Optimization on matrix manifolds and application to image segmentation on the Stiefel manifold

Marco Sutti

Postdoctoral fellow at NCTS

MeDA Seminar

November 10, 2022

Overview

Paper: Optimization Methods on Riemannian Manifolds and Their Application to Shape Space, W. Ring and B. Wirth, SIAM J. Optim., 2012 22:2, 596–62.
$\rightsquigarrow$ Hereafter: [Ring/Wirth 2012].

Contributions:

- Convergence and convergence rates of BFGS quasi-Newton methods.
- Convergence and convergence rates of Fletcher–Reeves nonlinear CG.
- Numerical applications (image segmentation, truss shape deformations).

This talk:

I. Optimization on matrix manifolds, fundamental ideas and tools of Riemannian geometry that we use in optimization algorithms.

II. Riemannian BFGS, fundamental ideas [Ring/Wirth 2012, §3.1].

III. Application to image segmentation [Ring/Wirth 2012, §4.2].

I. Optimization on matrix manifolds

Optimization problems on matrix manifolds

- We can state the optimization problem as

$$\min_{x \in \mathcal{M}} f(x),$$

where $f: \mathcal{M} \to \mathbb{R}$ is the objective function and $\mathcal{M}$ is some matrix manifold.

- Matrix manifold: any manifold that is constructed from $\mathbb{R}^{n \times p}$ by taking either embedded submanifolds or quotient manifolds.
    - Examples of embedded submanifolds: orthogonal Stiefel manifold, manifold of symplectic matrices, manifold of fixed-rank matrices, …
    - Example of quotient manifold: the Grassmann manifold.

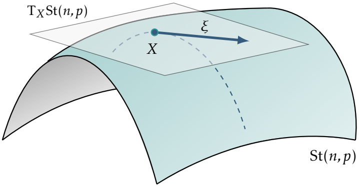
- Motivation: exploit the underlying geometric structure, take into account the constraints explicitly!

Manifold optimization: [Edelman et al. 1998, Absil et al. 2008, Boumal 2022], …

The Stiefel manifold and its tangent space

- Set of matrices with orthonormal columns:

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



- Tangent space to $\mathcal{M}$ at x: set of all tangent vectors to $\mathcal{M}$ at x, denoted $\mathrm{T}_x\mathcal{M}$. For $\mathrm{St}(n,p)$,

$$\mathrm{T}_X\mathrm{St}(n,p) = \{X\Omega + X_\perp K : \Omega = -\Omega^\top,\ K \in \mathbb{R}^{(n-p)\times p}\},$$

where $X_\perp \in \mathbb{R}^{n\times(n-p)}$ is orthonormal and $\mathrm{span}(X_\perp) = \big(\mathrm{span}(X)\big)^\perp$.

- Dimension: since $\dim\big(\mathrm{St}(n,p)\big) = \dim\big(\mathrm{T}_X\mathrm{St}(n,p)\big)$, we have

$$\dim(\mathrm{St}(n,p)) = \dim(\mathcal{S}_{\mathrm{skew}}) + \dim(\mathbb{R}^{(n-p)\times p}) = np - \tfrac{1}{2}p(p+1).$$

Stiefel manifold: [Stiefel 1935]

Riemannian manifold

A manifold $\mathcal{M}$ endowed with a smoothly-varying inner product (called Riemannian metric g) is called Riemannian manifold.

$\rightsquigarrow$ A couple $(\mathcal{M}, g)$, i.e., a manifold with a Riemannian metric on it.

$\rightsquigarrow$ For the Stiefel manifold:

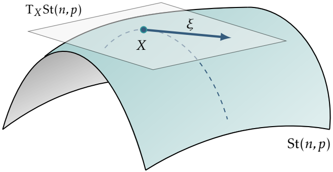
- Embedded metric inherited by $\mathrm{T}_X\mathrm{St}(n,p)$ from the embedding space $\mathbb{R}^{n\times p}$

$$\langle \xi, \eta \rangle = \mathrm{Tr}(\xi^\top \eta), \qquad \xi, \eta \in \mathrm{T}_X\mathrm{St}(n,p).$$

- Canonical metric by seeing $\mathrm{St}(n,p)$ as a quotient of the orthogonal group $\mathrm{O}(n)$: $\mathrm{St}(n,p) = \mathrm{O}(n)/\mathrm{O}(n-p)$

$$\langle \xi, \eta \rangle_\mathrm{c} = \mathrm{Tr}(\xi^\top (I - \tfrac{1}{2} X X^\top) \eta), \qquad \xi, \eta \in \mathrm{T}_X\mathrm{St}(n,p).$$

