	**11.15**	$\mathbf{0.0120 \pm 0.0004}$
Cauchy Distribution	24.96	0.7524 ± 0.0008	26.46	0.7438 ± 0.0079	**24.91**	$\mathbf{0.0167 \pm 0.0004}$
Pareto Distribution	11.63	0.6514 ± 0.0004	31.71	0.6544 ± 0.0026	**11.03**	$\mathbf{0.0134 \pm 0.0004}$
VAE MNIST	304.26	463.7502 ± 0.0112	296.61	463.6756 ± 0.4144	**287.44**	$\mathbf{18.3540 \pm 0.0084}$
VAE CelebA	3173.79	1596.7343 ± 0.0697	3227.14	1445.5969 ± 106.3969	**3170.45**	$\mathbf{40.1097 \pm 0.0064}$

Table 2 The table shows the sum of squared geodesic length for the estimated Fréchet mean for *ADAM*, *RMSprop Momentum* and *GEORCE-FM* on a GPU for Finsler manifolds constructed with a Riemannian background metric equipped with the wind field in Eq. 24. The methods were terminated if the ℓ^2-norm of the average gradient across the data was less than 10^{-4} or $1,000$ iterations have been reached. $\mathrm{E}(n)$ denotes an Ellipsoid of dimension n, while $\mathcal{P}(n)$ denotes the space of $n \times n$ symmetric positive definite matrices. When the computational time was longer than 24 hours, or if the method returns nan, the value is set to $-$. Since the VAE's are learned manifolds based on data, and therefore can be unstable, the methods are terminated if the average gradient across the data was less than 10^{-3}. Further, to avoid memory complexity for the learned manifolds in storing the neural network, the geodesics are updated sequentially for VAE MNIST and VAE CelebA, while for the other manifolds the geodesics are updated in parallel. The VAE's use 10 datapoints, while the remaining manifolds use 100 synthetically generated datapoints. The data is described in Appendix D.

	Adaptive Estimation of the Fréchet Mean for Riemannian Manifolds		
	Batch 10%		
Manifold	ADAM	RMSprop Momentum	GEORCE-FM
$\mathbb{S}^2$	2370.92	**2033.23**	2179.35
$\mathbb{S}^3$	2339.52	**1378.82**	1430.56
$\mathbb{S}^5$	2025.96	851.12	**821.47**
$\mathbb{S}^{10}$	1277.13	394.19	**374.00**
$\mathbb{S}^{20}$	810.52	188.29	**173.39**
$\mathbb{S}^{50}$	378.83	83.76	**68.65**
$\mathbb{S}^{100}$	192.71	51.20	**33.50**
E (2)	1364.47	1482.42	**1308.92**
E (3)	1446.98	**846.17**	1163.30
E (5)	1333.88	504.87	**439.33**
E (10)	846.91	237.06	**211.01**
E (20)	556.93	116.52	**101.55**
E (50)	267.66	52.28	**40.59**
E (100)	−/−	−/−	**19.83**
$\mathbb{T}^2$	14986.02	**14839.74**	14953.47
$\mathbb{H}^2$	**1255.85**	1256.54	1256.70
Paraboloid	**1040.85**	1142.48	1041.53
Hyperbolic Paraboloid	**1109.81**	1117.19	1110.46
Gaussian Distribution	1509.58	1667.65	**1437.59**
Fréchet Distribution	308.39	26604758.00	**306.09**
Cauchy Distribution	648.87	645.29	**585.35**
Pareto Distribution	273.26	798.23	**269.40**
VAE MNIST	7886.67	6890.60	**6462.85**
VAE CelebA	−/−	−/−	**80255.43**

Table 3 The table shows the sum of squared geodesic length for the estimated Fréchet mean for *ADAM*, *RMSprop Momentum* and *GEORCE-FM* on a GPU using only mini-batches of data. E(n) denotes an Ellipsoid of dimension n, while $\mathcal{P}(n)$ denotes the space of $n \times n$ symmetric positive definite matrices. When the computational time was longer than 24 hours, or if the method returns nan, the value is set to $-$. The VAE's use 100 datapoints, while the remaining manifolds use $1,000$ synthetically generated datapoints. The data is described in Appendix D.

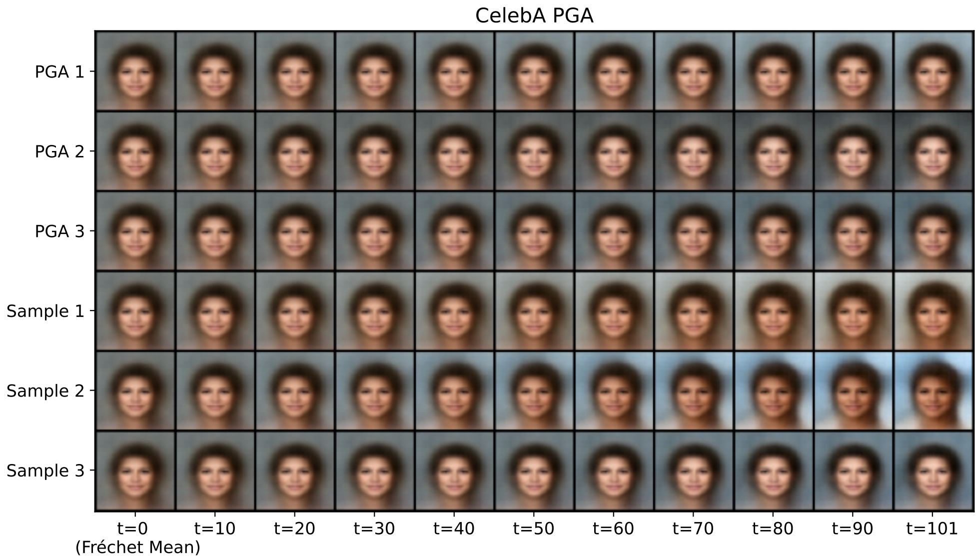

Fig. 9 The first three rows shows the three-most variance explaining principal geodesics for the 10 images in Fig. 8. The bottom three rows shows samples using the three principal geodesics by $\mathrm{Exp}\left(\mu, \sum_{k=1}^{3} \alpha_k v_k\right)$, where μ is the Fréchet mean, $\{v_k\}_{k=1}^{3}$ are the initial direction of the principal geodesics and $\alpha_k \sim \mathcal{N}(0,1)$. We compute the exponential map using the ODE in Eq. 3 with the explicit Runge-Kutta method of order 5 (4) [54].

the logarithmic map as $\mathrm{Log}_\mu(a_i) \approx u_{0,i}$. In Fig. 9, we illustrate the principal geodesics and samples for the VAE on the CelebA dataset [52] for the images in Fig. 8.

7 Conclusion

In this paper, we have introduced *GEORCE-FM*; an algorithm to compute geodesics and the Fréchet mean simultaneously, which significantly reduces runtime for Riemannian and Finslerian manifolds. We have shown that the algorithm has global convergence and local quadratic convergence. In addition, we have derived an adaptive extension to make it scalable for a large number of data and have proven that it converges in expectation to the Fréchet mean. Empirically, we have shown that *GEORCE-FM* exhibits superior runtime and accuracy of the Fréchet mean compared to other optimization methods.

Limitations

Estimating geodesics and the Fréchet mean at the same time can be numerically expensive, as it requires storing multiple matrices for each grid point, which makes it less efficient in memory. Although this can be solved by only computing the metric matrix function when needed, this will increase the runtime as the metric matrix function would have to be computed multiple times again in each iteration. Thus, the method provides a trade-off between memory efficiency and runtime efficiency, where the optimal choice depends on the number of data and dimensionality. In the adaptive update scheme for *GEORCE-FM* we have proposed rather simple update formulas. In future research, it should be studied whether the adaptive scheme can be improved by incorporating, e.g., momentum of the update similar to other stochastic optimizers. Our algorithm also assumes access to a local coordinate system, which can be a limitation for manifolds, where this is not practically possible.

Supplementary information. The supplementary information in the paper consists of the appendix for proofs, details on experiments and manifolds as well as additional experiments.

Acknowledgements. We acknowledge the Python library JAX [55], from which our implementation has been built.

Declarations

- Funding: Research grant 42062 from VILLUM FONDEN; funding from the European Research Council (ERC) under the European Union's Horizon research and innovation programme (grant agreement 101125993); funding from the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606).

- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use): NA
- Ethics approval and consent to participate: NA
- Consent for publication: Currently under review.
- Data availability: All data used are described in the appendix.
- Materials availability: NA
- Code availability: The code can be found at https://github.com/FrederikMR/georce_fm
- Author contribution: FMR derived the algorithm and implemented the method as well as benchmarks. SH and SM supervised the project with comments and corrections. SM and FMR proved Lemma 1. All authors reviewed the manuscript.

References

[1] Bortoli, V. D. *et al.* Oh, A. H., Agarwal, A., Belgrave, D. & Cho, K. (eds) *Riemannian score-based generative modelling.* (eds Oh, A. H., Agarwal, A., Belgrave, D. & Cho, K.) *Advances in Neural Information Processing Systems* (2022). URL https://openreview.net/forum?id=oDRQGo8I7P.

[2] Jo, J. & Hwang, S. J. Generative modeling on manifolds through mixture of riemannian diffusion processes (2024).

[3] Mathieu, E. & Nickel, M. Riemannian continuous normalizing flows (2020).

[4] Rozen, N., Grover, A., Nickel, M. & Lipman, Y. Moser flow: divergence-based generative modeling on manifolds (2021).

[5] Hsieh, M. A. *et al.* Reactive motion generation on learned riemannian manifolds. *Int. J. Rob. Res.* **42**, 729–754 (2023). URL https://doi.org/10.1177/02783649231193046.

[6] Simeonov, A. *et al.* Neural descriptor fields: Se(3)-equivariant object representations for manipulation (2022).

[7] Feiten, W., Lang, M. & Hirche, S. Rigid motion estimation using mixtures of projected gaussians 1465–1472 (2013).

[8] Watson, J. L. *et al.* Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv* (2022). URL https://www.biorxiv.org/content/early/2022/12/10/2022.12.09.519842.

[9] Detlefsen, N., Hauberg, S. & Boomsma, W. Learning meaningful representations of protein sequences. *Nature Communications* **13** (2022). Publisher Copyright: © 2022, The Author(s).

[10] Shapovalov, M. & Dunbrack, R. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure (London, England : 1993)* **19**, 844–58 (2011).

[11] Fréchet, M. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré* **10**, 215–310 (1948). URL http://www.numdam.org/item/AIHP_1948__10_4_215_0/.

[12] Fletcher, T. Geodesic regression on riemannian manifolds. *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy - Geometrical and Statistical Methods for Modelling Biological Shape Variability* 75–86 (2011).

[13] Fletcher, P., Lu, C., Pizer, S. & Joshi, S. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging* **23**, 995–1005 (2004).

[14] Pennec, X. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision* **25**, 127–154 (2006).

[15] Hauberg, S. Principal curves on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**, 1915–1921 (2016).

[16] Henry, G. & Rodriguez, D. Kernel density estimation on riemannian manifolds: Asymptotic results (2009). URL https://doi.org/10.1007/s10851-009-0145-2.

[17] Banerjee, M., Chakraborty, R., Ofori, E., Vaillancourt, D. & Vemuri, B. C. Nonlinear regression on riemannian manifolds and its applications to neuro-image analysis (2015). URL https://doi.org/10.1007/978-3-319-24553-9_88.

[18] Udriste, C. *Convex Functions and Optimization Methods on Riemannian Manifolds* Mathematics and Its Applications (Springer Netherlands, 2013). URL https://books.google.com/books?id=uX_rCAAAQBAJ.

[19] do Carmo, M. *Riemannian Geometry* Mathematics (Boston, Mass.) (Birkhäuser, 1992). URL https://books.google.dk/books?id=uXJQQgAACAAJ.

[20] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization (2014). URL http://arxiv.org/abs/1412.6980. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[21] Shanno, D. & Kettler, P. Optimal conditioning of quasi-newton methods. *Mathematics of Computation - Math. Comput.* **24**, 657–657 (1970).

[22] Broyden, C. G. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics* **6**, 76–90 (1970). URL https://doi.org/10.1093/imamat/6.1.76.

[23] Fletcher, R. A new approach to variable metric algorithms. *The Computer Journal* **13**, 317–322 (1970). URL https://doi.org/10.1093/comjnl/13.3.317.

[24] Goldfarb, D. A family of variable-metric methods derived by variational means. *Mathematics of Computation* **24**, 23–26 (1970). URL https://api.semanticscholar.org/CorpusID:790344.

[25] Rygaard, F. M. & Hauberg, S. Georce: A fast new control algorithm for computing geodesics (2025). URL https://arxiv.org/abs/2505.05961. arXiv:2505.05961.

[26] Ohta, S. *Comparison Finsler Geometry* Springer Monographs in Mathematics (Springer International Publishing, 2021). URL https://books.google.dk/books?id=4Mh6zgEACAAJ.

[27] Kendall, W. Probability, convexity, and harmonic maps with small image i: Uniqueness and fine existence. *Proceedings of The London Mathematical Society - PROC LONDON MATH SOC* **s3-61**, 371–406 (1990).

[28] Karcher, H. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics* **30**, 509–541 (1977). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160300502.

[29] Ziezold, H. *On Expected Figures and a Strong Law of Large Numbers for Random Elements in Quasi-Metric Spaces*, 591–602 (Springer Netherlands, Dordrecht, 1977). URL https://doi.org/10.1007/978-94-010-9910-3_63.

[30] Bhattacharya, R. & Patrangenaru, V. Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics* **31**, 1 – 29 (2003). URL

https://doi.org/10.1214/aos/1046294456.

[31] Huckemann, S. F. & Eltzner, B. Backward nested descriptors asymptotics with inference on stem cell differentiation. *The Annals of Statistics* **46**, 1994 – 2019 (2018). URL https://doi.org/10.1214/17-AOS1609.

[32] Brooks, D., Schwander, O., Barbaresco, F., Schneider, J.-Y. & Cord, M. Wallach, H. *et al.* (eds) *Riemannian batch normalization for spd neural networks.* (eds Wallach, H. *et al.*) *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Inc., 2019). URL https://proceedings.neurips.cc/paper_files/paper/2019/file/6e69ebbfad976d4637bb4b39de261bf7-Paper.pdf.

[33] Lou, A. *et al.* Differentiating through the fréchet mean (2020).

[34] Eichfelder, G., Hotz, T. & Wieditz, J. An algorithm for computing fréchet means on the sphere. *Optimization Letters* **13** (2019). URL https://doi.org/10.1007/s11590-019-01415-y.

