Algorithms for Riemannian Optimization

K. A. Gallivan[1](*) P.-A. Absil[2] Chunhong Qi[1] C. G. Baker[3]

[1]Florida State University

[2]Catholic University of Louvain

[3]Oak Ridge National Laboratory

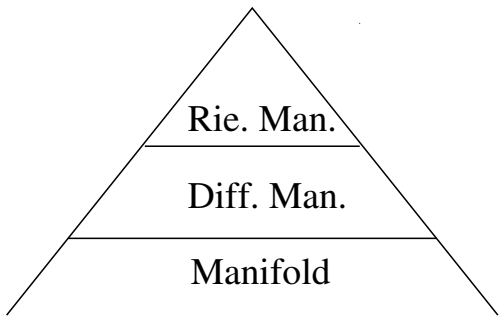
ICCOPT, July 2010

Why Riemannian Manifolds?

A manifold is a set that is
locally Euclidean.

A Riemannian manifold is a differentiable manifold with a Riemannian metric:

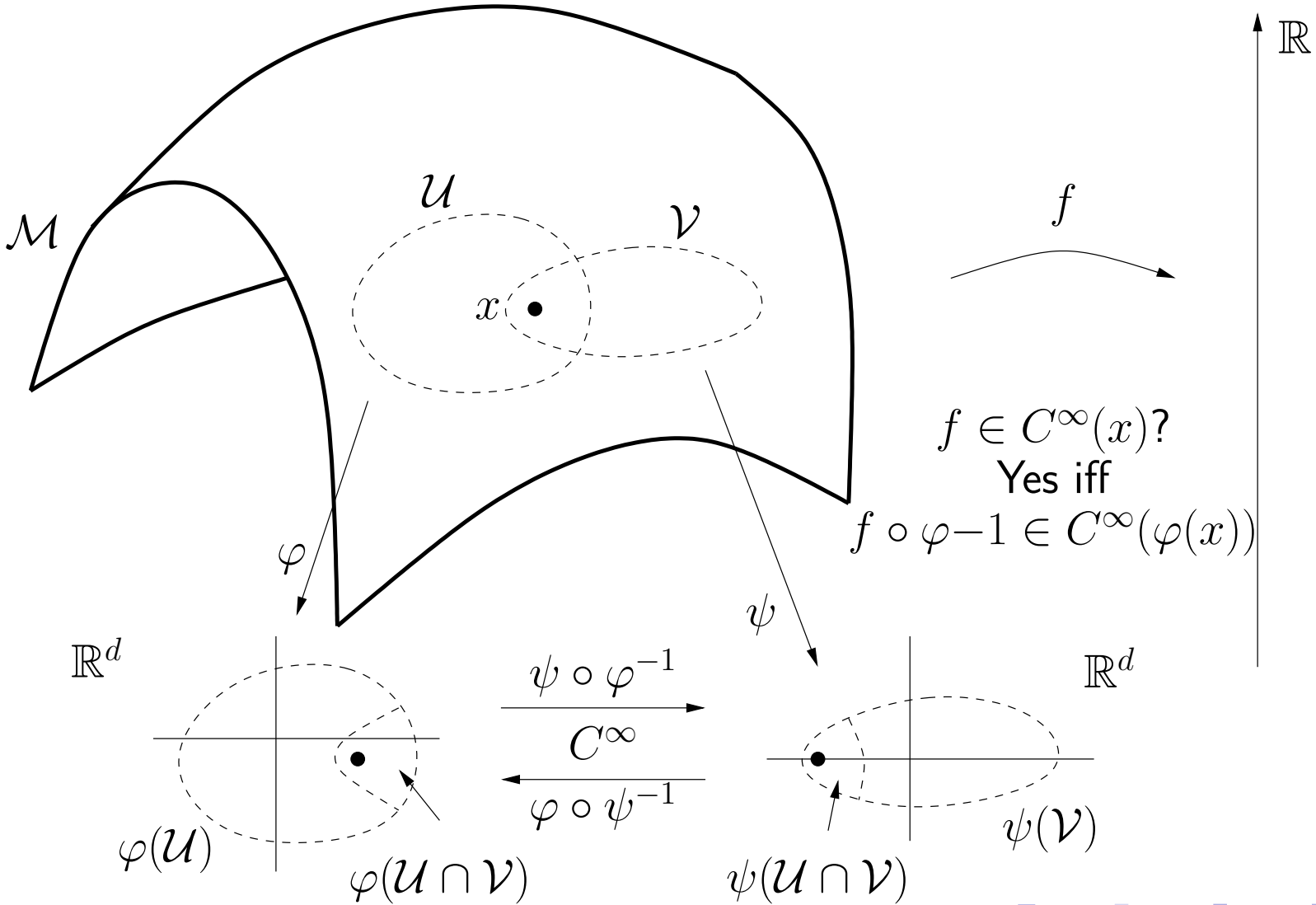
- The manifold gives us topology.
- Differentiability gives us calculus.
- The Riemannian metric gives us geometry.



Riemannian manifolds strike a balance between power and practicality.

Riemannian Manifolds

Roughly, a Riemannian manifold is a smooth set with a smoothly-varying inner product on the tangent spaces.



Noteworthy Manifolds and Applications

- Stiefel manifold $\mathrm{St}(p,n)$ and orthogonal group $O_p = \mathrm{St}(p,p)$ and Unit sphere $S^{n-1} = \mathrm{St}(1,n)$

$$\mathrm{St}(p,n) = \{X \in \mathbb{R}^{n\times p} : X^T X = I_p\}$$

  $d = np - \frac{p(p+1)}{2}$, Applications: computer vision; principal component analysis; independent component analysis...
- Grassmann manifold $\mathrm{Gr}(p,n)$

  Set of all p-dimensional subspaces of $\mathbb{R}^n$

  $d = np - p^2$ Applications: various dimension reduction problems...
- Oblique manifold $\mathbb{R}_*^{n\times p}/\mathcal{S}_{\mathrm{diag}+}$

$$\mathbb{R}_*^{n\times p}/\mathcal{S}_{\mathrm{diag}+} \simeq \{Y \in \mathbb{R}_*^{n\times p} : \mathrm{diag}(Y^T Y) = I_p\} = S^{n-1}\times\cdots\times S^{n-1}$$

  $d = (n-1)p^2$ Applications: blind source separation ...

Noteworthy Manifolds and Applications

- Set of fixed-rank PSD matrices $\mathrm{S}_+(p,n)$. A quotient representation:

$$X \sim Y \Leftrightarrow \exists Q \in O_p : Y = XQ$$

  Applications: Low-rank approximation of symmetric matrices; low-rank approximation of tensors...
- Flag manifold $\mathbb{R}_*^{n\times p}/\mathcal{S}_{\mathrm{upp}_*}$
  Elements of the flag manifold can be viewed as a p-tuple of linear subspaces $(\mathcal{V}_1,\ldots,\mathcal{V}_p)$ such that $\dim(\mathcal{V}_i) = i$ and $\mathcal{V}_i \subset \mathcal{V}_{i+1}$.
  Applications: analysis of QR algorithm...
- Shape manifold $O_n\backslash\mathbb{R}_*^{n\times p}$

$$Y \sim X \Leftrightarrow \exists U \in O_n : Y = UX$$

  Applications: shape analysis

Iterations on the Manifold

Consider the following generic update for an iterative Euclidean optimization algorithm:

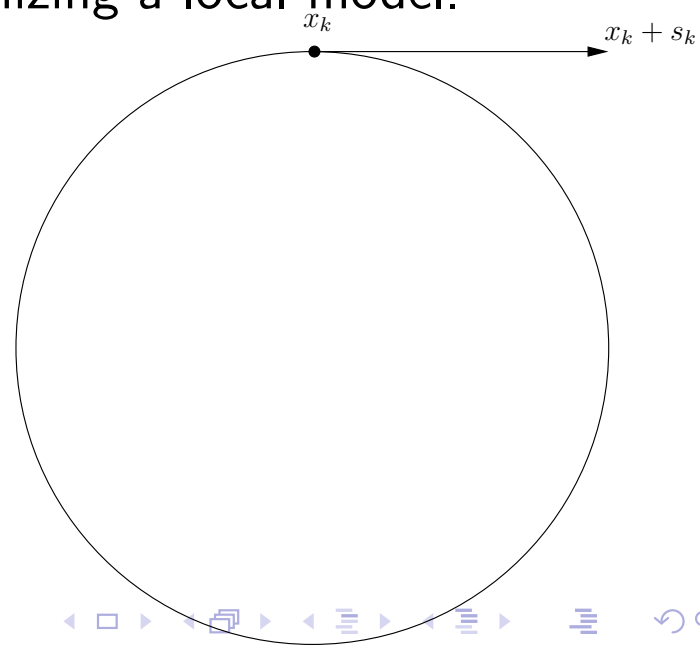
$$x_{k+1} = x_k + \Delta x_k = x_k + \alpha_k s_k \ .$$

This iteration is implemented in numerous ways, e.g.:

- Steepest descent: $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$
- Newton's method: $x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$
- Trust region method: Δx_k is set by optimizing a local model.

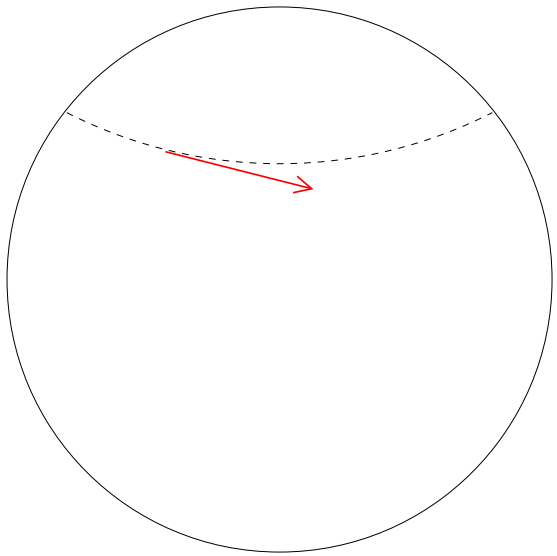
Riemannian Manifolds Provide

- Riemannian concepts describing directions and movement on the manifold
- Riemannian analogues for gradient and Hessian



Tangent Vectors

- The concept of direction is provided by tangent vectors.
- Intuitively, tangent vectors are tangent to curves on the manifold.
- Tangent vectors are an intrinsic property of a differentiable manifold.