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



Embedded metric:

$$\langle \xi, \eta \rangle = \mathrm{Tr}(\xi^\top \eta).$$

Canonical metric:

$$\langle \xi, \eta \rangle_{\mathrm{c}} = \mathrm{Tr}(\xi^\top (I - \tfrac{1}{2} X X^\top) \eta).$$

Length of a tangent vector $\xi = X\Omega + X_\perp K$:

$$\|\xi\|_{\mathrm{F}} = \sqrt{\langle \xi, \xi \rangle} = \sqrt{\|\Omega\|_{\mathrm{F}}^2 + \|K\|_{\mathrm{F}}^2}. \qquad \|\xi\|_{\mathrm{c}} = \sqrt{\langle \xi, \xi \rangle_{\mathrm{c}}} = \sqrt{\tfrac{1}{2}\|\Omega\|_{\mathrm{F}}^2 + \|K\|_{\mathrm{F}}^2}.$$

Example for $p = 3$: $\quad \Omega = \begin{bmatrix} 0 & a & b \\ -a & 0 & c \\ -b & -c & 0 \end{bmatrix}, \quad$ then $\quad \|\Omega\|_{\mathrm{F}}^2 = 2a^2 + 2b^2 + 2c^2.$

Riemannian gradient

Let $f : \mathcal{M} \to \mathbb{R}$. E.g., the objective function in an optimization problem.

$\rightsquigarrow$ For any embedded submanifold:

- Riemannian gradient: projection onto $\mathrm{T}_X\mathcal{M}$ of the Euclidean gradient

$$\operatorname{grad} f(X) = \mathrm{P}_{\mathrm{T}_X\mathcal{M}}(\nabla f(X)).$$

$\rightsquigarrow$ For the Stiefel manifold, the orthogonal projection of a given matrix $M \in \mathbb{R}^{n\times p}$ onto the tangent space is

$$\mathrm{P}_{\mathrm{T}_X\mathrm{St}(n,p)}(M) = X\,\mathrm{skew}(X^\top M) + (I - XX^\top)M.$$

$\rightsquigarrow$ $\nabla f(X)$ is the Euclidean gradient of $f(X)$. E.g., for $f(X) = -\frac{1}{2}\operatorname{Tr}(X^\top A X)$, one has $\nabla f(X) = -AX$.

Matrix and vector calculus: The Matrix Cookbook, www.matrixcalculus.org, …

Riemannian exponential and logarithm

- Let $x \in \mathcal{M}$, $\xi \in \mathrm{T}_x\mathcal{M}$, and $\gamma(t)$ the geodesic such that $\gamma(0) = x$, $\dot{\gamma}(0) = \xi$. The exponential mapping $\mathrm{Exp}_x : \mathrm{T}_x\mathcal{M} \to \mathcal{M}$ is defined as $\mathrm{Exp}_x(\xi) := \gamma(1)$.
- Corollary: $\mathrm{Exp}_x(t\xi) := \gamma(t)$, for $t \in [0, 1]$.
- $\forall x, y \in \mathcal{M}$, the mapping $\mathrm{Exp}_x^{-1}(y) \in \mathrm{T}_x\mathcal{M}$ is called logarithm mapping.

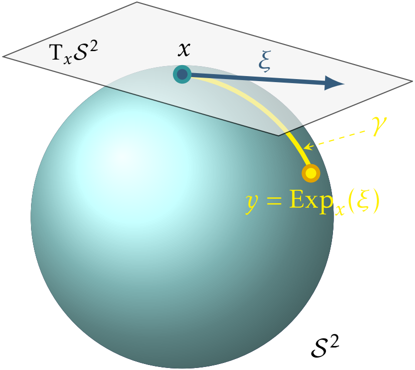
Example. Let $\mathcal{M} = \mathcal{S}^{n-1}$, then the exponential mapping at $x \in \mathcal{S}^{n-1}$ is

$$y = \mathrm{Exp}_x(\xi) = x\cos(\|\xi\|) + \frac{\xi}{\|\xi\|}\sin(\|\xi\|),$$

and the Riemannian logarithm is

$$\mathrm{Log}_x(y) = \xi = \arccos(x^\top y)\frac{\mathrm{P}_x y}{\|\mathrm{P}_x y\|},$$

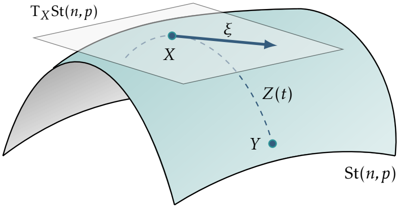
where $y \equiv \gamma(1)$ and P_x is the projector onto $\left(\mathrm{span}(x)\right)^\perp$, i.e., $\mathrm{P}_x = I - xx^\top$.