[35] Kierzenka, J. & Shampine, L. F. A bvp solver based on residual control and the maltab pse. *ACM Trans. Math. Softw.* **27**, 299–316 (2001). URL https://doi.org/10.1145/502800.502801.

[36] Noakes, L. A global algorithm for geodesics. *Journal of the Australian Mathematical Society. Series A. Pure Mathematics and Statistics* **65**, 37–50 (1998).

[37] Kaya, C. Y. & Noakes, J. L. Leapfrog for optimal control. *SIAM Journal on Numerical Analysis* **46**, 2795–2817 (2008). URL https://doi.org/10.1137/060675034.

[38] Kaya, C. & Noakes, J. Geodesics and an optimal control algorithm (1997).

[39] Hennig, P. & Hauberg, S. Kaski, S. & Corander, J. (eds) *Probabilistic Solutions to Differential Equations and their Application to Riemannian Statistics.* (eds Kaski, S. & Corander, J.) *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Vol. 33 of *Proceedings of Machine Learning Research*, 347–355 (PMLR, Reykjavik, Iceland, 2014). URL https://proceedings.mlr.press/v33/hennig14.html.

[40] Arvanitidis, G., Hauberg, S., Hennig, P. & Schober, M. Chaudhuri, K. & Sugiyama, M. (eds) *Fast and robust shortest paths on manifolds learned from data.* (eds Chaudhuri, K. & Sugiyama, M.) *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Vol. 89 of *Proceedings of Machine Learning Research*, 1506–1515 (PMLR, 2019). URL https://proceedings.mlr.press/v89/arvanitidis19a.html.

[41] Shao, H., Kumar, A. & Thomas Fletcher, P. The riemannian geometry of deep generative models (2018).

[42] Arvanitidis, G., Hansen, L. K. & Hauberg, S. Latent space oddity: on the curvature of deep generative models (2018). URL https://openreview.net/forum?id=SJzRZ-WCZ.

[43] Eltzner, B., Hansen, P., Huckemann, S. F. & Sommer, S. Diffusion means in geometric spaces. *Bernoulli* (2021). URL https://api.semanticscholar.org/CorpusID:235187429.

[44] Nielsen, F. Fast proxy centers for the jeffreys centroid: The jeffreys–fisher–rao center and the gauss–bregman inductive center. *Entropy* **26** (2024). URL https://www.mdpi.com/1099-4300/26/12/1008.

[45] Fletcher, P. T., Venkatasubramanian, S. & Joshi, S. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage* **45**, S143–S152 (2009). Epub 2008 Nov 13.

[46] Sommer, S., Arnaudon, A., Kuhnel, L. & Joshi, S. Cardoso, M. J. *et al.* (eds) *Bridge simulation and metric estimation on landmark manifolds.* (eds Cardoso, M. J. *et al.*) *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, 79–91 (Springer International Publishing, Cham, 2017).

[47] Rygaard, F. M., Markvorsen, S., Hauberg, S. & Sommer, S. Score matching riemannian diffusion means (2025). URL https://arxiv.org/abs/2502.13106. arXiv:2502.13106.

[48] Miyamoto, H. K., Meneghetti, F. C. C., Pinele, J. & Costa, S. I. R. On closed-form expressions for the fisher-rao distance. *Information Geometry* **7** (2024).

[49] Armijo, L. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics* **16**, 1–3 (1966).

[50] Arnaudon, M. & Nielsen, F. Medians and means in finsler geometry. *LMS Journal of Computation and Mathematics* **15**, 23–37 (2012).

[51] Ruder, S. An overview of gradient descent optimization algorithms (2017). URL https://arxiv.org/abs/1609.04747. arXiv:1609.04747.

[52] Liu, Z., Luo, P., Wang, X. & Tang, X. Deep learning face attributes in the wild (2015).

[53] Piro, L., Tang, E. & Golestanian, R. Optimal navigation strategies for microswimmers on curved manifolds. *Physical Review Research* **3** (2021). URL http://dx.doi.org/10.1103/PhysRevResearch.3.023125.

[54] Dormand, J. & Prince, P. A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics* **6**, 19–26 (1980). URL https://www.sciencedirect.com/science/article/pii/0771050X80900133.

[55] Bradbury, J. *et al.* JAX: composable transformations of Python+NumPy programs (2018). URL http://github.com/google/jax.

[56] Gallot, S., Hulin, D. & Lafontaine, J. *Riemannian Geometry* Universitext (Springer Berlin Heidelberg, 2004). URL https://books.google.dk/books?id=6F4Umpws_gUC.

[57] Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes (2014). http://arxiv.org/abs/1312.6114v10.

[58] Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* **29**, 141–142 (2012).

[59] Duchi, J., Hazan, E. & Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* **12**, 2121–2159 (2011).

Appendix

A Proofs and derivations

A.1 Equivalence between Fréchet means and energy formulation for Riemannian manifolds

Lemma 7 *Let μ^{FM} denote the set of minima of eq. 1, and let μ^{Energy} denote the set of minima of eq. 8 with $w_i = 1$ for all $i \in \{1, \dots, N\}$. Then*

$$\mu^{\mathrm{Energy}} = \mu^{\mathrm{FM}}.$$

Proof Let $\mathcal{L}_\gamma(a, y)$ denote the length for a curve γ between points $a, y \in \mathcal{M}$. Similarly, let $\mathcal{E}_\gamma(a, y)$ denote the energy for a curve γ between points $a, y \in \mathcal{M}$. By the proof of Lemma 2.3 page 194 in [19] it follows that

$$\mathcal{E}_\gamma(a, y) \propto \left(\mathcal{L}_\gamma(a, y)\right)^2,$$

if and only if γ is a geodesic. This implies that

$$\arg\min_{y \in \mathcal{M}} \sum_{i=1}^{N} \mathrm{dist}^2(y, a_i) = \arg\min_{y \in \mathcal{M}} \sum_{i=1}^{N} \min_{\gamma_i} \left(\mathcal{L}_{\gamma_i}(y, a_i)\right)^2 = \arg\min_{y \in \mathcal{M}} \sum_{i=1}^{N} \min_{\gamma_i} \mathcal{E}_{\gamma_i}(y, a_i),$$

since minimizing the energy is a geodesic [56]. In the latter equation, we minimize the starting point of the minimizing curves that connect the start and data points. This can equivalently be minimized jointly as

$$\arg\min_{\substack{y \in \mathcal{M} \\ \{\gamma_i\}_{i=1}^N}} \sum_{i=1}^{N} \mathcal{E}_\gamma(a_i, y),$$

where we have changed the order of the start and end point, since the Riemannian energy is symmetric. □

A.2 Equivalence between Fréchet means and energy formulation for Finslerian manifolds

Lemma 8 *Let μ^{FM} denote the set of minima of eq. 18, and let μ^{Energy} denote the set of minima given by*

$$\mu^{\mathrm{Energy}}, \{\gamma_i\}_{i=1}^N = \arg\min_{\substack{y \in \mathcal{M} \\ \{\gamma_i\}_{i=1}^N}} \sum_{i=1}^{N} \mathcal{E}^{(F)}_{\gamma_i}(y, a_i), \tag{26}$$

where the energy $\mathcal{E}^{(F)}$ is defined using the fundamental tensor g, i.e.

$$\mathcal{E}^{(F)}_\gamma(y, a_i) = \int_0^1 g\left(\gamma(t), \dot{\gamma}(t)\right) \mathrm{d}t,$$

with $\gamma(0) = y$ and $\gamma(1) = a_i$, Then

$$\mu^{\mathrm{Energy}} = \mu^{\mathrm{FM}}.$$

Proof Let $\mathcal{L}_{\gamma}^{(F)}(a,y)$ denote the length for a curve γ between points $a,y \in \mathcal{M}$. Similarly, let $\mathcal{E}_{\gamma}^{(F)}(a,y)$ denote the energy for a curve γ between points $a,y \in \mathcal{M}$. By Cauchy-Schwarz inequality it follows competently similar to Lemma 2.3 in [19]

$$\left(\mathcal{L}_{\gamma}^{(F)}(a,y)\right)^2 \leq c\mathcal{E}_{\gamma}^{(F)}(a,y),$$

where it is assumed that γ is paraemetrized by the interval $[0,c]$. Equality holds if and only if the $g(\dot{\gamma}(t),\dot{\gamma}(t))$ is constant. By [26] 3.3 page 24 then critical points of the energy functional are both locally length minimizing and of constant speed. Thus if and only if γ is a geodesic, then

$$\underset{y\in\mathcal{M}}{\arg\min}\sum_{i=1}^{N}\mathrm{dist}_F^2(y,a_i) = \underset{y\in\mathcal{M}}{\arg\min}\sum_{i=1}^{N}\min_{\gamma_i}\left(\mathcal{L}_{\gamma_i}(y,a_i)\right)^2 = \underset{y\in\mathcal{M}}{\arg\min}\sum_{i=1}^{N}\min_{\gamma_i}\mathcal{E}_{\gamma_i}^{(F)}(y,a_i),$$

since minimizing the energy is a geodesic [56]. In the latter equation, we minimize the starting point of the minimizing curves that connect the start and data points. This can equivalently be minimized jointly as

$$\underset{\substack{y\in\mathcal{M}\\ \{\gamma_i\}_{i=1}^N}}{\arg\min}\sum_{i=1}^{N}\mathcal{E}_{\gamma}^{(F)}(y,a_i).$$

□

A.3 Assumptions

For the proof of local quadratic convergence we will assume the same regularity of the discretized energy functional as in [25]. We re-state the assumptions below

Assumption 9 ([25]) *We assume the following regarding the discretized energy functional.*

- *We assume that the discretized energy functional $E(z)$ is locally strictly convex in the (local) minimum point $z^* = (x^*,u^*)$ in the sense that*

  $$\exists\epsilon>0:\ \forall z\in B_\epsilon(z^*),z\neq z^*:\ \forall\alpha]0,1[:\ E\left((1-\alpha)z+\alpha z^*\right)<(1-\alpha)E(z)+\alpha E(z^*),$$

  where $B_\epsilon = \{z \mid \|z-z^*\| < \epsilon\}$.
- *Assume that the discretized energy functional $E(z)$ is smooth and thus locally Lipschitz, and consider the first order Taylor approximation of the discretized energy functional*

  $$\Delta E = \langle \nabla E(z_0), \Delta z\rangle + \mathcal{O}(\Delta z)\|\Delta z\|,$$

  Then the term $\mathcal{O}(\Delta z)\|\Delta z\|$ can be re-written as $\tilde{\mathcal{O}}\left(\|\Delta z\|^2\right)$ locally.

In the proofs for global and local quadratic convergence, we let $\|\cdot\|$, $\langle\cdot,\cdot\angle$ and ∇ denote the Euclidean norm, inner product and gradient, respectively.

A.4 Necessary conditions

Proposition 10 *The necessary conditions for a minimum in Eq. 10 is*

$$
\begin{aligned}
&2w_iG(x_{t,i})u_{t,i} + \mu_{t,i} = 0, \quad t = 0,\dots,T-1,\ i = 1,\dots,N,\\
&x_{t+1,i} = x_{t,i} + u_{t,i}, \quad t = 0,\dots,T-1,\ i = 1,\dots,N,\\
&\nabla_x \left[w_i u_{t,i}^\top G(x) u_{t,i}\right]\Big|_{x=x_{t,i}} + \mu_t = \mu_{t-1}, \quad t = 1,\dots,T-1,\\
&0 = \sum_{i=1}^N \mu_{T-1,i},\\
&x_{0,i} = a_i, x_{T,i} = y, \quad i = 1,\dots,N.
\end{aligned} \tag{27}
$$

where $\mu_{t,i} \in \mathbb{R}^d$ *for* $t = 0,\dots,T-1$ *and* $i = 1,\dots,N$*.*

Proof The corresponding Hamiltonian function of eq. 10 for $t = 0,\dots,T-1$ and $i = 1,\dots,N$ is

$$
\begin{aligned}
H_t\left(x_t,u_t,\mu_t\right) &= \sum_{i=1}^N H_{t,i}\left(x_{t,i},u_{t,i},\mu_{t,i}\right),\\
H_{t,i}\left(x_{t,i},u_{t,i},\mu_{t,i}\right) &= w_i u_{t,i}^\top G(x_{t,i})u_{t,i} + \mu_{t,i}^\top\left(x_{t,i}+u_{t,i}\right).
\end{aligned} \tag{28}
$$

The time-discrete version of Pontryagin's maximum problem on for eq. 10 gives the following optimization problem with the slight modification that the endpoint, $x_{T,i}$, replaced by the variable Fréchet mean y compared to [25]

$$\min_{u_{t,i}} \quad \sum_{i=1}^N \sum_{t=0}^T H_{t,i}(x_{t,i},u_{t,i},\mu_{t,i}) \tag{29}$$

$$\text{s.t.} \quad x_{t+1,i} = x_{t,i} + u_{t,i}, \quad t = 0,\dots,T-2,\ i = 1,\dots,N, \qquad \text{(state equation)}, \tag{30}$$

$$y = x_{T-1,i} + u_{T-1,i}, \quad i = 1,\dots,N, \qquad \text{(state equation)}, \tag{31}$$

$$\nabla_{x_{t,i}} H_{t,i}(x_{t,i},u_{t,i},\mu_{t,i}) = \mu_{t-1,i}, \quad t = 1,\dots,T-1,\ i = 1,\dots,N \qquad \text{(co-state equation)}, \tag{32}$$

$$0 = \sum_{i=1}^N \mu_{T-1,i}, \quad i = 1,\dots,N, \qquad \text{(co-state equation)}, \tag{33}$$

$$x_{0,i} = a_i, \quad i = 1,\dots,N. \tag{34}$$

The minimization problem can be decomposed into the following sub-problems for each $i \in \{1,\dots,N\}$, where the equality constraints on the control variables are relaxed in Lagrange sense

$$\min_{u_{t,i}} H_{t,i}\left(x_{t,i},u_{t,i},\mu_{t,i}\right), \quad t = 0,\dots,T-1.$$

We observe that $H_{t,i}\left(x_{t,i},u_{t,i},\mu_{t,i}\right)$ is strictly convex in $u_{t,i}$, since $G(x_{t,i})$ is positive definite. The stationary point wrt. $u_{t,i}$ is therefore also the global minimum point, which gives the following equations including state and co-state equations from eq. 10

$$\begin{aligned}
&2w_iG(x_{t,i})u_{t,i} + \mu_{t,i} = 0, \quad t = 0, \dots, T-1,\, i = 1, \dots, N, \\
&x_{t+1,i} = x_{t,i} + u_{t,i}, \quad t = 0, \dots, T-2,\, i = 1, \dots, N, \\
&y = x_{T-1,i} + u_{T-1,i}, \quad i = 1, \dots, N, \\
&\nabla_x \left[w_i u_{t,i}^\top G(x) u_{t,i} \right] \Big|_{x=x_{t,i}} + \mu_t = \mu_{t-1}, \quad t = 1, \dots, T-1, \\
&0 = \sum_{i=1}^N \mu_{T-1,i}, \\
&x_{0,i} = a_i, x_T = b, \quad i = 1, \dots, N.
\end{aligned} \tag{35}$$