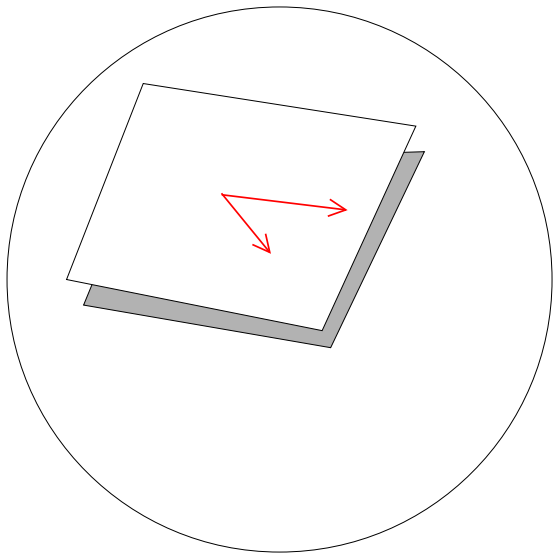


Definition

The tangent space T_xM is the vector space comprised of the tangent vectors at $x \in M$. The Riemannian metric is an inner product on each tangent space.

Tangent Vectors

- The concept of direction is provided by tangent vectors.
- Intuitively, tangent vectors are tangent to curves on the manifold.
- Tangent vectors are an intrinsic property of a differentiable manifold.



Definition

The tangent space T_xM is the vector space comprised of the tangent vectors at $x \in M$. The Riemannian metric is an inner product on each tangent space.

Riemannian gradient and Riemannian Hessian

Definition

The Riemannian gradient of f at x is the unique tangent vector in T_xM satisfying $\forall \eta \in T_xM$

$$\mathrm{D}\, f(x)[\eta] = \langle \mathrm{grad}\, f(x), \eta \rangle$$

and $\mathrm{grad}\, f(x)$ is the direction of steepest ascent.

Definition

The Riemannian Hessian of f at x is a symmetric linear operator from T_xM to T_xM defined as

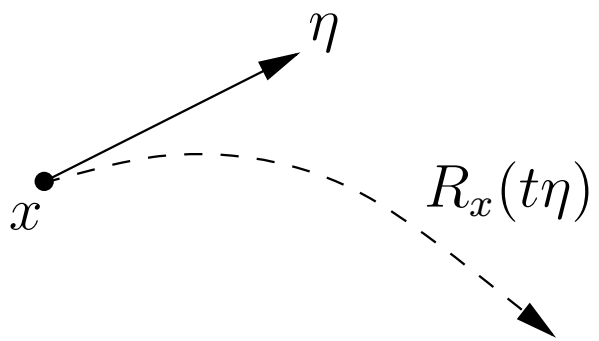
$$\mathrm{Hess}\, f(x) : T_xM \to T_xM : \eta \to \nabla_\eta \mathrm{grad}\, f$$

Retractions

Definition

A retraction is a mapping R from TM to M satisfying the following:

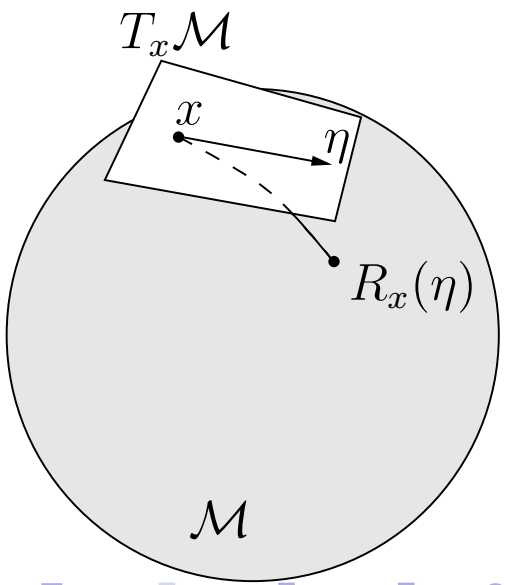
- R is continuously differentiable
- $R_x(0) = x$
- $\mathrm{D}\, R_x(0)[\eta] = \eta$



- maps tangent vectors back to the manifold
- lifts objective function f from M to T_xM, via the pullback

$$\hat{f}_x = f \circ R_x$$

- defines curves in a direction
- exponential map $\mathrm{Exp}(t\eta)$ defines "straight lines" geodesic

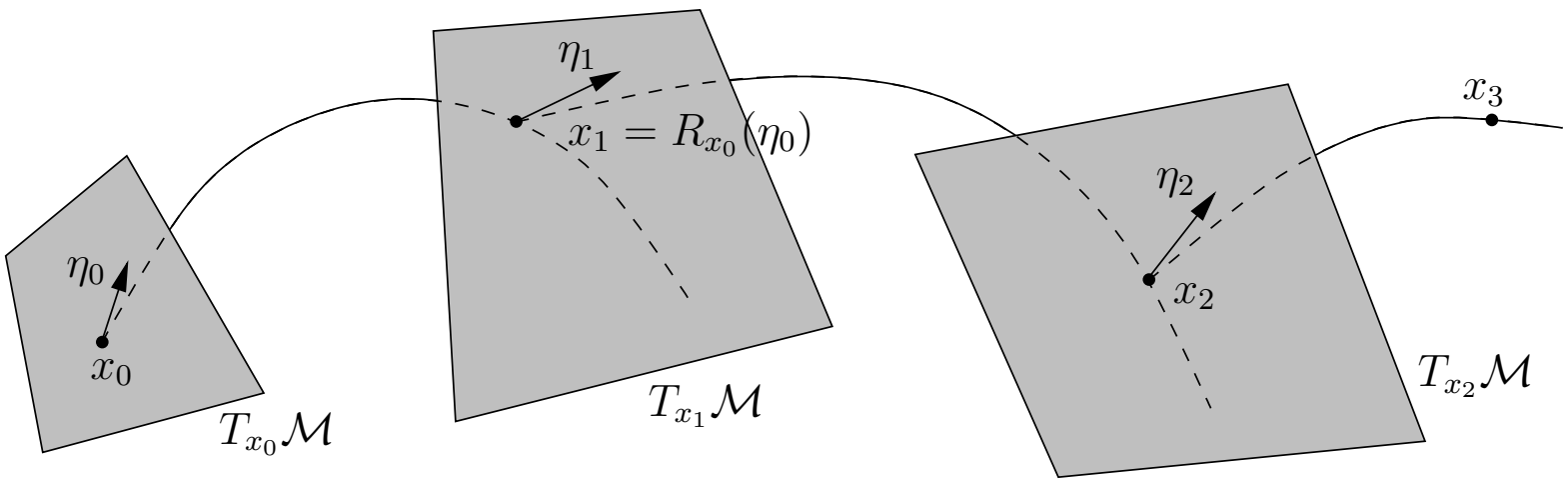


Generic Riemannian Optimization Algorithm

1. At iterate $x \in M$, define $\hat{f}_x = f \circ R_x$.
2. Find $\eta \in T_x M$ which satisfies certain condition.
3. Choose new iterate $x_+ = R_x(\eta)$.
4. Goto step 1.

A suitable setting

This paradigm is sufficient for describing many optimization methods.



Categories of Riemannian optimization methods

Retraction-based: local information only

Line search-based: use local tangent vector and $R_x(t\eta)$ to define line

- Steepest decent: geodesic in the direction $-\operatorname{grad} f(x)$
- Newton

Local model-based: series of flat space problems

- Riemannian Trust region (RTR)
- Riemannian Adaptive Cubic Overestimation (RACO)

Retraction and transport-based: information from multiple tangent spaces

- Conjugate gradient and accelerated iteration: multiple tangent vectors
- Quasi-Newton e.g. Riemannian BFGS: transport operators between tangent spaces

Basic principles

All/some elements required for optimizing a cost function (M, g):

- an efficient numerical representation for points x on M , for tangent spaces T_xM , and for the inner products $g_x(\cdot, \cdot)$ on T_xM ;
- choice of a retraction $R_x : T_xM \to M$;
- formulas for $f(x), \operatorname{grad} f(x)$ and $\operatorname{Hess} f(x)$;
- formulas for combining information from multiple tangent spaces.

Parallel transport

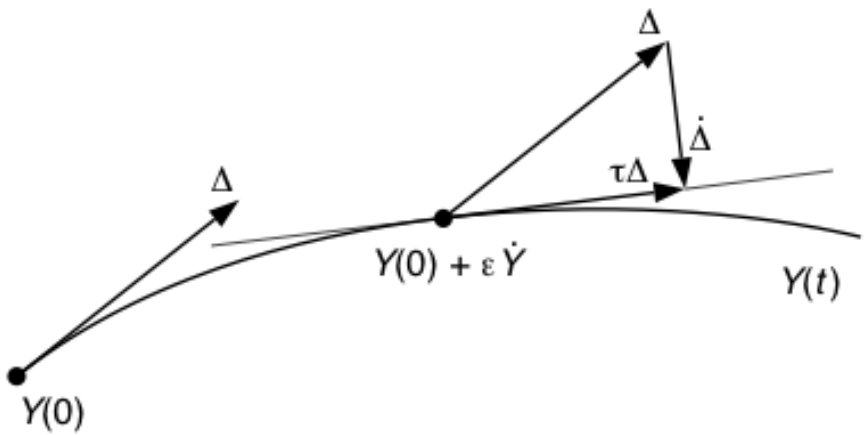


Figure: Parallel transport

- Parallel transport one tangent vector along some curve $Y(t)$.
- It is often along the geodesic $\gamma_\eta(t) : \mathbb{R} \to M : t \to \mathrm{Exp}_x(t\eta_x)$.
- In general, geodesics and parallel translation require solving an ordinary differential equation.