Riemannian exponential and logarithm on $\mathrm{St}(n,p)$

- Explicit expression (with the canonical metric) of the Riemannian exponential on the Stiefel manifold $\mathrm{St}(n,p)$:

$$Y = \mathrm{Exp}_X(\xi) = Z(1) = [\,X \;\; X_\perp\,]\exp\left(\begin{bmatrix} X^\top \xi & -(X_\perp^\top \xi)^\top \\ X_\perp^\top \xi & O \end{bmatrix}\right)\begin{bmatrix} I_p \\ O_{(n-p)\times p} \end{bmatrix}.$$



- Recall: there is no explicit expression for the Riemannian logarithm on the Stiefel manifold (see talk of Oct. 27, 2022).

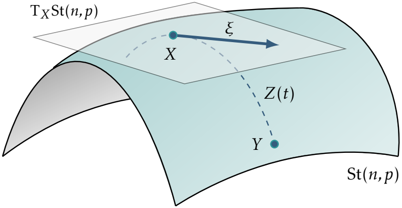
Riemannian distance

- Definition: given x, $y \in \mathcal{M}$, the Riemannian distance $\mathrm{dist}(x, y)$ is defined as

$$\mathrm{dist}(x,y) = \min_{\substack{\gamma:\,[0,1]\to\mathcal{M}\\ \gamma(0)=x,\,\gamma(1)=y}} L[\gamma], \quad \text{where} \quad L[\gamma] = \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}\,\mathrm{d}t.$$

- Property: given x, $y \in \mathcal{M}$, and $\xi \in \mathrm{T}_x\mathcal{M}$ such that $\mathrm{Exp}_x(\xi) = y$, the Riemannian distance $\mathrm{dist}(x, y)$ equals the length of $\xi \equiv \dot{\gamma}(0) \in \mathrm{T}_x\mathcal{M}$, i.e.,

$$\mathrm{dist}(x,y) = \|\xi\| = \sqrt{\langle \xi, \xi \rangle}.$$



Equivalent to: Compute the length of the Riemannian logarithm of y with base point x, i.e.,

$$\mathrm{Log}_x(y) = \xi.$$

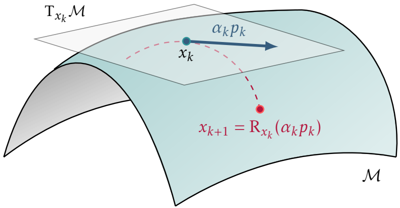
Line search on a manifold

- Recall (e.g., from here, §1.1): line-search methods in $\mathbb{R}^n$ are based on the update formula
$$x_{k+1} = x_k + \alpha_k p_k,$$
where $\alpha_k \in \mathbb{R}$ is the step size and $p_k \in \mathbb{R}^n$ is the search direction.

$\rightsquigarrow$ On nonlinear manifolds:

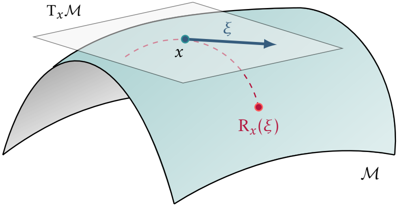
- p_k will be a tangent vector to $\mathcal{M}$ at x_k, i.e., $p_k \in \mathrm{T}_{x_k}\mathcal{M}$.
- Search along a curve in $\mathcal{M}$ whose tangent vector at $\alpha = 0$ is p_k.

$\rightsquigarrow$ Retraction.



Retractions/1

- Move in the direction of ξ while remaining constrained to $\mathcal{M}$.
- Smooth mapping $R_x : T_x\mathcal{M} \to \mathcal{M}$ with a local condition that preserves gradients at x.



- The Riemannian exponential mapping is also a retraction, but it is not computationally efficient.
- Retractions: first-order approximation of the Riemannian exponential!