□

A.5 Update scheme

Proposition 11 *The update scheme for u_t, μ_t and x_t in eq. 12 can be reduced to*

$$\begin{aligned}
&y = W^{-1}V, \\
&\mu_{T-1,i} = \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1} \left(2w_i(a_i - y) - \sum_{t=0}^{T-1} G_{t,i}^{-1} \sum_{t>j}^{T-1} \nu_{j,i} \right), \quad i = 1, \dots, N, \\
&u_{t,i} = -\frac{1}{2w_i} G_{t,i}^{-1} \left(\mu_{T-1,i} + \sum_{j>t}^{T-1} \nu_{j,i} \right), \quad t = 0, \dots, T-1,\, i = 1, \dots, N \\
&x_{t+1,i} = x_{t,i} + u_{t,i}, \quad t = 0, \dots, T-2,\, i = 1, \dots, N, \\
&x_{0,i} = a_i \quad i = 1, \dots, N,
\end{aligned}$$

where

$$\begin{aligned}
W &= \sum_{i=1}^N w_i \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1}, \\
V &= \sum_{i=1}^N w_i \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1} a_i - \frac{1}{2} \sum_{i=1}^N \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1} \sum_{t=0}^{T-1} G_{t,i}^{-1} \sum_{j>t}^{T-1} \nu_{j,i}.
\end{aligned}$$

Proof The equation system solved in *GEORCE-FM* in iteration $k+1$ is

$$\begin{aligned}
&\nu_{t,i} + \mu_{t,i} = \mu_{t-1,i}, \quad t = 1, \dots, T-1,\, i = 1, \dots, N, \\
&2w_iG_{t,i}u_{t,i} + \mu_{t,i} = 0, \quad t = 0, \dots, T-1,\, i = 1, \dots, N, \\
&\sum_{t=0}^{T-1} u_{t,i} = y - a_i, \quad i = 1, \dots, N, \\
&\sum_{i=1}^N \mu_{T-1,i} = 0, \\
&\nu_{t,i} := \nabla_x \left(w_i u_{t,i}^\top G(x) u_{t,i} \right) \Big|_{y=x_{t,i}^{(k)}, u_{t,i}=u_{t,i}^{(k)}}, \quad t = 1, \dots, T-1,\, i = 1, \dots, N \\
&G_{t,i} := G\left(x_{t,i}^{(k)} \right), \quad t = 0, \dots, T-1,
\end{aligned}$$

where the related state variables (before line-search) are found by the state equation

$$x_{t+1,i} = x_{t,i} + u_{t,i}, t = 0, \dots, T-2, i = 1, \dots, N, x_{0,i} = a_i, \quad i = 1, \dots, N.$$

As we will see in the following $y = x_{T-1,i} + u_{T-1,i}$ for $i = 1, \dots, N$ is omitted in the update step, since y will be determined initially in each iteration. By re-arranging the terms we see that

$$\mu_{t,i} = \mu_{T-1,i} + \sum_{j>t}^{T-1} \nu_{j,t}, \quad t = 0, \dots, T-1, i = 1, \dots, N,$$

$$u_{t,i} = -\frac{1}{2} G_{t,i}^{-1} \mu_{t,i} = -\frac{1}{2w_i} G_{t,i}^{-1} \left(\mu_{T-1,i} + \sum_{j>t}^{T-1} \nu_{j,i} \right), \quad t = 0, \dots, T-1, i = 1, \dots, N,$$

$$-\frac{1}{2\omega_i} \sum_{t=0}^{T-1} G_{t,i}^{-1} \left(\mu_{T-1,i} + \sum_{j>t}^{T-1} g_{ji} \right) = y - a_i, \quad i = 1, \dots, N,$$

where it is exploited that $G_{t,i}$ is positive definite for all t, i and has a well defined inverse. These equations can be used to provide explicit solutions for the state and control variables. First observe that

$$\mu_{T-1,i} = \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1} \left(2w_i(a_i - y) - \sum_{t=0}^{T-1} G_{t,i}^{-1} \sum_{j>t}^{T-1} \nu_{j,i} \right), \quad i = 1, \dots, N.$$

Applying this result, we deduce that

$$\sum_{i=1}^{N} \mu_{T-1,i} = 0 \Leftrightarrow \sum_{i=1}^{N} \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1} \left(2w_i(a_i - y) - \sum_{t=0}^{T-1} G_{t,i}^{-1} \sum_{j>t}^{T-1} \nu_{j,i} \right) = 0,$$

which implies that

$$y = W^{-1} V.$$

where

$$W = \sum_{i=1}^{N} w_i \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1},$$

$$V = \sum_{i=1}^{N} w_i \left(\sum_{t=0}^{T-1} G_{ti}^{-1} \right)^{-1} a_i - \frac{1}{2} \sum_{i=1}^{N} \left(\sum_{t=0}^{T-1} G_{t,i}^{-1} \right)^{-1} \sum_{t=0}^{T-1} G_{ti}^{-1} \sum_{j>t}^{T-1} g_{ji}.$$

Thus, we get the update scheme in Proposition 3. □

A.6 Global convergence

We prove that *GEORCE-FM* has global convergence similar to the original *GEORCE* by extending the proof to estimating multiple geodesics and end point. We follow the same approach as in [25] and adapt the Lemma's and proofs to the general setting.

Lemma 12 *Assume that $\left(x^{(j)}, u^{(j)}\right)$ is a feasible solution, then the following properties hold.*

- *There exists a unique solution, $\left(x^{(j+1)}, u^{(j+1)}\right)$, to the system of equations in eq. 11 based on $\left(x^{(j)}, u^{(j)}\right)$.*

- *All linear combinations* $(1-\alpha)\left(x^{(j)}, u^{(j)}\right)+\alpha\left(x^{(j+1)}, u^{(j+1)}\right)$ *for* $0 \leq \alpha \leq 1$ *are feasible solutions.*

Proof Since the metric tensors, $\left\{G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)\right\}_{s=0,\dots,T-1,i=1,\dots,N}$ are positive definite matrices, then the matrices are regular, and the inverse is unique, and so are sum of the inverse matrices. That means that there exists a unique solution to the system of equations in iteration j in eq. 41 wrt. $\left\{u_{s,i}\right\}_{s=0,\dots,T-1,i=1,\dots,N}$ and $\left\{x_{s,i}^{(j)}\right\}_{s=0,\dots,T-1,i=1,\dots,N}$ by the state equations.

At first, assume that $\left(x^{(j+1)}, u^{(j+1)}\right)$ is well-defined in the domain of the local chart.

Since the solution, $\left(x^{(j+1)}, u^{(j+1)}\right)$ is also a feasible solution according to eq. 41, then the linear combination of two feasible solutions will also be a feasible solution, i.e.

$$\sum_{s=0}^{T-1} u_{s,i}^{(j+1)} = y - a_i, \quad i=1,\dots,N.$$

The new solution based on the linear combination gives

$$(a-\alpha)\sum_{s=0}^{T-1} u_{s,i}^{(j)} + \alpha \sum_{s=0}^{T-1} u_{s,i}^{(j+1)} = (1-\alpha)(y-a_i) + \alpha(y-a_i) = y - a_i, \quad i=1,\dots,N,$$

$$(1-\alpha)\left(a_i, x_{1,i}^{(j)}, \dots, x_{T-1,i}^{(j)}, y\right) + \alpha\left(a_i, x_{1,i}^{(j)}, \dots, x_{T-1,i}^{(j)}, y\right)$$
$$= \left(a_i, (1-\alpha)x_{1,i}^{(j)} + \alpha x_{1,i}^{(j)}, \dots, (1-\alpha)x_{T-1,i}^{(j)} + \alpha x_{T-1,i}^{(j+1)}, y\right), \quad i=1,\dots,N,$$

where

$$(1-\alpha)x_{s,i}^{(j)} + \alpha x_{s,i}^{(j+1)} = (1-\alpha)\left(a_i + \sum_{k=0}^{t-1} u_{k,i}^{(j)}\right) + \alpha\left(a_i + \sum_{k=0}^{t-1} u_{k,i}^{(j+1)}\right)$$
$$= a_i + \sum_{k=0}^{t-1}(1-\alpha)u_{k,i}^{(j)} + \alpha u_{k,i}^{(j+1)}, \quad i=1,\dots,N.$$

This shows the linear combination in the state variable is feasible as it produces a consistent state variable in terms of the start and end point, and furthermore each state variable is feasible if they were determined from the linear combination of the control vectors, which proves that the linear combinations of the state and control variables are also feasible solutions.

In contrast, if $\left(x^{(j+1)}, u^{(j+1)}\right)$ is not well-defined in the local chart, then the line-search in *GEORCE-FM* will be able to determine a point belonging to the manifold and that will be on the line between $\left(x^{(j+1)}, u^{(j+1)}\right)$ and $\left(x^{(j)}, u^{(j)}\right)$ as the local chart is defined on an open set and $x^{(j)}$ is well-defined in the local chart similar to *GEORCE* [25]. The new solution is therefore a feasible solution following the argument above with this modification. □

Lemma 13 *Let* $\left\{x_{s,i}^{(j)}, u_{s,i}^{(j)}\right\}_{s=0,\dots,T,i=1,\dots,N}$ *denote the (feasible) solution after iteration* j *in GEORCE-FM. If* $\left\{x_{s,i}^{(j)}, u_{s,i}^{(j)}\right\}_{s=0,\dots,T,i=0,\dots,N}$ *decreases the objective function in the sense that there exists an* $\eta > 0$ *such that for all* $0 < \alpha \leq \eta \leq 1$*, then*

$$E\left(x^{(j)} + \alpha\left(x^{(j+1)} - x^{(j)}\right), u^{(j)} + \alpha\left(u^{(j+1)} - u^{(j)}\right)\right) < E\left(x^{(j)}, u^{(j)}\right),$$

where $x^{(j)}[i] = \left(a_i, x_{1,i}^{(j)}, \dots, x_{T-1,i}^{(j)}, y^{(j)}\right)$ *and* $u^{(j)}[i] = \left(u_{0,i}, u_{1,i}^{(j)}, \dots, u_{T-1,i}^{(j)}, u_{T,i}^{(j)}\right)$.

Proof Since E is a smooth function, the first order Taylor approximation is well defined

$$\begin{aligned}\Delta E_i\left(x,u\right) = &\sum_{s=1}^{T}\sum_{i=1}^{N}\left(\left\langle \nabla_x E(x,u)\big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)}, \Delta x_{s,i}\right\rangle + \mathcal{O}\left(\Delta x_{s,i}\right)\left\|x_{s,i}\right\|\right)\\ &+\sum_{s=1}^{T}\sum_{i=1}^{N}\left(\left\langle \nabla_u E(x,u)\big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)}, \Delta u_{s,i}\right\rangle + \mathcal{O}\left(\Delta u_{s,i}\right)\left\|u_{s,i}\right\|\right),\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ is the standard Euclidean inner product and $\lim_{v\to 0}\mathcal{O}\left(v\right)=0$. The perturbation, $\left(\Delta x_{s,i}, \Delta u_{s,i}\right)$ is restricted to be a step of *GEORCE-FM* including the end point. From the optimality conditions in eq. 41 we have that

$$\nabla_{x_{s,i}} E\left(x,u\right)\big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)} = \mu_{s-1,i} - \mu_{s,i}, \quad s=1,\dots,T-1,\ i=1,\dots,N.$$

$$\nabla_{x_{T,i}} E\left(x,u\right)\big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)} = \mu_{T-1,i}, \quad i=1,\dots,N.$$

$$\nabla_{u_{s,i}} E\left(x,u\right)\big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j+1)}\right)} = 2\omega_i G_{s,i} u_{s,i}^{(j+1)} + \omega_i \zeta_{s,i} = -\mu_{s,i}, \quad s=1,\dots,T-1,\ i=1,\dots,N.$$

Applying this to the first order Taylor expansion we have that

$$\begin{aligned}\Delta E(x,u) = &\sum_{s=1}^{T-1}\sum_{i=1}^{N}\left(\left\langle \mu_{s-1,i}-\mu_{s,i}, \Delta x_{s,i}\right\rangle + \mathcal{O}\left(\Delta x_{s,i}\right)\left\|\Delta x_{s,i}\right\|\right)\\ &+\sum_{i=1}^{N}\left\langle \mu_{T-1,i}, \Delta x_{T,i}\right\rangle\\ &+\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle \nabla_{u_{s,i}} E(x,u)\big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)}, \Delta u_{s,i}\right\rangle + \mathcal{O}\left(\Delta u_{s,i}\right)\left\|\Delta u_{s,i}\right\|\right).\end{aligned}$$

Re-arranging the terms we see that

$$\begin{aligned}
\Delta E(x,u) &= \sum_{s=0}^{T-2}\sum_{i=1}^{N}\left\langle \mu_{s,i}, \Delta x_{s+1,i}\right\rangle \\
&- \sum_{s=1}^{T-1}\sum_{i=1}^{N}\left(\left\langle \mu_{s,i}, \Delta x_{s,i}\right\rangle + \mathcal{O}\left(\Delta x_{s,i}\right)\left\|\Delta x_{s,i}\right\|\right) \\
&+ \sum_{i=1}^{N}\left\langle \mu_{T-1,i}, \Delta x_{T,i}\right\rangle \\
&+ \sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle \nabla_{u_{s,i}} E(x,u)\Big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)}, \Delta u_{s,i}\right\rangle + \mathcal{O}\left(\Delta u_{s,i}\right)\left\|\Delta u_{s,i}\right\|\right) \\
&= \sum_{s=1}^{T-2}\sum_{i=1}^{N}\left\langle \mu_{s,i}, \Delta x_{s+1,i} - \Delta x_{s}\right\rangle \\
&- \sum_{i=1}^{N}\left\langle \mu_{T-1,i}, \Delta x_{T-1,i}\right\rangle - \sum_{s=1}^{T-1}\sum_{i=1}^{N}\mathcal{O}\left(\Delta x_{s,i}\right)\left\|\Delta x_{s,i}\right\| \\
&+ \sum_{i=1}^{N}\left\langle \mu_{T-1,i}, \Delta x_{T-1,i}\right\rangle \\
&+ \sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle \nabla_{u_{s,i}} E(x,u)\Big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)}, \Delta u_{s,i}\right\rangle + \mathcal{O}\left(\Delta u_{s,i}\right)\left\|u_{s,i}\right\|\right)
\end{aligned}$$