Vector transport

Definition

We define a **vector transport** on a manifold M to be a smooth mapping

$$TM \oplus TM \to TM : (\eta_x, \xi_x) \mapsto \mathcal{T}_{\eta_x}(\xi_x) \in TM$$

satisfying the following properties for all $x \in M$.

- (Associated retraction) There exists a retraction R, called the *retraction associated with* $\mathcal{T}$, s.t. the following diagram holds

$$\begin{array}{ccc} (\eta_x, \xi_x) & \xrightarrow{\mathcal{T}} & \mathcal{T}_{\eta_x}(\xi_x) \\ \downarrow & & \downarrow \pi \\ \eta_x & \xrightarrow[R]{} & \pi\left(\mathcal{T}_{\eta_x}(\xi_x)\right) \end{array}$$

  where $\pi\left(\mathcal{T}_{\eta_x}(\xi_x)\right)$ denotes the foot of the tangent vector $\mathcal{T}_{\eta_x}(\xi_x)$.
- (Consistency) $\mathcal{T}_{0_x}\xi_x = \xi_x$ for all $\xi_x \in T_xM$;
- (Linearity) $\mathcal{T}_{\eta_x}(a\xi_x + b\zeta_x) = a\mathcal{T}_{\eta_x}(\xi_x) + b\mathcal{T}_{\eta_x}(\zeta_x)$.

Vector transport

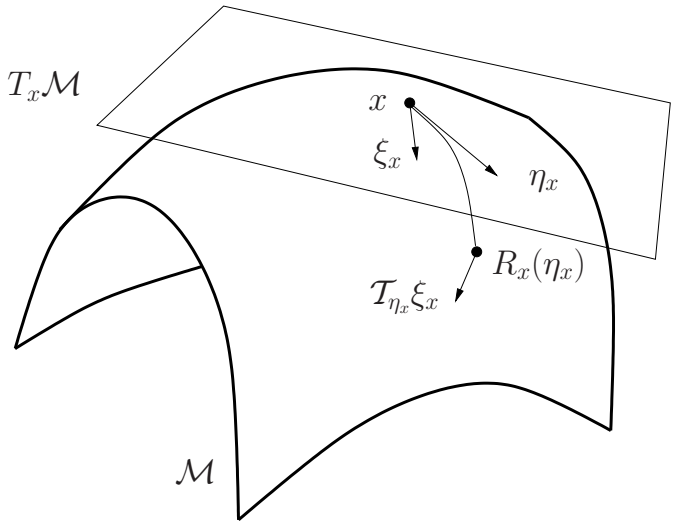


Figure: Vector transport.

- Want to be efficient while maintaining rapid convergence.
- Parallel transport is an isometric vector transport for $\gamma(t) = R_x(t\eta_x)$ with additional properties.
- When $\mathcal{T}_{\eta_x}\xi_x$ exists, if η and ξ are two vector fields on M, this defines

$$\left((\mathcal{T}_\eta)^{-1}\xi\right)_x := (\mathcal{T}_{\eta_x})^{-1}\left(\xi_{R_x(\eta_x)}\right) \in T_xM.$$

Retraction/Transport-based Riemannian Optimization

Benefits

- Increased generality does not compromise the important theory
- Can easily employ classical optimization techniques
- Less expensive than or similar to previous approaches
- May provide theory to explain behavior of algorithms in a particular application – or closely related ones

Possible Problems

- May be inefficient compared to algorithms that exploit application details

Retraction-based Riemannian optimization

Equivalence of the pullback $\hat{f}_x = f \circ R_x$

	Exp_x	R_x
$\mathrm{grad}\, f(x) = \mathrm{grad}\, \hat{f}_x(0)$	yes	yes
$\mathrm{Hess}\, f(x) = \mathrm{Hess}\, \hat{f}_x(0)$	yes	no
$\mathrm{Hess}\, f(x) = \mathrm{Hess}\, \hat{f}_x(0)$ at critical points	yes	yes

Sufficient Optimality Conditions

If $\mathrm{grad}\, \hat{f}_x(0) = 0$ and $\mathrm{Hess}\, \hat{f}_x(0) > 0$,
then $\mathrm{grad}\, f(x) = 0$ and $\mathrm{Hess}\, f(x) > 0$,
so that x is a local minimizer of f.

Retraction-based Riemannian optimization

Equivalence of the pullback $\hat{f}_x = f \circ R_x$

	Exp_x	R_x
$\mathrm{grad}\, f(x) = \mathrm{grad}\, \hat{f}_x(0)$	yes	yes
$\mathrm{Hess}\, f(x) = \mathrm{Hess}\, \hat{f}_x(0)$	yes	no
$\mathrm{Hess}\, f(x) = \mathrm{Hess}\, \hat{f}_x(0)$ at critical points	yes	yes

Sufficient Optimality Conditions

If $\mathrm{grad}\, \hat{f}_x(0) = 0$ and $\mathrm{Hess}\, \hat{f}_x(0) > 0$,
then $\mathrm{grad}\, f(x) = 0$ and $\mathrm{Hess}\, f(x) > 0$,
so that x is a local minimizer of f.

Some History of Optimization On Manifolds (I)

Luenberger (1973), *Introduction to linear and nonlinear programming*. Luenberger mentions the idea of performing line search along geodesics, "which we would use if it were computationally feasible (which it definitely is not)".

Gabay (1982), *Minimizing a differentiable function over a differential manifold*. Stepest descent along geodesics; Newton's method along geodesics; Quasi-Newton methods along geodesics. On Riemannian submanifolds of $\mathbb{R}^n$.

Some History of Optimization On Manifolds (II)

Smith (1993-94), *Optimization techniques on Riemannian manifolds*. Levi-Civita connection ∇; Riemannian exponential; parallel translation. But Remark 4.9: If Algorithm 4.7 (Newton's iteration on the sphere for the Rayleigh quotient) is simplified by replacing the exponential update with the update

$$x_{k+1} = \frac{x_k + \eta_k}{\|x_k + \eta_k\|}$$

then we obtain the Rayleigh quotient iteration.

Helmke and Moore (1994), *Optimization techniques on Riemannian manifolds*. dynamical systems, flows on manifolds, SVD, balancing, eigenvalues

Some History of Optimization On Manifolds (III)

The "pragmatic era" begins:

Manton (2002), *Optimization algorithms exploiting unitary constraints* "The present paper breaks with tradition by not moving along geodesics". The geodesic update $\mathrm{Exp}_x \eta$ is replaced by a projective update $\pi(x+\eta)$, the *projection* of the point $x+\eta$ onto the manifold.

Adler, Dedieu, Shub, et al. (2002), *Newton's method on Riemannian manifolds and a geometric model for the human spine*. The exponential update is relaxed to the general notion of *retraction*. The geodesic can be replaced by any (smoothly prescribed) curve tangent to the search direction.

Absil, Mahony, Sepulchre (2007) Nonlinear CG using retractions.

Some History of Optimization On Manifolds (IV)

Absil, Baker, Gallivan (2004-07), Theory and implementations of Riemannian Trust Region method. Retraction-based approach. Matrix manifold problems, software repository

`http://www.math.fsu.edu/~cbaker/GenRTR`

Anasazi Eigenproblem package in Trilinos Library at Sandia National Laboratory

Dresigmeyer (2007), Nelder-Mead using geodesics.

Absil, Gallivan, Qi (2007-10), Theory and implementations of Riemannian BFGS and Riemannian Adaptive Cubic Overestimation. Retraction-based and vector transport-based.

Trust-Region Method

1a. At iterate x, define pullback $\hat{f}_x = f \circ R_x$

1. Construct quadratic model m_x of f around x
2. Find (approximate) solution to

$$\eta = \operatorname*{argmin}_{\|\eta\| \leq \Delta} m_x(\eta)$$

3. Compute $\rho_x(\eta)$:

$$\rho_x(\eta) = \frac{f(x) - f(x+\eta)}{m_x(0) - m_x(\eta)}$$

4. Use $\rho_x(\eta)$ to adjust Δ and accept/reject new iterate:

$$x_+ = x + \eta$$

Convergence Properties

Retains convergence of Euclidean trust-region methods:

- robust global and superlinear convergence (Baker, Absil, Gallivan)

Riemannian Trust-Region Method

1a. At iterate x, define pullback $\hat{f}_x = f \circ R_x$

1b. Construct quadratic model m_x of $\hat{f}_x$

2. Find (approximate) solution to

$$\eta = \underset{\eta \in T_x M,\, \|\eta\| \leq \Delta}{\operatorname{argmin}} \; m_x(\eta)$$

3. Compute $\rho_x(\eta)$:

$$\rho_x(\eta) = \frac{\hat{f}_x(0) - \hat{f}_x(\eta)}{m_x(0) - m_x(\eta)}$$

4. Use $\rho_x(\eta)$ to adjust Δ and accept/reject new iterate:

$$x_+ = R_x(\eta)$$

Convergence Properties

Retains convergence of Euclidean trust-region methods:

- robust global and superlinear convergence (Baker, Absil, Gallivan)

RTR Applications

Large-scale Generalized Symmetric Eigenvalue Problem and SVD (Absil, Baker, Gallivan 2004-08)

Blind source separation on both Orthogonal group and Oblique manifold (Absil and Gallivan 2006)

Low-rank approximate solution symmetric positive definite Lyapanov $AXM + MXA = C$ (Vandereycken and Vanderwalle 2009)

Best low-rank approximation to a tensor (Ishteva, Absil, Van Huffel, De Lathauwer 2010)