Constructing retractions: [Absil/Malick 2012]

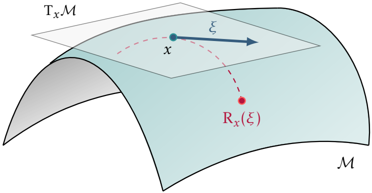
Retractions/2

Properties:

(i) $R_x(0_x) = x$, where 0_x is the zero element of $T_x\mathcal{M}$.

(ii) With the identification $T_{0_x} T_x\mathcal{M} \simeq T_x\mathcal{M}$, R_x satisfies the local rigidity condition

$$DR_x(0_x) = \mathrm{id}_{T_x\mathcal{M}}.$$



Two main purposes:

- Turn points of $T_x\mathcal{M}$ into points of $\mathcal{M}$.
- Transform cost functions $f : \mathcal{M} \to \mathbb{R}$ defined in a neighborhood of $x \in \mathcal{M}$ into cost functions $f_{R_x} := f \circ R_x$ defined on the vector space $T_x\mathcal{M}$.

Retractions on embedded submanifolds

Let $\mathcal{M}$ be an embedded submanifold of a vector space $\mathcal{E}$. Thus $\mathrm{T}_x\mathcal{M}$ is a linear subspace of $\mathrm{T}_x\mathcal{E} \simeq \mathcal{E}$. Since $x \in \mathcal{M} \subseteq \mathcal{E}$ and $\xi \in \mathrm{T}_x\mathcal{M} \subseteq \mathrm{T}_x\mathcal{E} \simeq \mathcal{E}$, with little abuse of notation we write $x + \xi \in \mathcal{E}$.

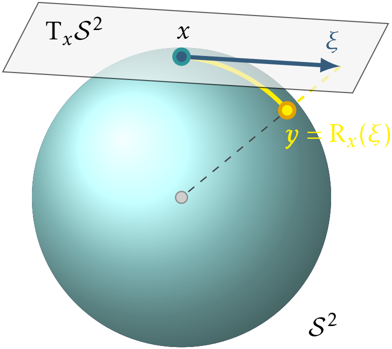
$\rightsquigarrow$ General recipe to define a retraction $\mathrm{R}_x(\xi)$ for embedded submanifolds:

- Move along ξ to get to $x + \xi$ in $\mathcal{E}$.
- Map $x + \xi$ back to $\mathcal{M}$. For matrix manifolds, use matrix decompositions.

Example. Let $\mathcal{M} = \mathcal{S}^{n-1}$, then the retraction at $x \in \mathcal{S}^{n-1}$ is

$$\mathrm{R}_x(\xi) = \frac{x + \xi}{\|x + \xi\|},$$

defined for all $\xi \in \mathrm{T}_x\mathcal{S}^{n-1}$. $\mathrm{R}_x(\xi)$ is the point on $\mathcal{S}^{n-1}$ that minimizes the distance to $x + \xi$.



Retractions on the Stiefel manifold

⇝ Based on matrix decompositions: given a generic matrix $A \in \mathbb{R}_*^{n\times p}$,

- Polar decomposition (∼ polar form of a complex number):

$$A = UP, \quad \text{with} \quad U \in \mathrm{St}(n,p), \quad P \in \mathcal{S}_{\mathrm{sym}^+}(p).$$

- QR factorization (∼ Gram−Schmidt algorithm):

$$A = QR, \quad \text{with} \quad Q \in \mathrm{St}(n,p), \quad R \in \mathcal{S}_{\mathrm{upp}^+}(p).$$

Let $X \in \mathrm{St}(n,p)$ and $\xi \in \mathrm{T}_X\mathrm{St}(n,p)$.

⇝ Retraction based on the polar decomposition:

$$\mathrm{R}_X(\xi) = (X+\xi)(I+\xi^\top\xi)^{-1/2}.$$

⇝ Retraction based on the QR factorization:

$$\mathrm{R}_X(\xi) = \mathrm{qf}(X+\xi),$$

where $\mathrm{qf}(A)$ denotes the Q factor of the QR factorization.

Line search on a manifold (reprise)

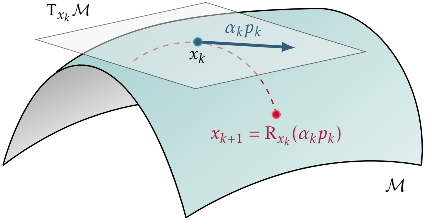
Line-search methods on manifolds are based on the update formula

$$x_{k+1} = \mathrm{R}_{x_k}(\alpha_k p_k),$$

where $\alpha_k \in \mathbb{R}$ and $p_k \in \mathrm{T}_{x_k}\mathcal{M}$.