Observe that

$$\begin{aligned}
\Delta x_{s+1,i} - \Delta x_{s,i} &= \left(x_{s+1,i}^{(j+1)} - x_{s+1,i}^{(j)}\right) - \left(x_{s,i}^{(j+1)} - x_{s,i}^{(j)}\right) \\
&= \left(x_{s+1,i}^{(j+1)} - x_{s,i}^{(j+1)}\right) - \left(x_{s+1,i}^{(j)} - x_{s,i}^{(j)}\right) \\
&= u_{s,i}^{(j+1)} - u_{s,i}^{(j)} \\
&= \Delta u_{s,i}.
\end{aligned}$$

Note that $\Delta x_{1,i} = \Delta x_{1,i} - \Delta x_{0,i}$, since $\Delta x_{0,i} = 0$. Using this we have that

$$\begin{aligned}
\Delta E(x,u) &= \sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle \mu_{s,i}, \Delta u_{s,i}\right\rangle + \mathcal{O}\left(\Delta x_{s,i}\right)\left\|\Delta x_{s,i}\right\|\right) \\
&+ \sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle \nabla_{s,i} E(x,u)\Big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)}, \Delta u_{s,i}\right\rangle + \mathcal{O}\left(\Delta u_{s,i}\right)\left\|\Delta u_{s,i}\right\|\right) \\
&= \sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle \mu_{s,i} + \nabla_{u_{s,i}} E(x,u)\Big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)}, \Delta u_{s,i}\right\rangle + \mathcal{O}\left(\Delta x_{s,i}\right)\left\|\Delta x_{s,i}\right\|\right) \\
&+ \sum_{t=0}^{T-1}\mathcal{O}\left(\Delta u_{t,i}\right)\left\|\Delta u_{t,i}\right\|.
\end{aligned}$$

Since

$$
\begin{aligned}
&\mu_{s,i} + \nabla_{u_{s,i}} E(x,u)\big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)} \\
&= -\nabla_{u_{s,i}} E(x,u)\big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j+1)}\right)} + \nabla_{u_{s,i}} E(x,u)\big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)} \\
&= -2\omega_i G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right) u_{s,i}^{(j+1)} - \omega_i \zeta_s + 2G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right) u_{s,i}^{(j)} + \omega_i \zeta_{s,i} \\
&= -2\omega_i G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)\left(u_{s,i}^{(j+1)} - u_{s,i}^{(j)}\right)
\end{aligned}
$$

Thus, the first order Taylor approximation becomes

$$
\Delta E(x,u) = \sum_{s=0}^{T-1} \sum_{i=1}^{N} \left(\left\langle -2\omega_i G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right) \Delta u_{s,i}, \Delta u_{s,i} \right\rangle + \mathcal{O}\left(\Delta x_{s,i}\right) \left\|\Delta x_{s,i}\right\| + \mathcal{O}\left(u_{s,i}\right) \left\|\Delta u_{s,i}\right\|\right).
$$

Since $G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)$ is positive definite, the first summation is negative assuming at least one non-zero vector $\left\{\Delta u_{s,i}\right\}_{s=0,\ldots,T-1,i=1,\ldots,N}$, which is the case as otherwise the solution $\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)$ is a local minimum point. Note that if $\left\{\left\|u_{s,i}\right\|\right\}_{s=0,\ldots,T-1,i=1,\ldots,N}$ converges to 0, then $\left\|\Delta x_{s,i}\right\|$ also converges to 0 from the state equations of $x_{s,i}$. Thus,

$$
\Delta E(x,u) = \sum_{s=0}^{T-1} \left(\left\langle -2\omega_i G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right) \Delta u_{s,i}, \Delta u_{s,i} \right\rangle + \mathcal{O}\left(\Delta u_{s,i}\right) \left\|\Delta u_{s,i}\right\|\right).
$$

Now scale all $\{\Delta u_{s,i}\}_{s=0,\ldots,T-1,i=1,\ldots,N}$ with a scalar $0 < \alpha \leq 1$. Note that the term $-2\omega_i G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right) \Delta u_{s,i}$ is unaffected by the scaling, since this term is equal to $\mu_{s,i} + \nabla_{u_{s,i}} E(x,u)\big|_{(x,u)=\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)}$, which is independent of α. The first order Taylor approximation is then

$$
\Delta E(x,u) = \sum_{s=0}^{T-1} \left(\left\langle -2\omega_i G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right) \Delta u_{s,i}, \alpha \Delta u_{s,i} \right\rangle + \mathcal{O}\left(\alpha \Delta u_{s,i}\right) \left\|\alpha \Delta u_{s,i}\right\|\right).
$$

In the limit it follows that there exists a scalar $\eta > 0$ such that $\frac{\Delta E(x,u)}{\alpha} < 0$ for some $0 < \alpha \leq \eta \leq 1$, i.e. $\Delta E(x,u) < 0$. From Lemma 12 it follows that $\left(x_{s,i}^{(j+1)}(\alpha), u_{s,i}^{(j+1)}(\alpha)\right)$ is a feasible solution. □

Proposition 14 *Let $E^{(i)}$ denote the discretized energy for the solution after iteration i (including line search) in GEORCE-FM. If the starting point, $\left(x^{(0)}, u^{(0)}\right)$ is feasible, then the series $\left\{E^{(i)}\right\}_i$ will converge to a local minimum.*

Proof If the starting point, $\left(x^{(0)}, u^{(0)}\right)$ in *GEORCE-FM* is feasible, then the points, $\left(x^{(j)}, u^{(j)}\right)$ after each iteration (including line search) in *GEORCE-FM* is also a feasible solution by Lemma 12.

The discretized energy functional is a positive function, i.e. a lower bound is higher or equal to 0. Assume that the series $\left\{E^{(j)}\right\}_j$ is not converging to a local minimum. From

Lemma 13 it follows that $\left\{E^{(j)}\right\}_j$ is decreasing for increasing j, and since $E^{(j)}$ has a lower bound, the series $E^{(j)}$ can not be non-converging. Assume on the contrary that the convergence value for $\left\{E^{(j)}\right\}_j$, is not a local minimum. Denote the value $\hat{E}$ and the convergence point $(\hat{x}, \hat{u})$. According to Lemma 13 *GEORCE-FM* will from the solution $(\hat{x}, \hat{u})$ in the following iteration produce a new point $(\tilde{x}, \tilde{u})$ such that $E(\tilde{x}, \tilde{u}) < \hat{F}$, which contradicts that the series $\left\{E^{(j)}\right\}_j$ is converging to $\hat{F}$. Thus, the series $\left\{E^{(j)}\right\}_j$ will converge to a local minimum. □

Note that by the same argument as in [25], then *GEORCE-FM* will not be able to jump between two local minima and will converge to only one local minimum.

A.7 Local convergence

Before proving local convergence we re-state the following Lemma from [25]

Lemma 15 ([25]) *Assume $f(x)$ is a smooth function. Let z^* be a (local) minimum point for f, and assume that f is locally strictly convex in z^*. Then*

$$\exists \epsilon > 0 : \quad \forall z \in B_\epsilon(z^*) \setminus \{z^*\} : \quad \langle \nabla f(z), z^* - z \rangle < 0.$$

In the following, we generalize the proof in [25] to the case of the Fréchet mean.

Proposition 16 *If the discretized energy functions is locally strictly convex in the (local) minimum point $z^* = (x^*, u^*)$ and locally $\alpha^* = 1$, i.e. no line-search, then GEORCE-FM has locally quadratic convergence, i.e.*

$$\exists \epsilon > 0 \, \exists c > 0 \, \forall z^{(j)} \in B_\epsilon\left(z^*\right) : \quad \left\| z^{(j+1)} - z^* \right\| \leq c \left\| z^{(j)} - z^* \right\|^2,$$

where $B_\epsilon\left(z^*\right)$ is an open ball with radius ϵ around z^* and $z^{(j)} = \left\{\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)\right\}_{s,i}$ is the i'th iteration of GEORCE-FM.

Proof Assume z^* is a local minimum point. Since *GEORCE-FM* has global convergence to a local minimum point by Proposition 14 we assume that the algorithm has convergence point $E\left(z^*\right)$. Consider the solution $z^{(j)} = \left(x^{(j)}, u^{(j)}\right)$ from *GEORCE-FM* and assume that for all the following iterations in *GEORCE-FM* the solution is in the open ball $B_\epsilon\left(z^*\right)$, and the locally strictly convex assumption in Assumptions 9 holds. Since the series, $\left\{z^{(j)}\right\}$, converge to z^*, then from a certain point k in the series all point $\left\{z^{(j)}\right\}_{j \leq k}$ will belong to the open ball $B_\epsilon\left(z^*\right)$.

It then follows that for some $K > 0$

$$E\left(z^{(j+1)}\right) - E\left(z^*\right) = \left\| z^{(j+1)} - z^* \right\|$$

since $z^{(j+1)} \in B_\epsilon\left(z^*\right)$. The left hand side can be rearranged into

$$E\left(z^{(j+1)}\right) - E\left(z^*\right) = \left(E\left(z^{(j+1)}\right) - E\left(z^{(j)}\right)\right) + \left(E\left(z^{(j)}\right) - E\left(z^*\right)\right).$$

By the proof of Lemma 13 we see

$$
\begin{aligned}
&E\left(z^{(j+1)}\right)-E\left(z^{*}\right)=\left(E\left(z^{(j+1)}\right)-E\left(z^{(j)}\right)\right)+\left(E\left(z^{(j)}\right)-E\left(z^{*}\right)\right)\\
&=\sum_{s=0}^{T-1}\sum_{i=0}^{N}\left(\left\langle -2\omega_i G\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\Delta u_{s,i},\Delta u_{s,i}\right\rangle+\mathcal{O}\left(\Delta u_{s,i}\right)\left\|\Delta u_{s,i}\right\|\right)\\
&-\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle -2\omega_i G\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\Delta u_{s,i},u_{s,i}^{*}-u_{s,i}^{(j)}\right\rangle+\mathcal{O}\left(u_{s,i}^{*}-u_{s,i}^{(j)}\right)\left\|u_{s,i}^{*}-u_{s,i}^{(j)}\right\|\right)\\
&=\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle -2\omega_i G\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\Delta u_{s,i},\Delta u_{s,i}-\left(u_{s,i}^{*}-u_{s,i}^{(j)}\right)\right\rangle\right)\\
&+\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\mathcal{O}\left(\Delta u_{s,i}\right)\left\|\Delta u_{s,i}\right\|+\mathcal{O}\left(u_{s,i}^{*}-u_{s,i}^{(j)}\right)\left\|u_{s,i}^{*}-u_{s,i}^{(j)}\right\|\right)\\
&=\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle -2\omega_i G\left(ts,i^{(j)},x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\Delta u_{s,i},u_{s,i}^{*}-u_{s,i}^{(j+1)}\right\rangle\right)\\
&+\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\mathcal{O}\left(\Delta u_{s,i}\right)\left\|\Delta u_{s,i}\right\|+\mathcal{O}\left(u_{s,i}^{*}-u_{s,i}^{(j)}\right)\left\|u_{s,i}^{*}-u_{s,i}^{(j)}\right\|\right).
\end{aligned}
$$

Observe that the term $2\omega_i G\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)$ is the gradient for $z_s^{(j)}$, which by the assumption of strongly unique minimum point in Assumptions 9 implies that

$$
\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left\langle 2\omega_i G\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\Delta u_{s,i},u_{s,i}^{*}-u_{s,i}^{(j+1)}\right\rangle<0.
$$

Combining the inequalities and utilizing that $E(z)$ is locally Lipschitz continuous by the Assumptions 9, then

$$
\mathcal{O}\left(\left\|z^{(j+1)}-z^{*}\right\|\right)=E\left(z^{(j+1)}\right)-E\left(z^{*}\right)\leq\mathcal{O}\left(\left\|z^{(j)}-z^{*}\right\|^{2}\right),
$$

such that there exists an $\epsilon>0$

$$
\forall z^{(j)}\in B_{\epsilon}\left(z^{*}\right):\quad\left\|z^{(j+1)}-z^{*}\right\|\leq c\left\|z^{(j)}-z^{*}\right\|^{2},
$$

where $c>0$. □

A.8 Adaptive convergence

Consider the general minimization for the energy functional for the Fréchet mean

$$
\min_z E(z)=\min_z \mathbb{E}_{k\in I}\left[E_k(z)\right],
$$

where I is a finite index set of possible occurences, and $E_k(z)$ is the energy functional related to occurence k. The metric tensor related to occurence k is denoted $G_k(z_k)$ and is assumed positive definite. It follows that $\mathbb{E}_{k\in I}\left[G_k(z_{s,i})\right]=G\left(z_{s,i}\right)$.