RTR Applications

- RTR tends to be:
  - robust and often yields final cost function values noticeably smaller than other methods
  - slightly to noticeably more expensive than less reliable or less successful methods
- solutions:
  - reduce cost of trust region adaptation and excessive reduction of cost function by exploiting knowledge of cost function and application, e.g., Implicit RTR for large scale eigenvalue problems (Absil, Baker, Gallivan 2006)
  - combine with linearly convergent initial linearly convergent but less expensive method, RTR and TRACEMIN for large scale eigenvalue problems (Absil, Baker, Gallivan, Sameh 2006)

Joint diagonalization on Orthogonal Group

- $x(t) = As(t)$, $x(t) \in \mathbb{R}^n$, $s(t) \in \mathbb{R}^p$, $A \in \mathbb{R}^{n \times p}$
- Measurements

$$X = \begin{bmatrix} x_1(t_1) & x_1(t_2) & \cdots & x_1(t_m) \\ \vdots & \vdots & \ddots & \vdots \\ x_n(t_1) & x_n(t_2) & \cdots & x_n(t_m) \end{bmatrix} = A \begin{bmatrix} s_1(t_1) & \cdots & s_1(t_m) \\ \vdots & \ddots & \vdots \\ s_p(t_1) & \cdots & s_p(t_m) \end{bmatrix}$$

- Goal: Find a matrix $W \in \mathbb{R}^{n \times p}$ such that the rows of

$$Y = W^T X \in \mathbb{R}^{p \times m}$$

look as statistically independent as possible, $YY^T \approx D \in \mathbb{R}^{p \times p}$

- Decompose $W = U\Sigma V^T$. We have

$$Y = V^T \underbrace{\Sigma U^T X}_{=:\tilde{X} \in \mathbb{R}^{p \times m}} .$$

Joint diagonalization on Orthogonal Group

- Whitening: Choose Σ and U such that $\tilde{X}\tilde{X}^T = I_p$. Then

$$Y = V^T \tilde{X} \in \mathbb{R}^{p \times m}$$

$$YY^T = V^T \tilde{X} \tilde{X}^T V = V^T V = I_p$$

- Independence and dimension reduction: Consider a collection of covariance-like matrix functions $C_i(Y) \in \mathbb{R}^{p \times p}$ such that $C_i(Y) = V^T C_i(\tilde{X}) V$. Choose V to make the $C_i(Y)$'s as diagonal as possible.
- Principle: Solve

$$\max_{V^T V = I_p} \sum_{i=1}^{N} \|\mathrm{diag}(V^T C_i(\tilde{X}) V)\|_F^2.$$

Joint diagonalization on Orthogonal Group

Blind source separation

Two mixed pictures are given as input to a blind source separation algorithm based on a trust-region method on $\mathrm{St}22$.

Nonorthogonal form $Y = W^T X$ on Oblique manifold also effective.

Joint diagonalization on Orthogonal Group

Joint diagonalization on Orthogonal Group

Riemannian BFGS: past and future

Previous work on BFGS on manifolds

- Gabay 1982 discussed a version using parallel translation
- Brace and Manton 2006 restrict themselves to a version on the Grassmann manifold and the problem of weighted low-rank approximations.
- Savas and Lim 2008 apply a version to the more complicated problem of best multilinear approximations with tensors on a product of Grassmann manifolds.

Our goals

- Make the algorithm more efficient.
- Understand its convergence properties.

Algorithm 1 The Riemannian BFGS (RBFGS) algorithm

1: Given: Riemannian manifold (M, g); vector transport $\mathcal{T}$ on M with associated retraction R; real-valued function f on M; initial iterate $\mathbf{x}_1 \in M$; initial Hessian approximation $\mathcal{B}_1$;
2: **for** k = 1, 2,... **do**
3: Obtain $\eta_k \in T_{\mathbf{x}_k} M$ by solving: $\eta_k = -\mathcal{B}_k^{-1} \text{grad } f(\mathbf{x}_k)$.
4: Perform a line search on $\mathbb{R} \ni \alpha \mapsto f(R_{\mathbf{x}_k}(\alpha \eta_k)) \in \mathbb{R}$ to obtain a step size α_k; set $\mathbf{x}_{k+1} = R_{\mathbf{x}_k}(\alpha_k \eta_k)$.
5: Define $s_k = \mathcal{T}_{\alpha \eta_k} \alpha \eta_k$ and $y_k = \text{grad } f(\mathbf{x}_{k+1}) - \mathcal{T}_{\alpha \eta_k} \text{grad } f(\mathbf{x}_k)$
6: Define the linear operator $\mathcal{B}_{k+1} : T_{\mathbf{x}_{k+1}} M \to T_{\mathbf{x}_{k+1}} M$ as follows

$$\mathcal{B}_{k+1} p = \tilde{\mathcal{B}}_k p - \frac{g(s_k, \tilde{\mathcal{B}}_k p)}{g(s_k, \tilde{\mathcal{B}}_k s_k)} \tilde{\mathcal{B}}_k s_k + \frac{g(y_k, p)}{g(y_k, s_k)} y_k, \ \forall p \in T_{\mathbf{x}_{k+1}} M$$

with $\tilde{\mathcal{B}}_k = \mathcal{T}_{\alpha_k \eta_k} \circ \mathcal{B}_k \circ (\mathcal{T}_{\alpha_k \eta_k})^{-1}$
7: **end for**

Other versions of the RBFGS algorithm

- An iterative method can be used to solve the system or a factorization transported/updated.
- choice dictates what properties, e.g., positive definiteness, must be preserved
- An alternative works with the inverse Hessian $\mathcal{H}_k = {\mathcal{B}_k}^{-1}$ approximation rather than the Hessian approximation B_k.
- Step 6 in algorithm 1 becomes:
  $\mathcal{H}_{k+1} =$
  $\tilde{\mathcal{H}}_k p - \frac{g(y_k, \tilde{\mathcal{H}}_k p)}{g(y_k, s_k)} s_k - \frac{g(s_k, p_k)}{g(y_k, s_k)} \tilde{\mathcal{H}}_k y_k + \frac{g(s_k, p) g(y_k, \tilde{\mathcal{H}}_k y_k)}{g(y_k, s_k)^2} s_k + \frac{g(s_k, s_k)}{g(y_k, s_k)} p$
  with
  $$\tilde{\mathcal{H}}_k = \mathcal{T}_{\eta_k} \circ \mathcal{H}_k \circ (\mathcal{T}_{\eta_k})^{-1}$$
- Makes it possible to cheaply compute an approximation of the inverse of the Hessian. This may make BFGS advantageous even in the case where we have a cheap exact formula for the Hessian but not for its inverse.

Global convergence of RBFGS

Assumption 1
(1) The objective function f is twice continuously differentiable
(2) The level set $\Omega = \{x \in M : f(x) \leq f(x_0)\}$ is convex. In addition, there exists positive constants n and N such that

$$ng(z,z) \leq g(G(x)z,z) \leq Ng(z,z) \text{ for all } z \in M \text{ and } x \in \Omega$$

where $G(x)$ denotes the lifted Hessian.

Theorem

Let $\mathcal{B}_0$ be any symmetric positive definite matrix, and let x_0 be starting point for which assumption 1 is satisfied.Then the sequence x_k generated by algorithm 1 converge to the minimizer of f.

Superlinear convergence of RBFGS

Generalized Dennis-Moré condition Let M be a manifold endowed with a C^2 vector transport $\mathcal{T}$ and an associated retraction R. Let F be a C^2 tangent vector field on M. Also let M be endowed with an affine connection ∇ and let $\mathbb{D}F(x)$ denote the linear transformation of T_xM defined by $\mathbb{D}F(x)[\xi_x] = \nabla_{\xi_x} F$ for all tangent vectors ξ_x to M at x. Let $\{\mathcal{B}_k\}$ be a sequence of bounded nonsingular linear transformation of $T_{x_k}M$, where $k = 0, 1, \cdots$, $x_{k+1} = R_{x_k}(\eta_k)$, and $\eta_k = -\mathcal{B}_k^{-1} F(x_k)$. Assume that $\mathbb{D}F(x^*)$ is nonsingular, $x_k \neq x^*, \forall k$, and $\lim_{k\to\infty} x_k = x^*$. Then $\{x_k\}$ converges superlinearly to x^* and $F(x^*) = 0$ if and only if

$$\lim_{k\to\infty} \frac{\|[\mathcal{B}_k - \mathcal{T}_{\xi_k} \mathbb{D}F(x^*) \mathcal{T}_{\xi_k}^{-1}]\eta_k\|}{\|\eta_k\|} = 0 \qquad (1)$$

where $\xi_k \in T_{x^*}M$ is defined by $\xi_k = R_{x^*}^{-1}(x_k)$, i.e. $R_{x^*}(\xi_k) = x_k$.

Superlinear convergence of RBFGS

Assumption 2 The lifted Hessian matrix $\mathrm{Hess}\, \widehat{f}_x$ is Lipschitz-continuous at 0_x uniformly in a neighbourhood of x^*, i.e., there exists $L_* > 0, \delta_1 > 0$, and $\delta_2 > 0$ such that, for all $x \in \mathcal{B}_{\delta_1}(x^*)$ and all $\xi \in \mathcal{B}_{\delta_2}(0_x)$, it holds that

$$\|\mathrm{Hess}\ \widehat{f}_x(\xi) - \mathrm{Hess}\ \widehat{f}_x(0_x)\|_x \leq L_* \|\xi\|_x$$

Theorem

Suppose that f is twice continuously differentiable and that the iterates generated by the RBFGS algorithm converge to a nondegenerate minimizer $x^* \in M$ at which Assumption 2 holds. Suppose also that $\sum_{k=1}^{\infty} \|x_k - x^*\| < \infty$ holds. Then x_k converges to x^* at a superlinear rate.

Implementation Choices

Approach 1: Realize $\mathcal{B}_k$ by an n-by-n matrix $B_k^{(n)}$.