Recipe for constructing a line-search method on a manifold:

- Choose a retraction R_{x_k}.
- Select a search direction p_k.
- Select a step length α_k (e.g., by using the Armijo condition).

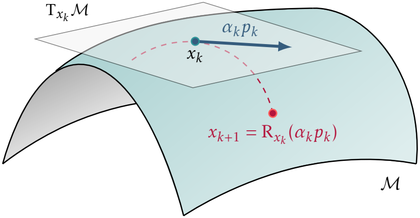


Remark: If $p_k = -\operatorname{grad} f(x_k)$, we get the Riemannian steepest descent.

Line search on a manifold (reprise)

Algorithm 1: Line-search minimization on manifolds.

Given $f : \mathcal{M} \to \mathbb{R}$, starting point $x_0 \in \mathcal{M}$;
$k \leftarrow 0$;
repeat
  choose a descent direction $p_k \in \mathrm{T}_{x_k}\mathcal{M}$;
  choose a retraction $\mathrm{R}_{x_k} : \mathrm{T}_{x_k}\mathcal{M} \to \mathcal{M}$;
  choose a step length $\alpha_k \in \mathbb{R}$;
  set $x_{k+1} = \mathrm{R}_{x_k}(\alpha_k p_k)$;
  $k \leftarrow k + 1$;
until x_{k+1} *sufficiently minimizes* f;



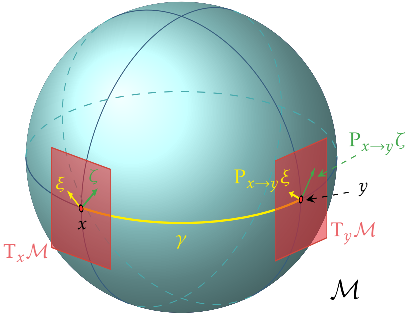
Parallel transport

- Given a Riemannian manifold $(\mathcal{M}, g)$ and $x,\ y \in \mathcal{M}$, the parallel transport $\mathrm{P}_{x\to y} : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_y\mathcal{M}$ is a linear operator that preserves the inner product:

$$\forall \xi, \zeta \in \mathrm{T}_x\mathcal{M}, \qquad \langle \mathrm{P}_{x\to y}\,\xi, \mathrm{P}_{x\to y}\,\zeta\rangle_y = \langle \xi, \zeta\rangle_x.$$

⚠ Caveat:

- Computing parallel transports, in general, requires numerically solving ODEs.
- One needs to choose a curve connecting x and y explicitly. If we choose a minimizing geodesic, this requires computing the Riemannian logarithm.



$\rightsquigarrow$ Computing the parallel transport might be too expensive in practice!

⚠ Remark: parallel transport with the Levi-Civita connection.
If we use other connections, we get different properties.

Transporters

- Transporter: "poor's man version of parallel transport".
- No need for a Riemannian connection. If x and y are close enough to one another, then one can define the linear map $\mathrm{T}_{y\leftarrow x}\colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_y\mathcal{M}$, with $\mathrm{T}_{x\leftarrow x}$ being the identity map.
- Useful in defining a Riemannian version of the classical BFGS algorithm.
- The differentials of a retraction provide a transporter via $\mathrm{T}_{y\leftarrow x} = \mathrm{DR}_x(v)$, where $v = \mathrm{R}_x^{-1}(y)$ [Boumal 2022, Prop. 10.64].
- For embedded submanifolds of a Euclidean space $\mathcal{E}$, a transporter can be defined as [Boumal 2022, Prop. 10.66]

$$\mathrm{T}_{y\leftarrow x} = \mathrm{P}_{\mathrm{T}_y\mathcal{M}}\Big|_{\mathrm{T}_x\mathcal{M}},$$

where $\mathrm{P}_{\mathrm{T}_y\mathcal{M}}$ is the orthogonal projector from $\mathcal{E}$ to $\mathrm{T}_y\mathcal{M}$, restricted to $\mathrm{T}_x\mathcal{M}$.