Lemma 17 *Let $z_{s,i}^{(j)} = \left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)$ be the feasible solution after iteration j in GEORCE-FM. If $z_{s,i}^{(j)}$ is not a local minimum point, then the feasible solution from the following iteration $z_{s,i}^{(j+1)}(k)$ for occurence k will decrease the objective function in the sense that there exists an $\eta > 0$ such that for all $0 < \alpha \leq \eta \leq 1$, then*

$$\mathbb{E}_{k\in I}\left[E_k\left(x^{(j)} + \alpha\left(x^{(j+1)}(k) - x^{(j)}\right), u^{(j)} + \alpha\left(u^{(j+1)}(k) - u^{(j)}\right)\right)\right] - E\left(x^{(j)}, u^{(j)}\right] < 0, \tag{36}$$

$$\mathbb{E}_{k\in I}\left[\Delta E_k\left(x^{(j)}, u^{(j)}\right)\right] \approx \sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle -2\omega_i G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)\left(\tilde{u}_{s,i}^{(j+1)} - u_{s,i}^{(j)}\right), \alpha\left(\tilde{u}_{s,i}^{(j+1)} - u_{s,i}^{(j)}\right)\right\rangle\right)$$
$$+ \sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\mathbb{E}_{k\in I}\left[\left\langle -2\omega_i G_k\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)\epsilon_{s,i}(k), \alpha\epsilon_{s,i}(k)\right\rangle\right]\right), \tag{37}$$

where $x^{(j)}[i] = \left(a_i, x_{s,i}^{(j)}, \dots, x_{T-1}^{(j)}, y\right)$, $u^{(j)}[i] = \left(u_{0,i}^{(j)}, u_{1,i}^{(j)}, \dots, u_{T-2,i}^{(j)}, u_{T-1,i}^{(j)}\right)$ and $\epsilon_{s,i}(k) = u_{s,i}^{(j+1)}(k) - \mathbb{E}_{k\in I}\left[u_{s,i}^{(j+1)}(k)\right]$ and $\tilde{u}_{s,i}^{(j+1)} = \mathbb{E}_{k\in I}\left[u_{s,i}^{(j+1)}(k)\right]$. Further, if the index set fulfills that

$$\sum_{s=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}_{k\in I}\left[\left(u_{s,i}^{(j+1)}(k) - u_{s,i}^{(j)}\right)^\top \omega_i G\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)\left(\hat{u}_{s,i}^{(j+1)} - u_{s,i}^{(j)}\right)\right] > 0, \tag{38}$$

where $u_{s,i}^{(j)}$ is not a (local) minimum point, and $\hat{u}_{s,i}^{(j+1)}$ is the GEORCE-FM iteration point, with no batching. If this property holds, then there exists an $\eta > 0$ such that for all α with $0 < \alpha \leq \eta \leq 1$, then

$$E\left(x^{(j)} + \alpha\left(\mathbb{E}_{k\in I}\left[x^{(j+1)}(k)\right] - x^{(j)}\right), u^{(i)} + \alpha\left(\mathbb{E}_{k\in I}\left[u^{(j+1)}(k)\right] - u^{(j)}\right)\right) - E\left(x^j, u^j\right) < 0. \tag{39}$$

Proof Assume that $z_{s,i}^{(j)} = \left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)$ is a feasible solution. For the occurrence k then the change in the objective function after a *GEORCE-FM* step with step size, α, is by the proof of Lemma 13

$$\begin{aligned}\Delta E_k\left(x^{(j)}, u^{(j)}\right) = &\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left\langle -2\omega_i G_k\left(x_{s,i}^{(j)}, u_{s,i}^{(j)}\right)\Delta u_{s,i}^{(k)}, \alpha\Delta u_{s,i}(k)\right\rangle \\ &+ \sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\sum_{l=0}^{s-1}\mathcal{O}\left(\alpha\Delta u_{l,i}(k)\right)\left\|\sum_{l=0}^{s-1}\alpha\Delta u_{l,i}(k)\right\|\right) \\ &+ \sum_{i=1}^{N}\mathcal{O}\left(\alpha\Delta u_{s,i}(k)\right)\left\|\alpha\Delta u_{s,i}(k)\right\|\end{aligned}$$

Thus, there exists an $\eta_k > 0$ such that for all $0 < \alpha < \eta_k$, then $\frac{\Delta E_k\left(x^{(j)}, u^{(j)}\right)}{\alpha} < 0$, which implies that $\mathbb{E}_{k\in I}\left[\Delta E_k\left(x^{(j)}, u^{(j)}\right)\right] < 0$. This proves eq. 36.

Define $\epsilon_{s,i} = u_{s,i}^{(j+1)}(k) - \mathbb{E}_{k\in I}\left[u_{s,i}^{(j+1)}(k)\right]$. Then $\mathbb{E}_{k\in I}\left[\epsilon_{s,i}(k)\right] = 0$. To simplify notation, set $\tilde{u}_{s,i}^{(j+1)} := \mathbb{E}_{k\in I}\left[u_{s,i}^{(j+1)}\right]$. Applying this to the incremental change for *GEORCE-FM* we see that

$$\begin{aligned}
&\mathbb{E}_{k\in I}\left[\Delta E_k\left(x^{(j)},u^{(j)}\right)\right]\\
&=\sum_{s=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}_{k\in I}\left[\left\langle -2\omega_i G_k\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\left(\tilde{u}^{(j+1)}+\epsilon_{s,i}(k)-u_{s,i}^{(j)}\right),\alpha\left(\tilde{u}_{s,i}^{(j+1)}+\epsilon_{s,i}(k)-u_{s,i}^{(j)}\right)\right\rangle\right]\\
&+\sum_{s=0}^{T-1}\mathbb{E}_{k\in I}\left[\mathcal{O}\left(\sum_{l=0}^{s-1}\sum_{i=1}^{N}\alpha\Delta u_{l,k}(k)\right)\left\|\sum_{l=0}^{s-1}\sum_{i=1}^{N}\alpha\Delta u_{l,i}(k)\right\|\right]\\
&+\sum_{s=0}^{T-1}\sum_{i=1}^{N}\mathcal{O}\left(\alpha\Delta u_{s,i}(k)\right)\left\|\alpha\Delta u_{s,i}(k)\right\|\\
&=\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left(\left\langle -2\omega_i G\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\left(\tilde{u}_{s,i}^{(j+1)}-u_{s,i}^{(j)}\right),\alpha\left(\tilde{u}^{(j+1)}-u_{s,i}^{(j)}\right)\right\rangle+\left\langle -2G_k\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\epsilon_{s,i}(k),\alpha\epsilon_{s,i}(k)\right\rangle\right)\\
&+\sum_{s=0}^{T-1}\mathbb{E}_{k\in I}\left[\mathcal{O}\left(\sum_{l=0}^{s-1}\sum_{i=1}^{N}\alpha\Delta u_{l,i}(k)\right)\left\|\sum_{l=0}^{s-1}\sum_{i=1}^{N}\alpha\Delta u_{l,i}(k)\right\|\right]\\
&+\sum_{s=0}^{T-1}\mathbb{E}_{k\in I}\left[\sum_{i=1}^{N}\mathcal{O}\left(\alpha\Delta u_{s,i}(k)\right)\left\|\alpha\Delta u_{s,i}(k)\right\|\right],
\end{aligned}$$

which proves eq. 37. The incremental change in the energy functional applying a *GEORCE-FM* step is by the proof of Lemma 13

$$\begin{aligned}
\Delta E\left(x^{(j)},u^{(j)}\right)&=\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left\langle -2\omega_i G\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\Delta u_{s,i},\alpha\Delta u_{s,i}\right\rangle\\
&+\sum_{s=0}^{T-1}\sum_{i=1}^{N}\mathcal{O}\left(\sum_{l=0}^{s-1}\alpha u_{l,i}\right)\left\|\sum_{l=0}^{s-1}\alpha\Delta u_{l,i}\right\|\\
&+\sum_{s=0}^{T-1}\sum_{i=1}^{N}\mathcal{O}\left(\alpha\Delta u_{s,i}\right)\left\|\alpha\Delta u_{s,i}\right\|.
\end{aligned}$$

Suppose the step is instead $\tilde{u}_{s,i}^{(j+1)}-u_{s,i}^{(j)}$, then the incremental energy function becomes

$$\begin{aligned}
\Delta E\left(x^{(j)},u^{(j)}\right)&=\sum_{s=0}^{T-1}\sum_{i=1}^{N}\left\langle -2\omega_i G\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\Delta u_{s,i},\alpha\left(\tilde{u}_{s,i}^{(j+1)}-u_{s,i}^{(j)}\right)\right\rangle\\
&+\sum_{s=0}^{T-1}\mathcal{O}\left(\sum_{l=0}^{s-1}\sum_{i=1}^{N}\alpha\left(\tilde{u}_{l,i}^{(j+1)}-u_{l,i}^{(j)}\right)\right)\left\|\sum_{l=0}^{s-1}\sum_{i=1}^{N}\alpha\left(\tilde{u}_{l,i}^{(j+1)}-u_{l,i}^{(j)}\right)\right\|\\
&+\sum_{s=0}^{T-1}\sum_{i=1}^{N}\mathcal{O}\left(\alpha\left(\tilde{u}^{(j+1)}-u_{s,i}^{(j)}\right)\right)\left\|\alpha\left(\tilde{u}_{s,i}^{(j+1)}-u_{s,i}^{(j)}\right)\right\|
\end{aligned}$$

Using the property in eq. 38, then

$$\begin{aligned}
&\sum_{s=0}^{T-1}\sum_{i=0}^{N}\left\langle -2\omega_i G\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\Delta u_{s,i},\alpha\left(\tilde{u}_{s,i}^{(j+1)}-u_{s,i}^{(j)}\right)\right\rangle\\
&=\sum_{s=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}_{k\in I}\left[\left\langle -2\omega_i G\left(x_{s,i}^{(j)},u_{s,i}^{(j)}\right)\left(\hat{u}_{s,i}^{(j+1)}-u_{s,i}^{(j)}\right),\alpha\left(u_{s,i}^{(j+1)}(k)-u_{s,i}^{(j)}\right)\right\rangle\right]\\
&<0.
\end{aligned}$$

Since the quadratic forms under the summation are negative, then along the same lines as above, there exists an $\eta > 0$ such that for all $0 < \alpha < \eta\Delta$ then $E\left(x^{(j)}, u^{(j)}\right) < 0$ with step $\tilde{u}_{s,i}^{(j+1)} - u_{s,i}^{(j)}$, which proves eq. 39. $\square$

Note that eq. 37 states that the average incremental change in $\mathbb{E}_{k\in I}\left[\Delta E_k\left(x^{(j)}, u^{(j)}\right)\right]$ in *GEORCE-FM* steps over the index set I has the same structure as for the energy functional, but the *GEORCE-FM* step for the energy functional is replaced by the average *GEORCE-FM* steps over the index set I- and an additional variance term.

Eq. 38 is trivially fulfilled if the index set contains all data points. Eq. 38 ensures that the steps generated by *GEORCE-FM* for index k on average over index set I "points in the same direction" as the *GEORCE-FM* step for the energy functional, i.e. that the scalar product between the average step and the *GEORCE-FM* step for the energy fucntional is positive wrt. the metric tensor G.

Proposition 18 *Assume Lemma 17 is fulfilled and eq. 38 holds. Then the series* $\left\{\mathbb{E}_{k\in I}\left[z^{(j)}(k)\right]\right\}_j$ *will converge to a local minimum point in expectation.*

Proof Since the energy functional has a lower bound of 0, and the energy functional $\left\{E\left(\mathbb{E}_{k\in I}\left[z^{(j)}(k)\right]\right)\right\}_j$ is a decreasing series by eq. 39, then the series will converge. Assume that the series is converging to a point $z^{(c)}$ that is not a local minimum. Then there exists a *GEORCE-FM* step $\Delta z^{(c)}(k)$ for the energy functionals F_k at the point $z^{(c)}$ such that $\mathbb{E}_{k\in I}\left[E_k\left(z^{(c)} + \alpha_1 \Delta z^{(c)}(k)\right)\right] < E\left(z^{(c)}\right)$ by eq. 36. It follows from eq. 39 that $E\left(z^{(c)} + \alpha_2 \mathbb{E}_{k\in I}\left[\Delta z^{(c)}(k)\right]\right) < E\left(z^{(c)}\right)$, which contradicts that $z^{(c)}$ is not a local minimum. $\square$

A.9 Fréchet mean for Finsler manifolds

Proposition 19 *The update scheme for* u_t, μ_t *and* x_t *is*

$$\begin{aligned}
&y = W^{-1}V, \\
&\mu_{T-1,i} = \left(\sum_{t=0}^{T-1} \tilde{G}_{t,i}^{-1}\right)^{-1}\left(2w_i(a_i - y) - \sum_{t=0}^{T-1} \tilde{G}_{t,i}^{-1}\left(\zeta_{t,i} + \sum_{t>j}^{T-1} \nu_{j,i}\right)\right), \quad i = 1, \dots, N, \\
&u_{t,i} = -\frac{1}{2w_i}\tilde{G}_{t,i}^{-1}\left(\mu_{T-1,i} + \zeta_{t,i} + \sum_{j>t}^{T-1} \nu_{j,i}\right), \quad t = 0, \dots, T-1, \, i = 1, \dots, N \\
&x_{t+1,i} = x_{t,i} + u_{t,i}, \quad t = 0, \dots, T-2, \, i = 1, \dots, N, \\
&x_{0,i} = a_i \quad i = 1, \dots, N,
\end{aligned} \tag{40}$$

where

$$W = \sum_{i=1}^{N} w_i \left(\sum_{t=0}^{T-1} \tilde{G}_{t,i}^{-1} \right)^{-1},$$

$$V = \sum_{i=1}^{N} w_i \left(\sum_{t=0}^{T-1} \tilde{G}_{ti}^{-1} \right)^{-1} a_i - \frac{1}{2} \sum_{i=1}^{N} \left(\sum_{t=0}^{T-1} \tilde{G}_{t,i}^{-1} \right)^{-1} \sum_{t=0}^{T-1} \tilde{G_{ti}}^{-1} \left(\zeta_{t,i} + \sum_{j>t}^{T-1} \nu_{j,i} \right).$$

Here $\nu_{t,i} := \nabla_y u_{t,i}^\top \tilde{G}(y, u_{t,i}) u_{t,i} \Big|_{y=x_{t,i}}$ *and* $\zeta_{t,i} := \nabla_v u_{t,i}^\top \tilde{G}(x_{t,i}, v) u_{t,i} \Big|_{v=u_{t,i}}$.