Let $\mathcal{B}_k$ be the linear operator $\mathcal{B}_k : T_{x_k}M \longrightarrow T_{x_k}M$, $B_k^{(n)} \in \mathbb{R}^{n\times n}$, s.t

$$\mathsf{i}_{x_k}(\mathcal{B}_k\eta_k) = B_k^{(n)}(\mathsf{i}_{x_k}(\eta_k)), \forall \eta_k \in T_{x_k}M,$$

$$\text{from } \mathcal{B}_k\eta_k = -\text{grad } f(x_k)$$

$$\text{we have } B_k^{(n)}(\mathsf{i}_{x_k}(\eta_k)) = -\mathsf{i}_{x_k}(\text{grad } f(x_k)).$$

Implementation Choices

Approach 2: Use bases.

Let $[E_{k,1}, \cdots, E_{k,d}] =: \underline{E}_k \in \mathbb{R}^{n \times d}$ be a basis of $T_{x_k} M$. We have

$$\underline{E}_k^+ B_k^{(n)} \underline{E}_k \, \underline{E}_k^+ \mathsf{i}_{x_k}(\eta_k) = -\underline{E}_k^+ \mathsf{i}_{x_k}(\text{grad } f(x_k))$$

where $\underline{E}_k^+ = (\underline{E}_k^T \underline{E}_k)^{-1} \underline{E}_k^T$

$$B_k^d = \underline{E}_k^+ B_k^{(n)} \underline{E}_k \in \mathbb{R}^{d \times d}$$

$$B_k^{(d)}(\eta_k)^{(d)} = -(\text{grad } f(x_k))^{(d)}$$

Additional Transport Constraints

BFGS: symmetry and positive definiteness of B_k are preserved

RBFGS: we want to know

1. When transport information between multiple tangent spaces
   - Are symmetry/positive definite of B_k preserved?
   - Is it possible?
   - Is it important?
2. Implementation efficiency
3. Projection frame work for embedded submanifold allows us to do

Additional Transport Constraints

Embedded submanifold: projection-based

- Nonisometric vector transport
- Isometric vector transport (symmetry preserving)
- Efficiency via multiple choices

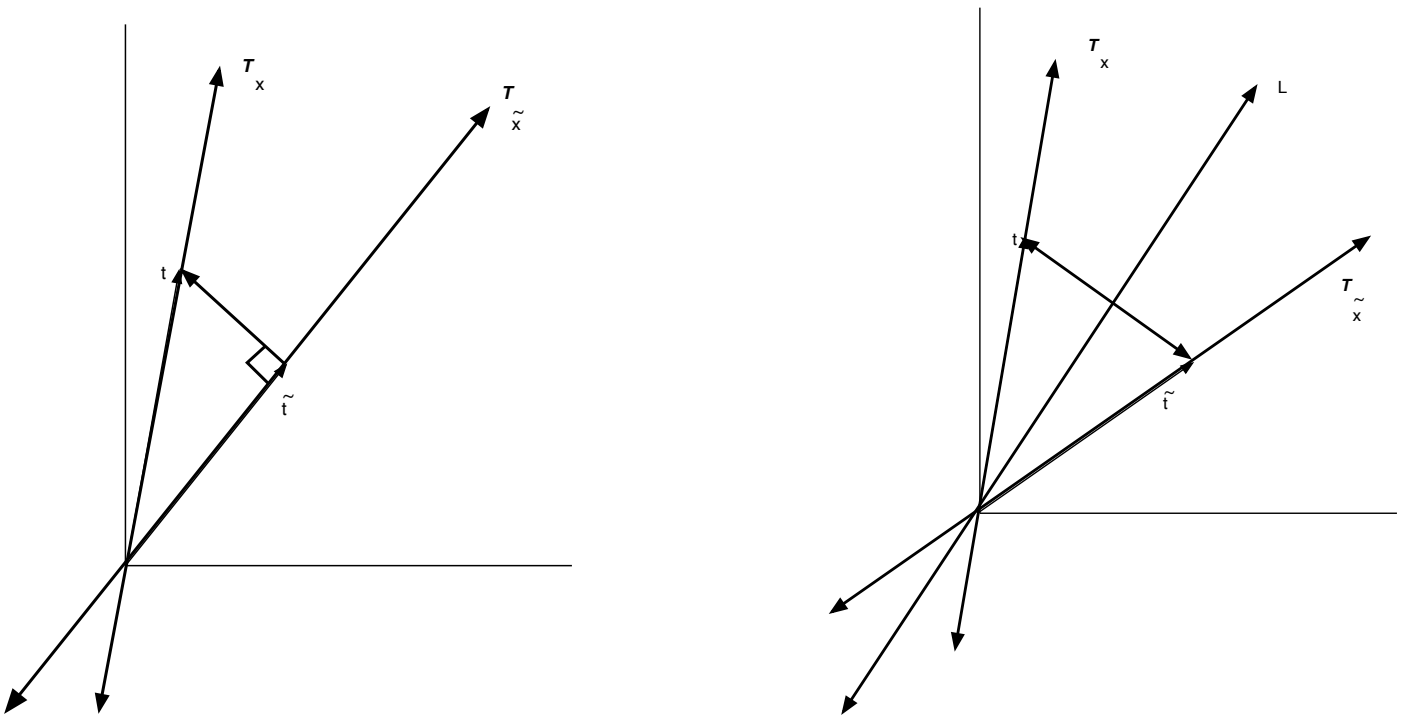


Figure: Orthogonal and oblique projections relating t and $\tilde{t}$

On the Unit Sphere S^{n-1}

- $\mathrm{Exp}(t\eta_x)$ only slightly more expensive than $R_x(t\eta_x)$
- Parallel transport only slightly more expensive than this implementation of this choice of vector transport
- key issue is therefore the effect on convergence of the use of vector transport and retraction

On compact Stiefel manifold $St(p, n)$

- $\mathrm{Exp}(t\eta_x)$ requires SVD/EVD that is expensive relative to QR decomposition-based $R_x(t\eta_x)$ as $p \to n$; only slightly more expensive when p is small or when a canonical basis-based $R_x(t\eta_x)$ is used.
- Parallel transport is much more expensive than most choices of vector transport; tolerance of integration of ODE is a key efficiency parameter.
- Vector transport efficiency is also a key consideration.
- key issue is therefore the effect on convergence of the use of vector transport and retraction

On the Unit Sphere S^{n-1}

Riemannian metric: $g(\xi, \eta) = \xi^T \eta$
The tangent space at x is:

$$T_x S^{n-1} = \{\xi \in \mathbb{R}^n \ : \ x^T \xi = 0\} = \{\xi \in \mathbb{R}^n : x^T \xi + \xi^T x = 0\}$$

Orthogonal projection to tangent space:

$$\mathsf{P}_x \xi_x = \xi - x x^T \xi_x$$

Retraction:

$$R_x(\eta_x) = (x + \eta_x)/\|(x + \eta_x)\|, \text{ where } \|\cdot\| \text{ denotes } \langle \cdot, \cdot \rangle^{1/2}$$

Exponential map:

$$\mathrm{Exp}(\eta_x) = x \cos(\|\eta_x\| t) + \frac{\eta_x}{\|\eta_x\|} \sin(\|\eta_x\| t)$$

Transport on the Unit Sphere S^{n-1}

Parallel Transport of $\xi \in T_x S^{n-1}$ along the geodesic from x in direction $\eta \in T_x S^{n-1}$:

$$P_{\gamma_\eta}^{t \leftarrow 0} \xi = \left(I_n + (\cos(\|\eta\| t) - 1) \frac{\eta \eta^T}{\|\eta\|^2} - \sin(\|\eta\| t) \frac{x \eta^T}{\|\eta\|} \right) \xi;$$

Vector Transport by orthogonal projection:

$$\mathcal{T}_{\eta_x} \xi_x = \left(I - \frac{(x + \eta_x)(x + \eta_x)^T}{\|x + \eta_x\|^2} \right) \xi_x$$

Inverse Vector Transport:

$$(\mathcal{T}_{\eta_x})^{-1}(\xi_{R_x(\eta_x)}) = \left(I - \frac{(x + \eta_x) x^T}{x^T (x + \eta_x)} \right) \xi_{R_x(\eta_x)}$$

Other vector transports possible.

Implementation on compact Stiefel manifold $St(p, n)$

View $St(p, n)$ as a Riemannian submanifold of $\mathbb{R}^{n \times p}$
Riemannian metric:

$$g(\xi, \eta) = \mathsf{tr}(\xi^T \eta)$$

The tangent space at X is:

$$T_X St(p, n) = \{Z \in \mathbb{R}^{n \times p} : X^T Z + Z^T X = 0\}.$$

Orthogonal projection to tangent space is :

$$\mathsf{P}_X \xi_X = (I - X X^T)\xi_X + X \mathsf{skew}(X^T \xi_X)$$

Retraction:

$$R_X(\eta_X) = \mathsf{qf}(X + \eta_X)$$

where $\mathsf{qf}(A) = Q \in \mathbb{R}_*^{n \times p}$, where $A = QR$

Parallel transport on Stiefel manifold Let $Y^TY = I_p$ and $A = Y^TH$ is skew-symmetric.
The geodesic from Y in direction H:

$$\gamma_H(t) = YM(t) + QN(t),$$

Q and R: the compact QR decomposition of $(I - YY^T)H$
$M(t)$ and $N(t)$ given by:

$$\begin{pmatrix} M(t) \\ N(t) \end{pmatrix} = \exp\left(t \begin{pmatrix} A & -R^T \\ R & 0 \end{pmatrix}\right) \begin{pmatrix} I_p \\ 0 \end{pmatrix}$$

Parallel transport on Stiefel manifold
The parallel transport of $\xi \neq H$ along the geodesic,$\gamma(t)$, from Y in direction H:

$$\begin{aligned} w(t) &= P_{\gamma}^{t \leftarrow 0} \xi \\ w'(t) &= -\frac{1}{2}\gamma(t)(\gamma'(t)^T w(t) + w(t)^T \gamma'(t)), w(0) = \xi \end{aligned}$$

In practice, the ODE is solved discretely.