Transporters: [Boumal 2022, §10.5]

II. Riemannian BFGS

(§3.1)

Riemannian BFGS quasi-Newton method

- Fundamental idea of quasi-Newton methods: instead of computing the approximate Hessian B_k from scratch at every iteration, we update it by using the newest information gained during the last iteration.
- The search direction p_k is chosen as the solution to

$$B_k(p_k, \cdot) = -\mathrm{D} f(x_k),$$

where $B_k \colon \mathrm{T}_x\mathcal{M} \times \mathrm{T}_x\mathcal{M} \to \mathbb{R}$ is updated according to

$$s_k = \alpha_k p_k = \mathrm{R}_{x_k}^{-1}(x_{k+1}), \qquad y_k = \mathrm{D} f_{\mathrm{R}_{x_k}}(s_k) - \mathrm{D} f_{\mathrm{R}_{x_k}}(0),$$

$$B_{k+1}(\mathrm{T}_k v, \mathrm{T}_k w) = B_k(v, w) - \frac{B_k(s_k, v) B_k(s_k, w)}{B_k(s_k, s_k)} + \frac{(y_k v)(y_k w)}{y_k s_k},$$

$\forall v, w \in \mathrm{T}_{x_k}\mathcal{M}$. Here, $\mathrm{T}_k \equiv \mathrm{T}_{x_k, x_{k+1}}$ denotes a transporter $\mathrm{T}_{x_k}\mathcal{M} \to \mathrm{T}_{x_{k+1}}\mathcal{M}$.

Riemannian BFGS: [Gabay 1982, Brace/Manton 2006, Qi/Gallivan/Absil 2010, Ring/Wirth 2012, Huang/Gallivan/Absil 2015, Huang/Absil/Gallivan 2016]

Euclidean BFGS vs Riemannian BFGS

Euclidean BFGS (see notes, §1.1)	Riemannian BFGS
$s_k = x_{k+1} - x_k,$	$s_k = \mathrm{R}_{x_k}^{-1}(x_{k+1}),$
$y_k = \nabla f_{k+1} - \nabla f_k,$	$y_k = \mathrm{D} f_{\mathrm{R}_{x_k}}(s_k) - \mathrm{D} f_{\mathrm{R}_{x_k}}(0),$
$B_{k+1} s_k = y_k,$	$B_{k+1}(\mathrm{T}_k s_k, \cdot) = y_k \mathrm{T}_k^{-1},$
$B_{k+1} = B_k - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k} + \frac{y_k y_k^\top}{y_k^\top s_k}.$	$B_{k+1}(\mathrm{T}_k v, \mathrm{T}_k w) = B_k(v, w) - \frac{B_k(s_k, v) B_k(s_k, w)}{B_k(s_k, s_k)} + \frac{(y_k v)(y_k w)}{y_k s_k}.$

Convergence and convergence rates of Riemannian BFGS

- Convergence of BFGS to the optimal value $f(x^*)$ [Prop. 10]:

$$f(x_k) - f(x^*) \leq \mu^{k+1}\left(f(x_0) - f(x^*)\right).$$

- Convergence of the iterates of BFGS to x^* [Cor. 11]:

$$\operatorname{dist}(x_k, x^*) \leq \sqrt{\frac{M}{m}}\sqrt{\mu}^{k+1}\operatorname{dist}(x_0, x^*).$$

- Convergence rate of BFGS [Cor. 13]: superlinear convergence, i.e.,

$$\lim_{k\to\infty}\frac{\operatorname{dist}(x_{k+1}, x^*)}{\operatorname{dist}(x_k, x^*)} = 0.$$

Compare with:

Riemannian steepest descent [Boumal 2022, Thm. 4.20] gives assumptions for the iterates x_k to converge to a local minimizer x^* at least linearly.

Riemannian Newton's method

$$\lim_{k\to\infty}\frac{\operatorname{dist}(x_{k+1}, x^*)}{\operatorname{dist}^2(x_k, x^*)} \leq C.$$

[Ring/Wirth 2012, Prop. 7]

III. Application to image segmentation on the Stiefel manifold

(§4.2)

Space of smooth closed curves/1

- Riemannian optimization in the space of smooth closed curves (§4.2).
- Younes et al. represent a curve $c\colon [0,1] \to \mathbb{C} \equiv \mathbb{R}^2$ by two functions $e,\ g\colon [0,1] \to \mathbb{R}$ via

$$c(\theta) = c(0) + \frac{1}{2} \int_0^\theta (e + \mathrm{i}\, g)^2 \, \mathrm{d}\theta.$$