Proof Consider the optimization problem

$$\min_{(x_{t,i},u_{t,i})} E(x) := \min_{(x_{t,i},u_{t,i})} \left\{ \sum_{i=1}^{N} w_i \sum_{t=0}^{T-1} u_{t,i}^T G(x_{t,i}, u_{t,i}) u_{t,i} \right\}$$
$$x_{t+1,i} = x_{t,i} + u_{t,i}, \quad t = 0, \ldots, T-1,\ i = 1, \ldots, N,$$
$$x_{0,i} = a_i, x_{T,i} = y, \quad i = 1, \ldots, N.$$

For iteration k we consider the following modified zero point problem derived using the same principle as in the Riemannian case.

$$\begin{aligned}
&\nu_{t,i} + \mu_{t,i} = \mu_{t-1,i}, \quad t = 1, \ldots, T-1,\ i = 1, \ldots, N,\\
&2w_i G_{t,i} u_{t,i} + \zeta_{t,i} + \mu_{t,i} = 0, \quad t = 0, \ldots, T-1,\ i = 1, \ldots, N,\\
&\sum_{t=0}^{T-1} u_{t,i} = y - a_i, \quad i = 1, \ldots, N,\\
&\sum_{i=1}^{N} \mu_{T-1,i} = 0,\\
&\nu_{t,i} := \nabla_x \left(w_i u_{t,i}^\top G(x) u_{t,i} \right) \Big|_{x=x_{t,i}^{(k)}, u_{t,i}=u_{t,i}^{(k)}}, \quad t = 1, \ldots, T-1,\ i = 1, \ldots, N\\
&\zeta_{t,i} := \nabla_u \left(w_i v^\top G(x, u) v \right) \Big|_{y=x_{t,i}^{(k)}, v=u_{t,i}^{(k)}, u=u_{t,i}^{(k)}}, \quad t = 1, \ldots, T-1,\ i = 1, \ldots, N\\
&G_{t,i} := G\left(x_{t,i}^{(k)} \right), \quad t = 0, \ldots, T-1.
\end{aligned} \tag{41}$$

From eq. 41 we get the following iterative scheme for updating y and $\{x_{t,i}\}_{t,i}$ By re-arranging the term we see that

$$\mu_{t,i} = \mu_{T-1,i} + \sum_{j>t}^{T-1} g_{jt}, \quad t = 0, \ldots, T-1, i = 1, \ldots, N,$$

$$u_{t,i} = -\frac{1}{2} G_{t,i}^{-1} \mu_{t,i} = -\frac{1}{2w_i} G_{t,i}^{-1} \left(\mu_{T-1,i} + \zeta_{t,i} + \sum_{j>t}^{T-1} \nu_{j,i} \right), \quad t = 0, \ldots, T-1, i = 1, \ldots, N,$$

$$-\frac{1}{2\omega_i} \sum_{t=0}^{T-1} G_{t,i}^{-1} \left(\mu_{T-1,i} + \zeta_{t,i} + \sum_{j>t}^{T-1} g_{ji} \right) = y - a_i, \quad i = 1, \ldots, N,$$

where it is exploited that $G_{t,i}$ is positive definite for all t, i and has a well defined inverse. These equations can be used to provide explicit solutions for the state and control variables. First observe that

$$\mu_{T-1,i} = \left(\sum_{t=0}^{T-1} G_{t,i}^{-1}\right)^{-1} \left(2w_i(a_i - y) - \sum_{t=0}^{T-1} G_{t,i}^{-1} \left(\zeta_{t,i} + \sum_{j>t}^{T-1} \nu_{j,i}\right)\right), \quad i = 1, \dots, N.$$

Applying this result, we deduce that

$$\sum_{i=1}^{N} \mu_{T-1,i} = 0 \Leftrightarrow \sum_{i=1}^{N} \left(\sum_{t=0}^{T-1} G_{t,i}^{-1}\right)^{-1} \left(2w_i(a_i - y) - \sum_{t=0}^{T-1} G_{t,i}^{-1} \left(\zeta_{t,i} + \sum_{j>t}^{T-1} \nu_{j,i}\right)\right) = 0,$$

which implies that

$$y = W^{-1}V,$$

where

$$W = \sum_{i=1}^{N} w_i \left(\sum_{t=0}^{T-1} G_{t,i}^{-1}\right)^{-1},$$

$$V = \sum_{i=1}^{N} w_i \left(\sum_{t=0}^{T-1} G_{ti}^{-1}\right)^{-1} a_i - \frac{1}{2} \sum_{i=1}^{N} \left(\sum_{t=0}^{T-1} G_{t,i}^{-1}\right)^{-1} \sum_{t=0}^{T-1} G_{ti}^{-1} \left(\zeta_{t,i} + \sum_{j>t}^{T-1} g_{ji}\right).$$

Thus, we get the update scheme in Proposition 6. □

B Algorithms

B.0.1 GEORCE

In algorithm 3 we re-state the *GEORCE* algorithm for Riemannian manifolds, and in algorithm 4 we re-state the *GEORCE* algorithm for Finsler manifolds from [25].

B.0.2 GEORCE-FM for Finsler

C Manifolds

In this section we give a brief background to the different manifolds used in the experiments in the paper. The manifolds are the same that can be found in [25], and we re-state the definitions here.

C.1 Riemannian Manifolds

The Riemannian manifolds used in the paper are the same as in [25], and we re-state the definitions from [25] in Table 4.

C.2 Finsler Manifolds

Let $\mathcal{M}$ be a differentiable manifold equipped with a Riemannian metric. Consider a force field $f : \mathcal{M} \to T_x\mathcal{M}$ acting on the manifold. If a microswimmer wants to find the shortest curve on the manifold, $\mathcal{M}$, under influence of f with constrant velocity v,

Algorithm 3 GEORCE for Riemannian Manifolds

Input: tol, T
Output: Geodesic estimate $x_{0:T}$
Set $x_t^{(0)} \leftarrow a + \frac{b-a}{T} t$, $u_t^{(0)} \leftarrow \frac{b-a}{T}$ for $t = 0., \dots, T$ and $i \leftarrow 0$
while do $\left\| \nabla_y E(y) \big|_{y = x_t^{(i)}} \right\|_2 > \text{tol}$
  $G_t \leftarrow G\left(x_t^{(i)}\right)$ for $t = 0, \dots, T-1$

  $\nu_t \leftarrow \nabla_y \left(u_t^{(i)} G(y) \, u_t^{(i)} \right) \Big|_{y = x_t^{(i)}}$ for $t =, 1 \dots, T-1$
  $\mu_{T-1} \leftarrow \left(\sum_{t=0}^{T-1} G_t^{-1} \right)^{-1} \left(2(a-b) - \sum_{t=0}^{T-1} G_t^{-1} \sum_{t>j}^{T-1} \nu_j \right)$
  $u_t \leftarrow -\frac{1}{2} G_t^{-1} \left(\mu_{T-1} + \sum_{j>t}^{T-1} \nu_j \right)$ for $t = 0, \dots, T-1$
  $x_{t+1} \leftarrow x_t + u_t$ for $t = 0, \dots, T-1$
  Using line search find α^* for the following optimization problem with the discrete energy functional E

$$
\begin{aligned}
\alpha^* = \arg\min_{\alpha} \quad & E\left(x_{0:T}\right) \quad \text{(exact line search)} \\
\text{s.t.} \quad & x_{t+1} = x_t + \alpha u_t + (1-\alpha) u_t^{(i)}, \quad t = 0, \dots, T-1, \\
& x_0 = a.
\end{aligned}
$$

  Set $u_t^{(i+1)} \leftarrow \alpha^* u_t + (1-\alpha^*) u_t^{(i)}$ for $t = 0, \dots, T-1$
  Set $x_{t+1}^{(i+1)} \leftarrow x_t^{(i+1)} + u_t^{(i+1)}$ for $t = 0, \dots, T-1$
  $i \leftarrow i + 1$
end while
return x_t for $t = 0, \dots, T-1$

then this corresponds to finding geodesics on a Randers metric [53], which is a special case of a Finsler manifold, where

$$
\begin{aligned}
F &= \sqrt{a_{ij} \dot{r}^i \dot{r}^j} + b_i \dot{r}^i \\
a_{ij} &= g_{ij} \lambda + f_i f_j \lambda^2 \\
b_i &= -f_i \lambda \\
f_i &= g_{ij} f^j \\
\lambda^{-1} &= v_0^2 - g_{ij} f^i f^j,
\end{aligned}
\tag{42}
$$

where g_{ij} denotes the elements of the Riemannian background metric, f^j denotes the jth element of the force field, while v_0 denotes the initial velocity Riemannian length of the microswimmer with constant velocity v on the surface. The Finsler metric is well defined if and only if $||f||_g < v_0$ [53].

Algorithm 4 GEORCE for Finsler Manifolds

Input: tol, T

Output: Geodesic estimate $x_{0:T}$

Set $x_t^{(0)} \leftarrow a + \frac{b-a}{T}t$, $u_t^{(0)} \leftarrow \frac{b-a}{T}$ for $t = 0., \dots, T$ and $i \leftarrow 0$

while do $\left\| \nabla_y E(y) \big|_{y=x_t^{(i)}} \right\|_2 > \mathrm{tol}$

$\quad G_t \leftarrow G\left(x_t^{(i)}, u_t^{(i)}\right)$ for $t = 0, \dots, T-1$

$\quad \nu_t \leftarrow \nabla_y \left(u_t^{(i)} G\left(y, u_t^{(i)}\right) u_t^{(i)}\right)\Big|_{y=x_t^{(i)}}$ for $t = 1, \dots, T-1$

$\quad h_t \leftarrow \nabla_y \left(u_t^{(i)} G\left(x_t^{(i)}, y\right) u_t^{(i)}\right)\Big|_{y=u_t^{(i)}}$ for $t = 1, \dots, T-1$

$\quad \mu_{T-1} \leftarrow \left(\sum_{t=0}^{T-1} G_t^{-1}\right)^{-1} \left(2(a-b) - \sum_{t=0}^{T-1} G_t^{-1}\left(h_t + \sum_{t>j}^{T-1} \nu_j\right)\right)$

$\quad u_t \leftarrow -\frac{1}{2} G_t^{-1}\left(\mu_{T-1} + h_t + \sum_{j>t}^{T-1} \nu_j\right)$ for $t = 0, \dots, T-1$

$\quad x_{t+1} \leftarrow x_t + u_t$ for $t = 0, \dots, T-1$

Using line search find α^* for the following optimization problem with the discrete energy functional E_F

$$
\begin{aligned}
\alpha^* = \arg\min_{\alpha} \quad & E_F\left(x_{0:T}\right) \quad \text{(exact line-search)} \\
\text{s.t.} \quad & x_{t+1} = x_t + \alpha u_t + (1-\alpha) u_t^{(i)}, \quad t = 0, \dots, T-1, \\
& x_0 = a.
\end{aligned}
$$

$\quad$ Set $u_t^{(i+1)} \leftarrow \alpha^* u_t + (1-\alpha^*) u_t^{(i)}$ for $t = 0, \dots, T-1$

$\quad$ Set $x_{t+1}^{(i+1)} \leftarrow x_t^{(i+1)} + u_t^{(i+1)}$ for $t = 0, \dots, T-1$

$\quad i \leftarrow i + 1$

end while

return x_t for $t = 0, \dots, T-1$

C.3 Variational-Autoencoders

The Variational-Autoencoders (VAE) [57] is generative model that learns the data manifold, where the manifold can be equipped with the pull-back metric. A variational-autoencoder consists of an encoder, $h_\Phi : \mathcal{X} \subset \mathbb{R}^D \to \mathcal{Z} \subset \mathbb{R}^d$ that encodes the data $x \in \mathcal{R}^D$ into a latent representation $z \in \mathcal{Z}$ with $d < D$ parameters Φ. The data is then decoded by $f_\theta : \mathcal{Z} \to \mathcal{X}$. The parameters, $\{\Phi, \theta\}$, are estimated by minimizing the Evidence Lower Bound (ELBO) [57]. The VAE can be seen as a Riemannian manifold with one chart with the Riemannian pull-back metric [41, 42]

$$G(z) = J_f^\top J_f,$$

where J_f denotes the Jacobian of f. We train a VAE for the MNIST data [58] and CelebA data [52]. We use the exact same architecture as in [25], which is a modification of the architecture in [41]. We re-state the architecture from [25] in Table 5 and Table 6.

Algorithm 5 GEORCE-FM for Finsler Manifolds

Input: tol, $a_{1:N,i}$, T.

Output: Geodesic estimate $x_{0:T}$.

Set $y^{(0)} \leftarrow a_0$, $x_{t,i}^{(0)} \leftarrow a_i + \frac{y^{(0)}-a_i}{T}t$ and $u_{t,i}^{(0)} \leftarrow \frac{y^{(0)}-a_i}{T}$ for $t = 0., \dots, T$ and $i = 1, \dots, N$.

while s **do**top criteria ¿ tol

$G_{t,i} \leftarrow G\left(x_{t,i}^{(k)}\right)$ for $t = 0, \dots, T-1$ and $i = 1, \dots, N$.

$g_{t,i} \leftarrow \nabla_x \left(u_{t,i}^{(k)} G\left(x\right) u_{t,i}^{(k)}\right)\Big|_{x=x_{t,i}^{(k)}}$ for $t = 1, \dots, T-1$ and $i = 1, \dots, N$.

$y \leftarrow W^{-1}V$ using Eq. 22.

$\mu_{T-1,i} \leftarrow \left(\sum_{t=0}^{T-1} G_{t,i}^{-1}\right)^{-1} \left(2w_i(a_i - y) - \sum_{t=0}^{T-1} G_{t,i}^{-1} \sum_{t>j}^{T-1} g_{j,i}\right)$ for $i = 1, \dots, N$ and $t = 1, \dots, T-1$.

$u_{t,i} \leftarrow -\frac{1}{2w_i} G_{t,i}^{-1} \left(\mu_{T-1,i} + \sum_{j>t}^{T-1} g_{j,i}\right)$ for $t = 0, \dots, T-1$ and $i = 1, \dots, N$.

$x_{t+1,i} \leftarrow x_{t,i} + u_{t,i}$ for $t = 0, \dots, T-2$ and $i = 1, \dots, N$.

Using line search find α^* for the following optimization problem for the discrete sum of energy E

$$
\begin{aligned}
\min_{\alpha} \quad & E\left(x_{0:T,1:N}, \tilde{u}_{0:T,1:N}\right) \quad \text{(exact line search)} \\
\text{s.t.} \quad & x_{t+1,i} = x_{t,i} + \alpha \tilde{u}_{t,i} + (1-\alpha) u_{t,i}^{(k)}, \quad t = 0, \dots, T-1,\ i = 1, \dots, N. \\
& \tilde{u}_{t,i} = \alpha u_{t,i} + (1-\alpha) u_{t,i}^{(k)}, \quad t = 0, \dots, T-1,\ i = 1, \dots, N. \\
& x_{0,i} = a_i.
\end{aligned}
$$

Set $u_{t,i}^{(k+1)} \leftarrow \alpha^* u_{t,i} + (1-\alpha^*) u_{t,i}^{(k)}$ for $t = 0, \dots, T-1$ and $i = 1, \dots, N$.

Set $x_{t+1,i}^{(k+1)} \leftarrow x_{t,i}^{(k+1)} + u_{t,i}^{(k+1)}$ for $t = 0, \dots, T-1$ and $i = 1, \dots, N$.

end while

return $x_{t,i}$ for $t = 0, \dots, T-1$ for $i = 1, \dots, N$.

D Data and methods

D.1 Data

We apply *GEORCE-FM* to the following real-world datasets

- MNIST data [58]: We train a VAE for MNIST, which consists of 28×28 images of handwritten digits between 0 and 9. The latent space is 8 dimensional. We use 60,000 images for training the VAE.
- CelebA data [52]: We train a VAE for CelebA dataset, which consists of 202,599 images of famous people of size $64 \times 64 \times 3$, where we use approximately 80% for training similar to [41].

In table 7 we summarize the data generation, which is done in local coordinates except for the VAE's.

D.2 Methods and Hyper-Parameters

In table 8 we provide a description of the methods used in the paper, and what the hyper-parameters have been set to.

E Experiments

In this section we provide details on hardware and additional experiments in the paper.