Vector transport on $St(p,n)$ Projection based nonisometric vector transport:

$$\mathcal{T}_{\eta_X}\xi_X = (I - YY^T)\xi_X + Y\text{skew}(Y^T\xi_X), \text{ where } Y := R_X(\eta_X)$$

Inverse vector transport:

$$(\mathcal{T}_{\eta_X})^{-1}\xi_Y = \xi_Y + YS, \text{ where } Y := R_X(\eta_X)$$

S is symmetric matrix such that $X^T(\xi_Y + YS)$ is skew-symmetric.

Rayleigh quotient minimization on S^{n-1}
Cost function on S^{n-1}

$$f : S^{n-1} \to \mathbb{R} \ : x \mapsto x^T A x, A = A^T$$

Cost function embedded in $\mathbb{R}^n$

$$\bar{f} : \mathbb{R}^n \to \mathbb{R} : x \mapsto x^T A x, \text{ so that } f = \bar{f}\Big|_{S^{n-1}}$$

$$T_x S^{n-1} = \{\xi \in \mathbb{R}^n \ : \ x^T \xi = 0\}, \qquad R_x(\xi) \ = \frac{x+\xi}{\|x+\xi\|}$$

$$\mathrm{D}\bar{f}(x)[\zeta] = 2\zeta^T A x \to \operatorname{grad} \bar{f}(x) = 2Ax$$

Projection onto $T_x\mathbb{R}^n$: $\quad \mathrm{P}_x\xi = \xi - xx^T\xi$

Gradient: $\quad \operatorname{grad} f(x) = 2\mathrm{P}_x(Ax)$

Numerical Result for Rayleigh Quotient on S^{n-1}

- Problem sizes $n = 100$ and $n = 300$ with many different initial points.
- All versions of RBFGS converge superlinearly to local minimizer.
- Updating L and B^{-1} combined with Vector transport display similar convergence rates.
- Vector transport Approach 1 and Approach 2 display the same convergence rate, but Approach 2 takes more time due to complexity of each step.
- The updated B^{-1} of Approach 2 and Parallel transport has better conditioning, i.e. more positive definite.
- Vector transport versions converge as fast or faster than Parallel transport.

A Procrustes Problem on $\mathsf{St}(p, n)$

$$f : \mathsf{St}(p, n) \to \mathbb{R} : X \to \|AX - XB\|_F$$

where A: $n \times n$ matix, $B : p \times p$ matix, $X^T X = I_p$.

Cost function embedded in $\mathbb{R}^{n \times p}$:

$$\bar{f} : \mathbb{R}^{n \times p} \to \mathbb{R} : X \to \|AX - XB\|_F, \quad \text{with} \quad f = \bar{f}\big|_{St(p,n)}$$

$$\operatorname{grad} f(X) = \mathsf{P}_X \operatorname{grad} \bar{f}(X) = Q - X\mathsf{sym}(X^T Q), \quad \text{where}$$
$$Q := A^T AX - A^T XB - AXB^T + XBB^T.$$

$$\begin{aligned} \operatorname{Hess} f(X)[Z] &= \mathsf{P}_X \mathsf{D}\operatorname{grad} f(X)[Z] \\ &= \mathsf{D}\operatorname{grad} f(X)[Z] - X\mathsf{sym}(X^T \mathsf{D}\operatorname{grad} f(X)[Z]) \end{aligned}$$

Numerical Result for Procrustes on $\mathrm{St}(p, n)$

- Problem sizes $(n, p) = (7, 4)$ and $(n, p) = (12, 7)$ with many different initial points.
- All versions of RBFGS converge superlinearly to local minimizer.
- Updating L and B^{-1} combined with Vector transport display B^{-1} is slightly faster converging.
- Vector transport Approach 1 and Approach 2 display the same convergence rate, but Approach 2 takes more time due to complexity of each step.
- The updated B^{-1} of Approach 2 and Parallel transport has better conditioning, i.e. more positive definite.
- Vector transport versions converge noticably faster than Parallel transport. This depends on numerical evaluation of ODE for Parallel transport.

Vector transports on S^{n-1}

- NI: nonisometric vector transport by orthogonal projection onto the new tangent space (see above)
- CB: a vector transport relying on the canonical bases between the current and next subspaces
- CBE: a mathematically equivalent but computationally efficient form of CB
- QR: the basis in the new suspace is obtained by orthogonal projection of the previous basis followed by Gram-Schmidt.

Rayleigh quotient, $n = 300$

	NI	CB	CBE	QR
Time (sec.)	4.0	20	4.7	15.8
Iteration	97	92	92	97

RBFGS: Rayleigh quotient on S^{n-1} The vector transport property is crucial in achieving the desired results.

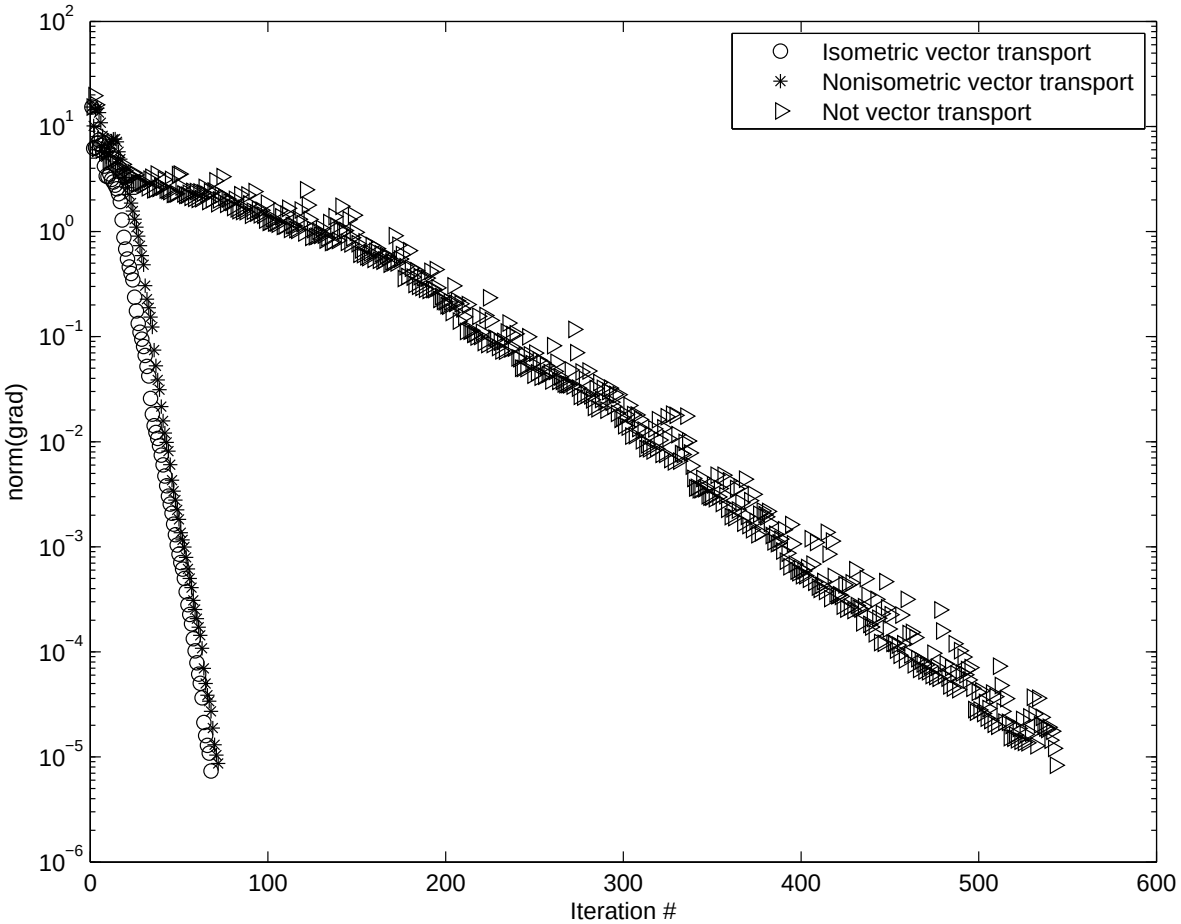


Figure: RBFGS with 3 transports for Rayleigh quotient. n=100.

RBFGS: Rayleigh quotient on S^{n-1} and Procrustes on $\mathrm{St}(p, n)$

Table: Vector transport vs. Parallel transport

	Rayleigh $n = 300$		Procrustes $(n, p) = (12, 7)$	
	Vector	Parallel	Vector	Parallel
Time (sec.)	4.0	4.2	24.0	304.0
Iteration	97	95	83	175