- Conditions: closed $c(1) = c(0)$, and of unit length, $\int_0^1 |c'(\theta)| \, \mathrm{d}\theta = 1$.
  $\rightsquigarrow$ e and g orthonormal in $L^2([0,1])$, thus (e, g) is an element of

$$\mathrm{St}\Big(L^2([0,1]), 2\Big) = \Big\{(e, g) \in L^2([0,1])\colon\ \|e\|_{L^2([0,1])} = \|g\|_{L^2([0,1])} = 1,\ (e, g)_{L^2([0,1])} = 0\Big\}.$$

Recall the inner product in $L^2([0,1])$: $\quad (e, g)_{L^2([0,1])} \coloneqq \int_0^1 e \cdot \bar{g} \, \mathrm{d}x,$

and the induced norm $\quad \|e\|_{L^2([0,1])} \coloneqq \sqrt{\int_0^1 |e(x)|^2 \, \mathrm{d}x}.$

[Younes/Michor/Shah/Mumford 2008]

Space of smooth closed curves/2

- Sundaramoorthi et al. represent a general closed curve c by an element $(c_0,\ \rho,\ (e,g))$ of $\mathbb{R}^2 \times \mathbb{R} \times \mathrm{St}\left(L^2([0,1]),\,2\right)$ via

$$c(\theta) = c_0 + \frac{\exp\rho}{2}\int_0^\theta (e + \mathrm{i}\,g)^2\,\mathrm{d}\theta.$$

  where c_0 is the curve centroid and $\exp\rho$ its length.

- Metric:

$$g_{[c]}(h,k) = h^t \cdot k^t + \lambda_\ell h^\ell k^\ell + \lambda_d \int_{[c]} \frac{\mathrm{d}h^d}{\mathrm{d}s} \cdot \frac{\mathrm{d}k^d}{\mathrm{d}s}\,\mathrm{d}s$$

  on the tangent space of curve variations $h,\ k\colon [c] \to \mathbb{R}^2$, where $[c]$ is the image of $c\colon [0,1] \to \mathbb{R}^2$, s denotes arclength, and weights $\lambda_\ell,\ \lambda_d > 0$.

- There is a closed formula for the exponential map [Sundaramoorthi et al. 2011].

[Sundaramoorthi/Mennucci/Soatto/Yezzi 2011]

Objective functional/1

Given a gray scale image $u\colon [0,1]^2 \to \mathbb{R}$, we would like to minimize the objective functional

$$f([c]) = a_1 \left(\int_{\mathrm{int}[c]} (u_i - u)^2 \,\mathrm{d}x + \int_{\mathrm{ext}[c]} (u_e - u)^2 \,\mathrm{d}x \right) + a_2 \int_{[c]} \mathrm{d}s,$$

where $a_1,\ a_2 > 0$, u_i and u_e are given gray values, and $\mathrm{int}[c]$ and $\mathrm{ext}[c]$ denote the interior and exterior of $[c]$.

Meaning:

- First two terms: indicate that $[c]$ should enclose the image region where u is close to u_i and far from u_e.
- Third term: acts as a regularizer and measures the curve length.

Image segmentation via active contours without edges: [Chan/Vese 2001]
Chan-Vese Segmentation in scikit-image

Objective functional/2

We interpret the curve c as an element of the manifold $\mathbb{R}^2 \times \mathbb{R} \times \mathrm{St}\Big(L^2([0,1]), 2\Big)$ and add a term that prefers a uniform curve parametrization:

$$f(c_0,\ \rho,\ (e,g)) = a_1 \left(\int_{\mathrm{int}[(c_0,\ \rho,\ (e,g))]} (u_i - u)^2 \,\mathrm{d}x + \int_{\mathrm{ext}[(c_0,\ \rho,\ (e,g))]} (u_e - u)^2 \,\mathrm{d}x \right) + a_2 \exp(\rho) + a_3 \int_0^1 (e^2 + g^2)^2 \,\mathrm{d}\theta,$$

Numerical implementation:

- e and g are discretized as piecewise constant functions on a uniform grid over $[0,1]$.
- The image u is given as pixel values on a uniform grid.

Numerical experiments/1

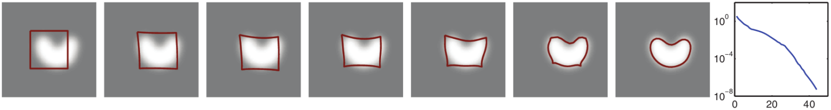


Fig. 3. *Curve evolution during BFGS minimization of f. The curve is depicted at steps 0, 1, 2, 3, 4, 7 and after convergence. Additionally we show the evolution of the function value $f(c_k) - \min_c f(c)$.*

Table 2

Iteration numbers for minimization of f with different methods. The iteration is stopped as soon as the derivative of the discretized functional f has ℓ^2-norm less than 10^{-3}. For the gradient flow discretization we employ a step size of 0.001, which is roughly the largest step size for which the curve stays within the image domain during the whole iteration.