E.1 Hardware

All figures have been computed on a *HP* computer with Intel Core i9-11950H 2.6 GHz 8C, 15.6" FHD, 720P CAM, 32 GB (2×16GB) DDR4 3200 So-Dimm, Nvidia Quadro TI2004GB Discrete Graphics, 1TB PCle NVMe SSD, backlit Keyboard, 150W PSU, 8cell, W11Home 64 Advanced, 4YR Onsite NBD.

The runtime estimates and the training of the VAE's have been computed on a GPU for at most 24 hours with a maximum memory of 10 GB. The *GPU* consists of 4 nodes on a *Tesla V100*.

E.2 Additional experiments

E.2.1 Fréchet Mean for VAE

In Fig. 10 we show the estimated Fréchet mean and geodesics using 10 reconstructed with a VAE for the pull-back metric of the decoder using *GEORCE-FM*. In Fig. 11 and Fig. 12 we show a similar plot, but where the pull-back metric of the decoder is a background with the generic force field in eq. 24 acting on the manifold, and thus converting the navigation problem into finding geodesics under a Finsler metric as described in Appendix C.

E.2.2 Additional Runtime Estimates Riemannian Manifolds

Additional Runtime Estimates Finsler Manifolds

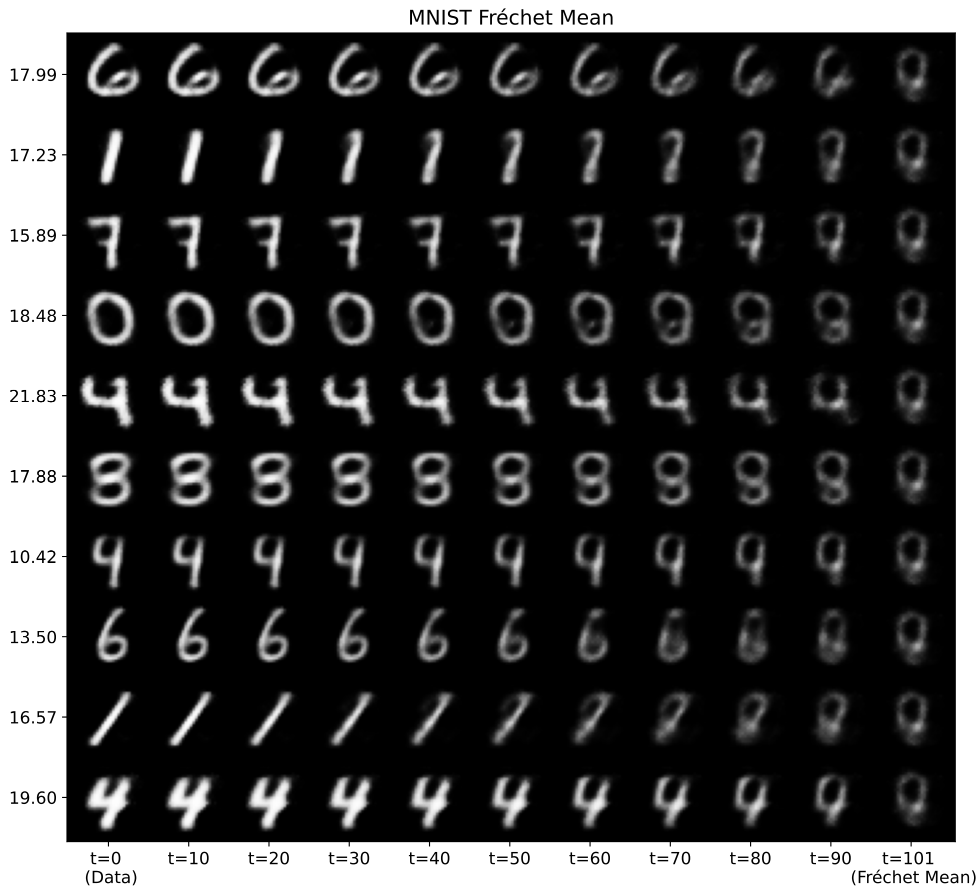

Fig. 10 The application of *GEORCE-FM* to manifold learning for a VAE estimating the Fréchet mean for 10 images of the MNIST dataset reconstructed using a VAE [58]. Each row shows 10 points on the geodesic, while the left most image is the data and the right most image is the estimated Fréchet mean using *GEORCE-FM*. The length of the estimated geodesic using *GEORCE-FM* is shown to the left, while the grid point number is shown at the bottom.

Manifold	Description	Parameters	Local Coordinates
$\mathbb{S}^n$	The n-sphere, $\{x \in \mathbb{R}^{n+1} \mid \|\|x\|\|_2 = 1\}$	-	Stereographic coordinates
$E(n)$	The n-Ellipsoid, $\{x \in \mathbb{R}^{n+1} \mid \|\|p \odot x\|\|_2 = 1\}$	Half-axes: $p = (0.5, \ldots, 1)$ equally distributed points.	Stereographic coordinates
$\mathbb{T}^2$	Torus	Major radius $R = 3.0$ and minor radius $r = 1.0$	Parameterized by (θ, ϕ) with $x = (R + r\cos\theta)\cos\phi, y = (R+r\cos\theta)\sin\phi, z = r\sin\theta$.
$\mathbb{H}^2$	2 dimensional Hyperbolic Space embedded into Minkowski pace $\mathbb{R}^3$	-	Parameterized by (α, β) with $x = \cosh\alpha, y = \sinh\alpha\cos\beta, z = \sinh\alpha\sin\beta$.
Paraboloid	Paraboloid of dimension n	-	Parameterized by standard coordinates by $\left(x^1, \ldots, x^n, \sum_{i=1}^n x_i^2\right)$
Hyperbolic Paraboloid	Hyperbolic Paraboloid of dimension 2	-	Parameterized by standard coordinates by $\left(x^1, \ldots, x^n, x_1^2 - x_2^2\right)$
$\mathcal{P}(n)$	Symmetric positive definite matrices of size n^2	-	Embedded into $\mathbb{R}^{n^2}$ by the mapping $f(x) = l(x)l(x)^T$, where $l : \mathbb{R}^{n(n+1)/2} \to \mathbb{R}^{n\times n}$ maps x into a lower triangle matrix consisting of the elements in x.
Gaussian Distribution	Parameters of the Gaussian distribution equipped with the Fischer-Rao metric	-	Parametrized by $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$.
Fréchet Distribution	Parameters of the Fréchet distribution equipped with the Fischer-Rao metric	-	Parametrized by $(\beta, \lambda) \in \mathbb{R}_+ \times \mathbb{R}_+$.
Cauchy Distribution	Parameters of the Cauchy distribution equipped with the Fischer-Rao metric	-	Parametrized by $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$.
Pareto Distribution	Parameters of the Pareto distribution equipped with the Fischer-Rao metric	-	Parametrized by $(\theta, \alpha) \in \mathbb{R}_+ \times \mathbb{R}_+$.
VAE MNIST	Variational-Autoencoder equipped with the pull-back metric for MNIST-data	-	Parametrized by the encoded latent space.
VAE CelebA	Variational-Autoencoder equipped with the pull-back metric for CelebA-data	-	Parametrized by the encoded latent space.

Table 4 Description of Riemannian Manifolds.

Encoder
Conv (output_channels $= 64$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Conv (output_channels $= 64$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Conv (output_channels $= 64$, kernel_shape $= 4 \times 4$, stride $= 1$, bias $=$ False)
GELU
Conv (output_channels $= 64$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Latent Parameters
μ: Linear(out_feature $= 8$, bias $=$ True), Identity
σ: Linear(out_feature $= 8$, bias $=$ True), Sigmoid
Decoder
Linear (out_feature $= 50$, bias $=$ True)
GELU
Conv2dTransposed (output_channels $= 64$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Conv2dTransposed (output_channels $= 32$, kernel_shape $= 4 \times 4$, stride $= 1$, bias $=$ False)
GELU
Conv2dTransposed (output_channels $= 16$, kernel_shape $= 4 \times 4$, stride $= 1$, bias $=$ False)
GELU
Linear (out_feature $= 784$, bias $=$ True)

Table 5 Architecture for MNIST

Encoder
Conv (output_channels $= 32$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Conv (output_channels $= 32$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Conv (output_channels $= 64$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Conv (output_channels $= 64$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Latent Parameters
μ: Linear(out_feature $= 32$, bias $=$ True), Identity
σ: Linear(out_feature $= 32$, bias $=$ True), Sigmoid
Decoder
Conv2dTransposed (output_channels $= 64$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Conv2dTransposed (output_channels $= 64$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Conv2dTransposed (output_channels $= 32$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Conv2dTransposed (output_channels $= 32$, kernel_shape $= 4 \times 4$, stride $= 2$, bias $=$ False)
GELU
Conv2dTransposed (output_channels $= 3$, kernel_shape $= 4 \times 4$, stride $= 4$, bias $=$ False)

Table 6 Architecture for CelebA

Manifold	**Data Generation**
$\mathbb{S}^n$	N normally distributed points with variance 1 and mean of n equally spaced points between 0 and 1
$E(n)$	N normally distributed points with variance 1 and mean of n equally spaced points between 0.5 and 1 (end point is excluded)
$\mathbb{T}^2$	N normally distributed points with variance 1 and mean $(0, 0)$
$\mathbb{H}^2$	N normally distributed points with variance 1 and mean $(0, 0)$
Paraboloid	N normally distributed points with variance 0.1 and mean $(1, 1)$
Hyperbolic Paraboloid	N normally distributed points with variance 0.1 and mean $(1, 1)$
$\mathcal{P}(n)$	N normally distributed points with variance 1.0 and mean $10I$, where I denotes the identity matrix.
Gaussian Distribution	$N/2$ normally with variance .1 and mean $(-1.0, 0.5)$ and $N/2$ normally with variance 0.1 and mean $(1.0, 1.0)$
Fréchet Distribution	$N/2$ normally with variance .1 and mean $(0.5, 0.5)$ and $N/2$ normally with variance 0.1 and mean $(1.0, 1.0)$
Cauchy Distribution	$N/2$ normally with variance .1 and mean $(-1.0, 0.5)$ and $N/2$ normally with variance 0.1 and mean $(1.0, 1.0)$
Pareto Distribution	$N/2$ normally with variance .1 and mean $(0.5, 0.5)$ and $N/2$ normally with variance 0.1 and mean $(1.0, 1.0)$
MNIST	N randomly selected images from the MNIST dataset.
CelebA	N randomly selected images from the CelebA dataset.

Table 7 The start and end point for the geodesics estimated for each manifold.

Method	**Description**	**Parameters**
GEORCE-FM	See algorithm in main paper	Backtracking: $\rho = 0.5$.
ADAM [20]	Moments based gradient descent method	$\alpha = 0.01$ (step size), $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$
SGD [51]	Stochastic gradient descent	$\gamma = 0.01$ (step size)
RMSprop Momentum	Stochastic gradient descent	$\alpha = 0.01$ (step size), momentum $= 0.9.$, $\epsilon = 10^{-8}$.
RMSprop [51]	Moments based gradient descent method	$\alpha = 0.01$ (step size), $\gamma = 0.9$, $\epsilon = 10^{-8}$
Adamax [20]	Corresponds to *ADAM* using the infinity norm in *ADAM*	$\alpha = 0.01$ (step size), $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$.
Adagrad [59]	Stochastic gradient method with adaptive updates of the step size	$\alpha = 0.01$ (step size), momentum $= 0.9$.

Table 8 The Hyper-parameters for estimating the Fréchet mean.

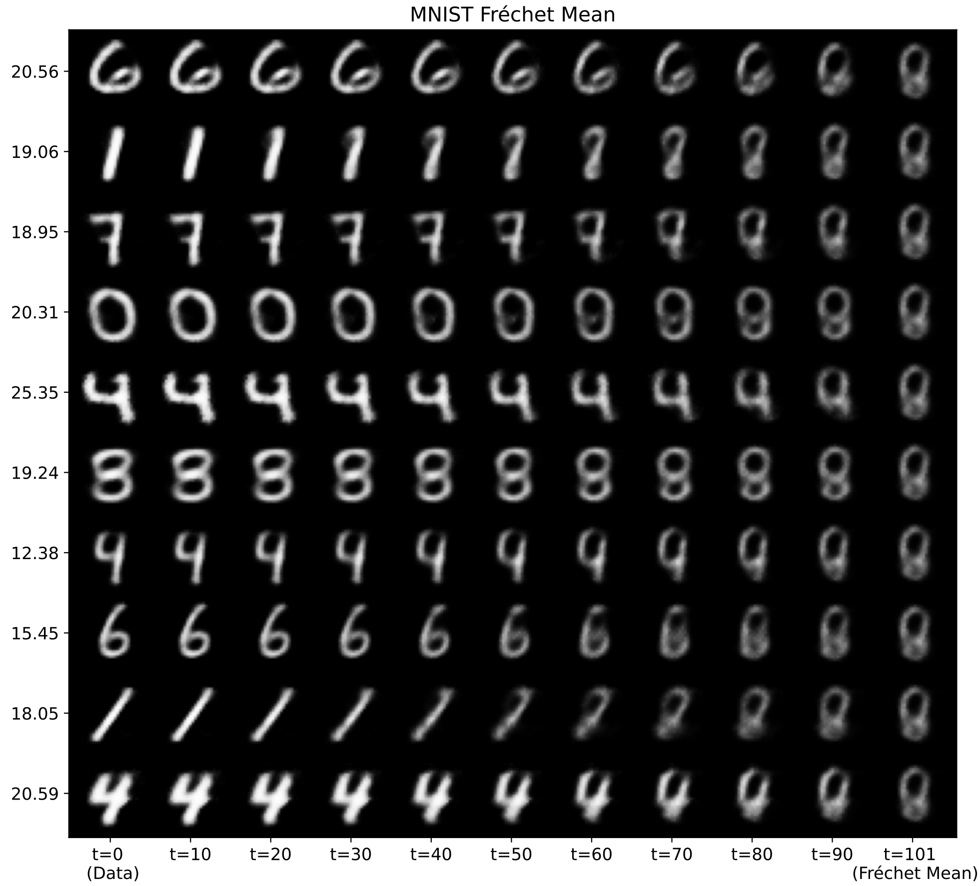

Fig. 11 The application of *GEORCE-FM* to manifold learning for a VAE estimating the Fréchet mean for 10 images of the MNIST dataset reconstructed using a VAE [58]. The learned manifold is equipped with the generic force field in eq, 24 such that it is a Finsler manifold as described in Appendix C. Each row shows 10 points on the geodesic, while the left most image is the data and the right most image is the estimated Fréchet mean using *GEORCE-FM*. The length of the estimated geodesic using *GEORCE-FM* is shown to the left, while the grid point number is shown at the bottom.