Procrustes Problem on $\mathrm{St}(p, n)$

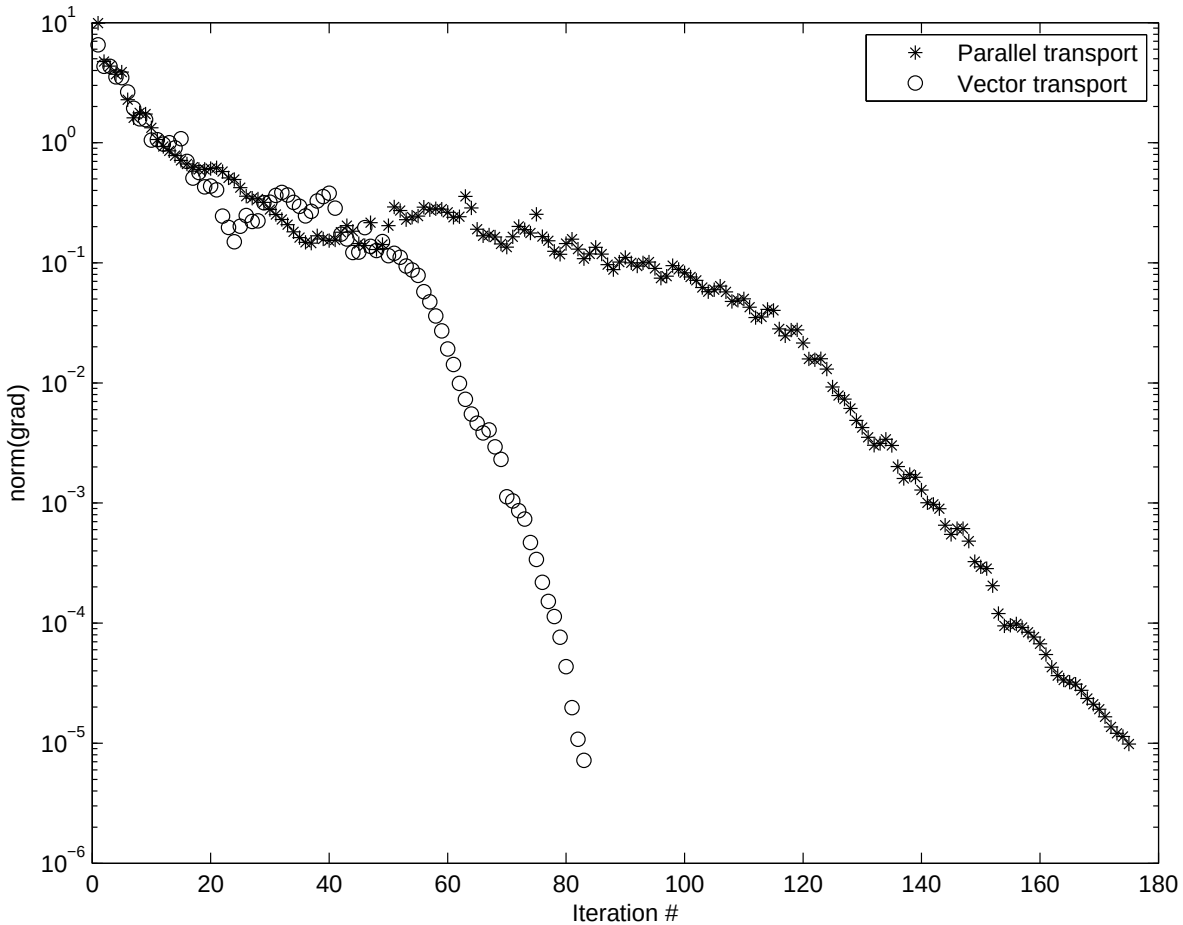


Figure: RBFGS parallel and vector transport for Procrustes. n=12, p=7.

RBFGS: Procrustes on $\mathrm{St}(p, n)$

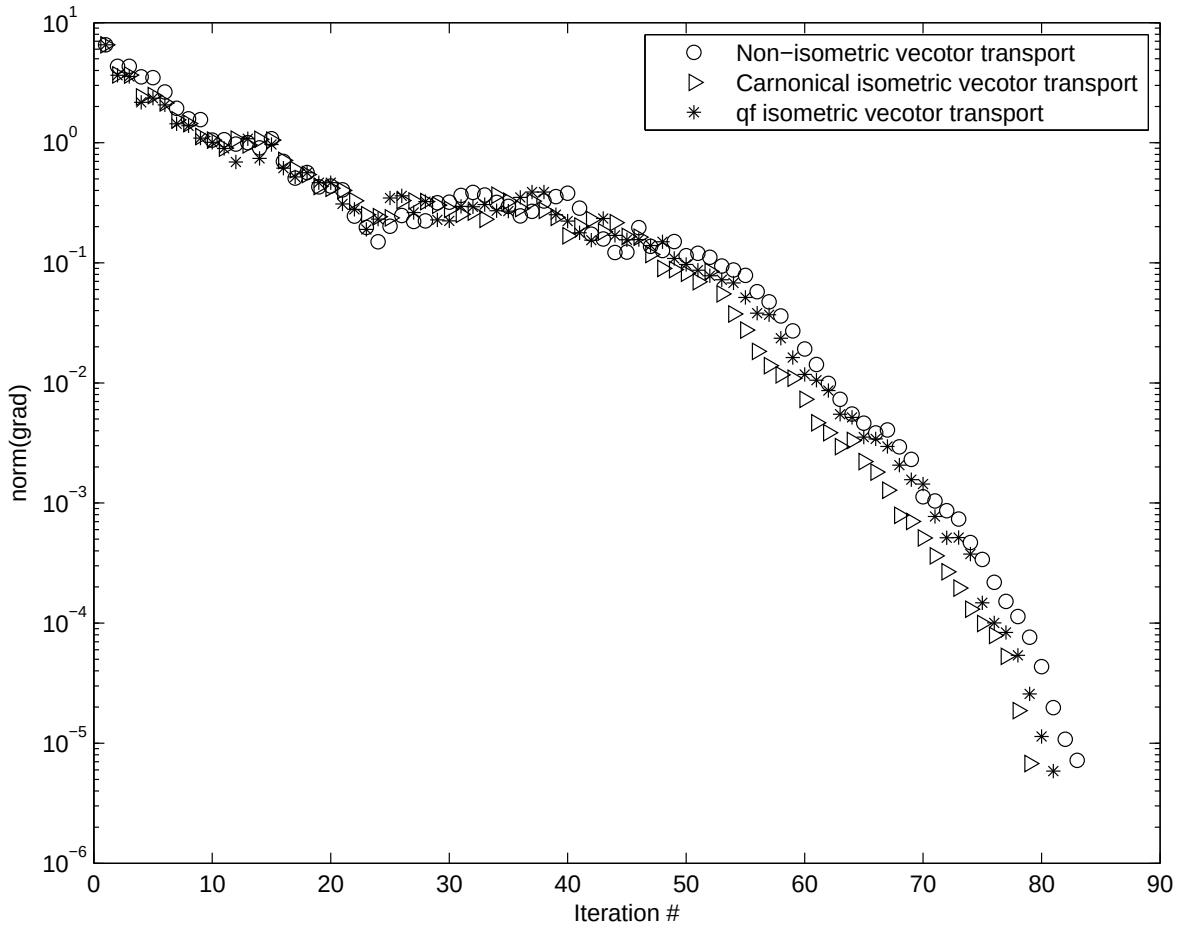


Figure: RBFGS using different vector transports for Procrustes. n=12, p=7.

RBFGS: Procrustes problem on $\mathrm{St}(p, n)$
Three vector transports have similar efficiencies and convergence rate.
Evidence that the nonisometric vector transport can converge effectively.

Table: Nonisometric vs. canonical isometric (SVD) vs. Isometric(qf)

	Procrustes $n = 12, p = 7$		
	Nonisometric	Canonical	Isometric(qf)
Time (sec.)	4.3	2.5	3.7
Iteration	83	79	81

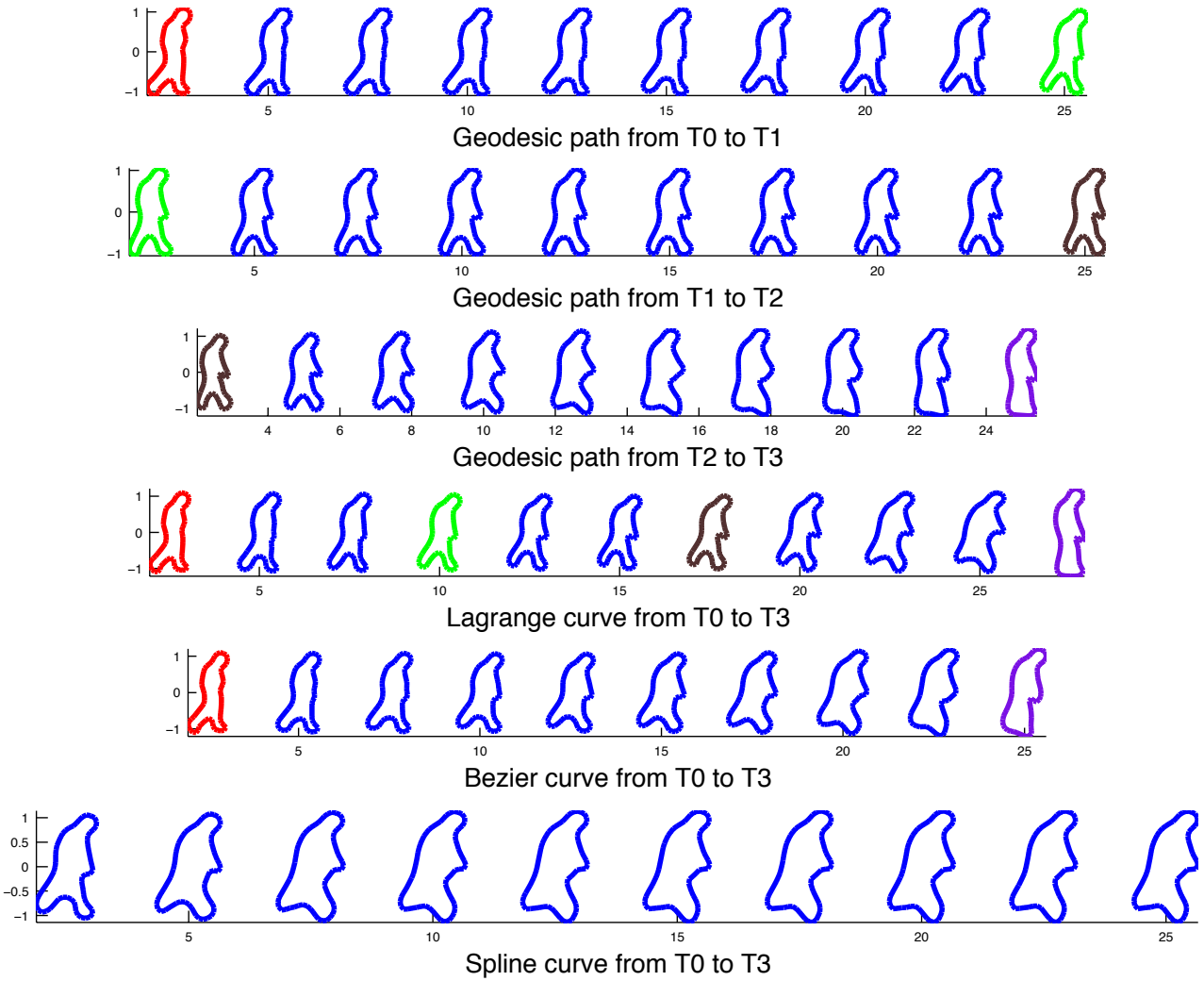
Application: Curve Fitting

- Shape manifold: landmarks and infinite dimensional forms
- Srivastava, Klassen, Gallivan (FSU), Absil and Van Dooren (UC Louvain), Samir (U Clermont-Ferrand)
- interpolation/fitting via geodesic ideas, e.g., Aitken interpolation, de Casteljau algorithm, generalizations for algorithms
- optimization problem (Leite and Machado)

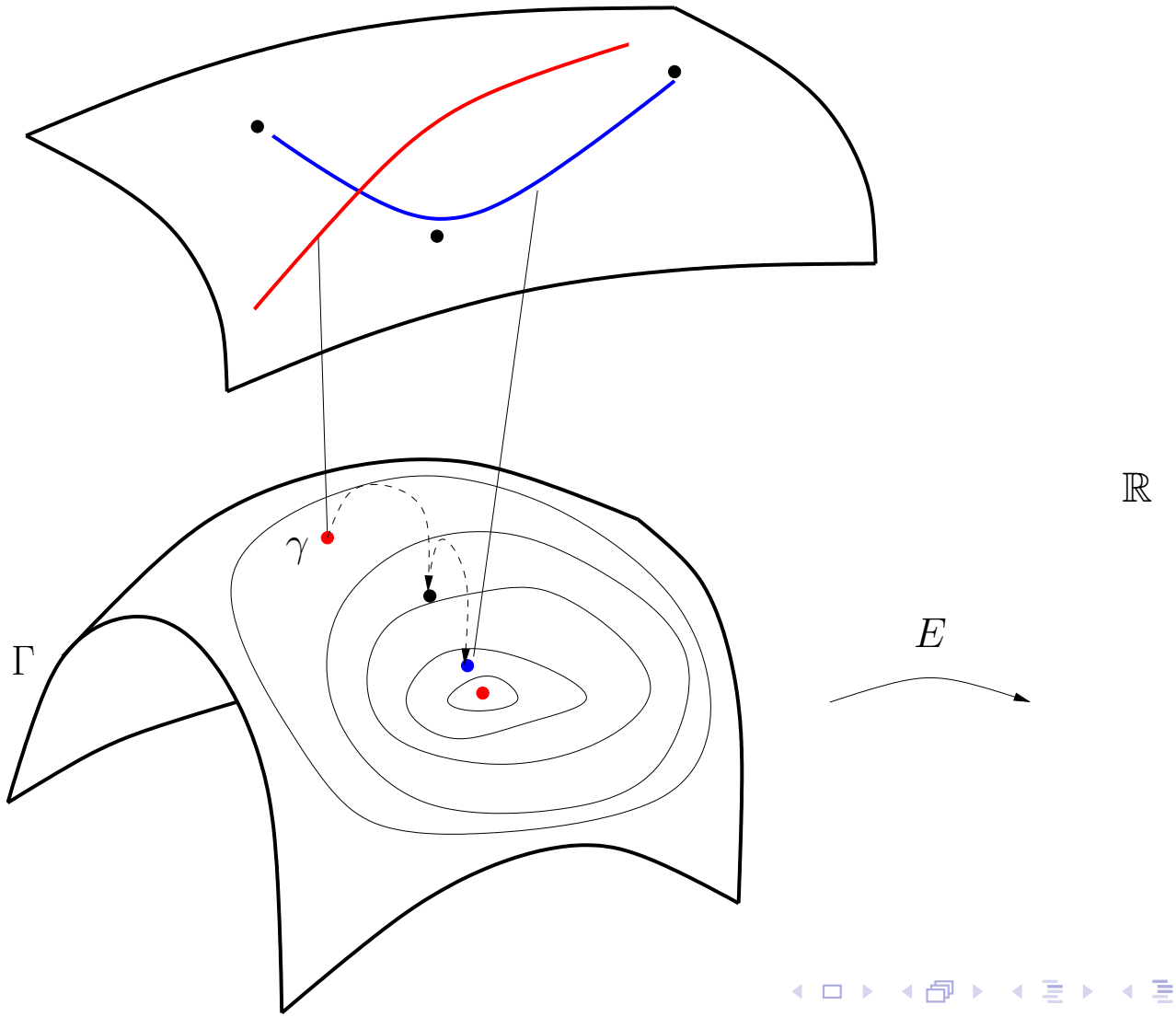
$$E_2 : \Gamma_2 \to \mathbb{R} : \gamma \mapsto E_2(\gamma) = E_{\mathrm{d}}(\gamma) + \lambda E_{\mathrm{s},2}(\gamma)$$
$$= \frac{1}{2}\sum_{i=0}^{N} d^2(\gamma(t_i), p_i) + \frac{\lambda}{2}\int_0^1 \langle \frac{\mathrm{D}^2\gamma}{\mathrm{d}\,t^2}, \frac{\mathrm{D}^2\gamma}{\mathrm{d}\,t^2} \rangle \,\mathrm{d}\,t, \tag{2}$$

where Γ_2 is a suitable set of curves on M.

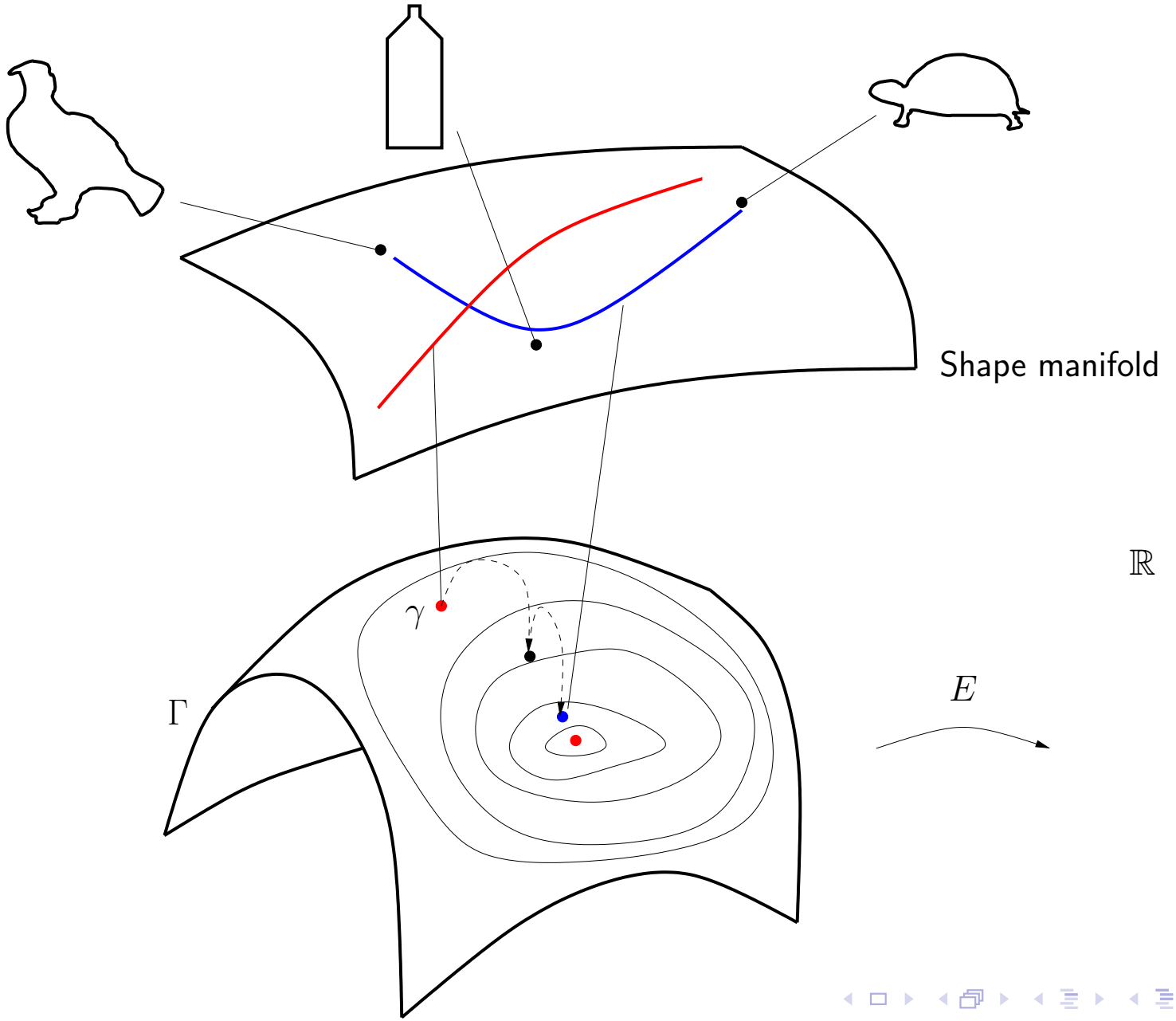
Fitting via de Casteljau-like algorithms



Curve fitting on manifolds



Curve fitting on shape manifold



Conclusions

Basic convergence theory complete for:

- Riemannian Trust Region
- Riemannian BFGS
- Riemannian Adaptive Cubic Overestimation

Software:

- Riemannian Trust Region for standard and large scale problems
- Riemannian BFGS for standard (current) and large scale problems (still needed)
- Riemannian Adaptive Cubic Overestimation (current)

Needed:

- Better understanding of the choice of retraction and transport relative to the structure of the cost function and convergence rate.
- Unified "computational" theory for retraction and transport.
- More application studies.