Fig. 12 The application of *GEORCE-FM* to manifold learning for a VAE estimating the Fréchet mean for 10 images of the CelebA dataset reconstructed using a VAE [52]. The learned manifold is equipped with the generic force field in eq, 24 such that it is a Finsler manifold as described in Appendix C. row shows 10 points on the geodesic, while the left most image is the data and the right most image is the estimated Fréchet mean using *GEORCE-FM*. The length of the estimated geodesic using *GEORCE-FM* is shown to the left, while the grid point number is shown at the bottom.

Riemannian Manifold	SGD (T=100) Length	SGD (T=100) Runtime	RMSprop (T=100) Length	RMSprop (T=100) Runtime	Adamax (T=100) Length	Adamax (T=100) Runtime
$\mathbb{S}^2$	**205.94**	2.1838 ± 0.0006	224.29	1.6724 ± 0.0014	227.18	$\mathbf{1.6636 \pm 0.0019}$
$\mathbb{S}^3$	**151.07**	1.6920 ± 0.0011	183.56	$\mathbf{1.6814 \pm 0.0011}$	188.34	1.7186 ± 0.0010
$\mathbb{S}^5$	194.97	$\mathbf{1.7908 \pm 0.0011}$	**179.87**	1.8230 ± 0.0010	193.20	1.8785 ± 0.0017
$\mathbb{S}^{10}$	203.93	$\mathbf{2.8978 \pm 0.0014}$	**95.59**	2.9154 ± 0.0012	118.48	2.9460 ± 0.0017
$\mathbb{S}^{20}$	143.37	$\mathbf{5.4790 \pm 0.0013}$	**48.80**	5.5168 ± 0.0012	69.48	5.5421 ± 0.0013
$\mathbb{S}^{50}$	64.22	23.7887 ± 0.0018	**19.21**	$\mathbf{22.8096 \pm 0.3671}$	28.00	23.9063 ± 0.0006
$\mathbb{S}^{100}$	32.36	97.5666 ± 4.6575	**10.58**	$\mathbf{84.6670 \pm 0.0665}$	13.50	84.8641 ± 0.0702
E (2)	137.83	$\mathbf{1.6446 \pm 0.0015}$	129.43	1.6579 ± 0.0016	**124.54**	1.6696 ± 0.0009
E (3)	**117.87**	$\mathbf{1.6822 \pm 0.0013}$	120.71	1.7087 ± 0.0008	120.92	1.7110 ± 0.0009
E (5)	**111.72**	$\mathbf{1.8051 \pm 0.0012}$	116.89	1.8220 ± 0.0014	125.84	1.8564 ± 0.0018
E (10)	108.97	2.9128 ± 0.0011	**68.20**	$\mathbf{2.9124 \pm 0.0013}$	75.54	2.9482 ± 0.0018
E (20)	82.70	$\mathbf{5.5250 \pm 0.0013}$	**37.48**	5.5792 ± 0.0009	44.89	5.5879 ± 0.0011
E (50)	41.87	$\mathbf{23.5703 \pm 0.0012}$	**14.97**	23.6562 ± 0.0020	18.62	23.7257 ± 0.0012
E (100)	22.31	$\mathbf{84.1223 \pm 0.0829}$	**7.95**	85.0741 ± 0.5396	9.21	84.4970 ± 0.1002
$\mathbb{T}^2$	–	$\mathbf{0.0219 \pm 0.0001}$	1250.43	1.6474 ± 0.0014	**1246.28**	1.6429 ± 0.0018
$\mathbb{H}^2$	**174.60**	$\mathbf{0.0225 \pm 0.0001}$	174.62	1.6579 ± 0.0013	174.61	1.6705 ± 0.0014
Paraboloid	$-/-$	–	112.73	1.6066 ± 0.0009	**104.19**	1.6068 ± 0.0016
Hyperbolic Paraboloid	$-/-$	–	114.56	1.6121 ± 0.0010	**114.40**	1.6126 ± 0.0014
Gaussian Distribution	243.03	0.2013 ± 0.0008	156.82	$\mathbf{0.1893 \pm 0.0014}$	**154.78**	0.2141 ± 0.0153
Fréchet Distribution	60189340.00	$\mathbf{0.0039 \pm 0.0001}$	32.89	0.1971 ± 0.0081	**31.51**	0.1932 ± 0.0004
Cauchy Distribution	77.17	$\mathbf{0.2181 \pm 0.0005}$	67.21	0.2299 ± 0.0014	**64.97**	0.2357 ± 0.0035
Pareto Distribution	25.48	$\mathbf{0.1966 \pm 0.0009}$	24.93	0.1967 ± 0.0073	**24.31**	0.2019 ± 0.0001
VAE MNIST	5687.01	$\mathbf{0.3928 \pm 0.0030}$	897.52	39.1152 ± 0.0014	**894.32**	39.1217 ± 0.0033
VAE CelebA	25601932831752192.00	$\mathbf{5.1663 \pm 0.0007}$	9483.42	857.7762 ± 0.0208	**9465.98**	862.3471 ± 0.0082

Table 9 The table shows the sum of squared lengths of the estimated Fréchet mean on a GPU. When the computational time was longer than 24 hours, the value is set to $-$.

	Adagrad (T=100)	
Riemannian Manifold	Length	Runtime
Manifold	Length	Runtime
$\mathbb{S}^2$	**226.54**	$\mathbf{1.6704 \pm 0.0012}$
$\mathbb{S}^3$	**184.13**	$\mathbf{1.6923 \pm 0.0006}$
$\mathbb{S}^5$	**178.95**	$\mathbf{1.8459 \pm 0.0005}$
$\mathbb{S}^{10}$	**106.87**	$\mathbf{2.9221 \pm 0.0014}$
$\mathbb{S}^{20}$	**64.50**	$\mathbf{5.5231 \pm 0.0024}$
$\mathbb{S}^{50}$	**26.57**	$\mathbf{23.8409 \pm 0.0004}$
$\mathbb{S}^{100}$	**12.87**	$\mathbf{104.3970 \pm 4.9584}$
E (2)	**124.39**	$\mathbf{1.6623 \pm 0.0012}$
E (3)	**120.60**	$\mathbf{1.6860 \pm 0.0012}$
E (5)	**120.13**	$\mathbf{1.8351 \pm 0.0011}$
E (10)	**69.30**	$\mathbf{2.9252 \pm 0.0013}$
E (20)	**39.57**	$\mathbf{5.5755 \pm 0.0011}$
E (50)	**16.46**	$\mathbf{23.6970 \pm 0.0009}$
E (100)	**7.99**	$\mathbf{85.1028 \pm 0.4041}$
$\mathbb{T}^2$	**1248.16**	$\mathbf{1.6418 \pm 0.0015}$
$\mathbb{H}^2$	**174.61**	$\mathbf{1.6710 \pm 0.0017}$
Paraboloid	**104.02**	$\mathbf{1.6051 \pm 0.0013}$
Hyperbolic Paraboloid	**114.40**	$\mathbf{1.6111 \pm 0.0022}$
Gaussian Distribution	**155.03**	$\mathbf{0.2229 \pm 0.0007}$
Fréchet Distribution	**31.44**	$\mathbf{0.2235 \pm 0.0005}$
Cauchy Distribution	**64.36**	$\mathbf{0.2480 \pm 0.0011}$
Pareto Distribution	**24.27**	$\mathbf{0.1998 \pm 0.0003}$
VAE MNIST	**884.30**	$\mathbf{39.0508 \pm 0.0012}$
VAE CelebA	**9452.99**	$\mathbf{835.0145 \pm 0.1027}$

Table 10 The table shows the sum of squared lengths of the estimated Fréchet mean on a GPU. When the computational time was longer than 24 hours, the value is set to $-$.

	SGD (T=100)		**RMSprop (T=100)**		**Adamax (T=100)**	
Finsler Manifold	Length	Runtime	Length	Runtime	Length	Runtime
$\mathbb{S}^2$	86.83	$\mathbf{2.3036 \pm 0.0018}$	83.81	2.3118 ± 0.0014	**83.68**	2.3185 ± 0.0015
$\mathbb{S}^3$	79.70	2.3912 ± 0.0003	79.16	$\mathbf{2.3890 \pm 0.0012}$	**78.49**	2.4210 ± 0.0014
$\mathbb{S}^5$	96.69	$\mathbf{3.1103 \pm 0.0013}$	84.52	3.1244 ± 0.0010	**84.43**	3.1469 ± 0.0014
$\mathbb{S}^{10}$	76.13	$\mathbf{5.4475 \pm 0.0019}$	**50.06**	5.4784 ± 0.0018	52.81	5.4966 ± 0.0032
$\mathbb{S}^{20}$	52.70	$\mathbf{12.3483 \pm 0.0015}$	**26.17**	12.3857 ± 0.0019	30.54	12.4081 ± 0.0013
$\mathbb{S}^{50}$	23.45	$\mathbf{54.9592 \pm 0.0016}$	**10.91**	55.0525 ± 0.0108	12.65	55.1001 ± 0.0013
$\mathbb{S}^{100}$	11.78	$\mathbf{179.9665 \pm 0.1075}$	**6.16**	180.1367 ± 0.0069	6.36	180.2978 ± 0.2581
E (2)	52.09	$\mathbf{2.3014 \pm 0.0023}$	49.72	2.3081 ± 0.0005	**46.92**	2.3178 ± 0.0034
E (3)	51.24	$\mathbf{2.3818 \pm 0.0010}$	50.88	2.3839 ± 0.0015	**49.31**	2.4363 ± 0.0017
E (5)	56.06	$\mathbf{3.0950 \pm 0.0020}$	**53.31**	3.0960 ± 0.0021	54.20	3.1361 ± 0.0014
E (10)	42.15	$\mathbf{5.0520 \pm 0.0022}$	**32.50**	5.0801 ± 0.0020	32.98	5.0919 ± 0.0018
E (20)	30.65	$\mathbf{12.4187 \pm 0.0018}$	**18.66**	12.4817 ± 0.0010	19.87	12.4936 ± 0.0014
E (50)	15.20	$\mathbf{54.6004 \pm 0.0015}$	**8.21**	54.6795 ± 0.0113	8.76	54.7518 ± 0.0011
E (100)	8.09	$\mathbf{176.7958 \pm 0.0389}$	4.99	176.9338 ± 0.0083	**4.84**	177.0561 ± 0.0270
$\mathbb{T}^2$	$-/-$	$-$	467.77	2.4788 ± 0.0014	**467.67**	2.4945 ± 0.0010
$\mathbb{H}^2$	74.76	$\mathbf{0.0296 \pm 0.0014}$	74.79	2.5035 ± 0.0019	**74.75**	2.5172 ± 0.0019
Paraboloid	inf	$\mathbf{0.0186 \pm 0.0001}$	**61.26**	2.2251 ± 0.0007	65.78	2.2251 ± 0.0008
Hyperbolic Paraboloid	$-$	$\mathbf{0.0206 \pm 0.0001}$	**52.62**	2.2064 ± 0.0004	62.22	2.2156 ± 0.0007
Gaussian Distribution	76.16	0.7477 ± 0.0019	68.64	$\mathbf{0.7352 \pm 0.0003}$	**64.75**	0.7513 ± 0.0016
Fréchet Distribution	35.58	$\mathbf{0.6141 \pm 0.0031}$	**11.74**	0.6541 ± 0.0019	15.06	0.6393 ± 0.0007
Cauchy Distribution	28.10	0.7585 ± 0.0044	26.92	$\mathbf{0.7365 \pm 0.0026}$	**25.73**	0.7628 ± 0.0020
Pareto Distribution	27.73	0.6621 ± 0.0011	**11.36**	$\mathbf{0.6529 \pm 0.0072}$	15.52	0.6633 ± 0.0035
VAE MNIST	$-/-$	$-$	316.46	463.9254 ± 0.0712	**313.52**	462.4614 ± 0.2502
VAE CelebA	$-$	$\mathbf{9.6821 \pm 0.0034}$	3213.31	1602.4119 ± 0.1665	**3202.72**	1597.8483 ± 0.3720

Table 11 The table shows the sum of squared lengths of the estimated Fréchet mean on a GPU. When the computational time was longer than 24 hours, the value is set to $-$.

	Adagrad (T=100)	
Finsler Manifold	Length	Runtime
$\mathbb{S}^2$	**83.67**	$\mathbf{2.3119 \pm 0.0006}$
$\mathbb{S}^3$	**78.23**	$\mathbf{2.4114 \pm 0.0014}$
$\mathbb{S}^5$	**81.83**	$\mathbf{3.1352 \pm 0.0013}$
$\mathbb{S}^{10}$	**49.19**	$\mathbf{5.4872 \pm 0.0015}$
$\mathbb{S}^{20}$	**28.47**	$\mathbf{12.3839 \pm 0.0012}$
$\mathbb{S}^{50}$	**12.06**	$\mathbf{55.1034 \pm 0.0018}$
$\mathbb{S}^{100}$	**6.14**	$\mathbf{180.1266 \pm 0.0070}$
E (2)	**46.76**	$\mathbf{2.3096 \pm 0.0007}$
E (3)	**49.15**	$\mathbf{2.3999 \pm 0.0022}$
E (5)	**53.28**	$\mathbf{3.1178 \pm 0.0022}$
E (10)	**31.17**	$\mathbf{5.0837 \pm 0.0031}$
E (20)	**17.89**	$\mathbf{12.4912 \pm 0.0016}$
E (50)	**7.90**	$\mathbf{54.7486 \pm 0.0008}$
E (100)	**5.22**	$\mathbf{176.9808 \pm 0.0056}$
$\mathbb{T}^2$	**467.59**	$\mathbf{2.4961 \pm 0.0019}$
$\mathbb{H}^2$	**74.73**	$\mathbf{2.5105 \pm 0.0021}$
Paraboloid	**65.71**	$\mathbf{2.2253 \pm 0.0005}$
Hyperbolic Paraboloid	**61.45**	$\mathbf{2.2111 \pm 0.0006}$
Gaussian Distribution	**65.12**	$\mathbf{0.7496 \pm 0.0003}$
Fréchet Distribution	**14.97**	$\mathbf{0.6548 \pm 0.0014}$
Cauchy Distribution	**25.70**	$\mathbf{0.7510 \pm 0.0018}$
Pareto Distribution	**16.07**	$\mathbf{0.6515 \pm 0.0016}$
VAE MNIST	**310.68**	$\mathbf{462.2765 \pm 0.0565}$
VAE CelebA	**3194.40**	$\mathbf{1644.1156 \pm 0.0458}$

Table 12 The table shows the sum of squared lengths of the estimated Fréchet mean on a GPU. When the computational time was longer than 24 hours, the value is set to $-$.