	Nongeodesic retraction	Geodesic retraction
Gradient flow	4207	4207
Gradient descent	1076	1064
BFGS quasi-Newton	44	45
Fletcher–Reeves NCG	134	220

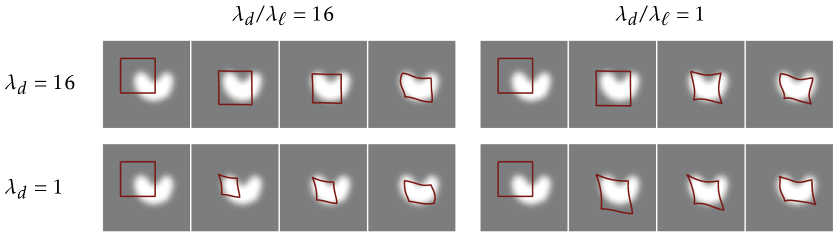
- Right column: geodesic retractions based on the matrix exponential [Sundaramoorthi et al. 2011].

Numerical experiments/2

- Experiments for different weights λ_ℓ and λ_d inside the metric

$$g_{[c]}(h,k) = h^t \cdot k^t + \lambda_\ell h^\ell k^\ell + \lambda_d \int_{[c]} \frac{\mathrm{d}h^d}{\mathrm{d}s} \cdot \frac{\mathrm{d}k^d}{\mathrm{d}s}\, \mathrm{d}s.$$

- A larger λ_d (top row) ensures a good curve positioning and scaling before starting major deformations. A small λ_d has a reverse effect (bottom row).
- The ratio between λ_d and λ_d/λ_ℓ decides whether the scaling or the positioning is adjusted first.



Numerical experiments/3

- Active contour segmentation on the widely used cameraman image.
- The iteration was stopped as soon as the derivative of the discretized objective functional f reached an ℓ^2-norm less than 10^{-2}.
- In the top row, BFGS needed 46 steps, while gradient descent needed 8325 steps.

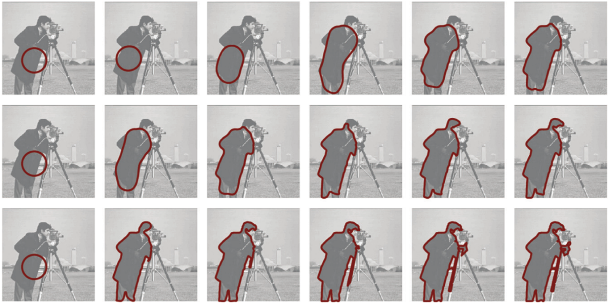


FIG. 5. *Segmentation of the cameraman image with different parameters (using the BFGS iteration and* $\lambda_l = \lambda_d = 1$*). Top:* $(a_1, a_2, a_3) = (50, 3 \cdot 10^{-1}, 10^{-3})$*, steps* 0, 1, 5, 10, 20, 46 *are shown. Middle:* $(a_1, a_2, a_3) = (50, 8 \cdot 10^{-2}, 10^{-3})$*, steps* 0, 10, 20, 40, 60, 116 *are shown. Bottom:* $(a_1, a_2, a_3) = (50, 10^{-2}, 10^{-3})$*, steps* 0, 50, 100, 150, 200, 250 *are shown. The curves were reparameterized every* 70 *steps. The bottom iteration was stopped as soon as the curve self-intersected.*

Conclusions

Pros and cons:

- ⊕ Solid, quite well-understood mathematical theory behind.
- ⊖ Cannot deal with self-intersecting curves.

This talk:

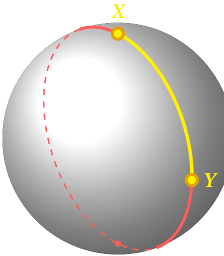
- Fundamental ideas and tools of Riemannian geometry that we use in optimization on Riemannian manifolds.
- Riemannian BFGS [Ring/Wirth 2012, §3.1].
- Application to image segmentation [Ring/Wirth 2012, §4.2].

⇝ Download slides: marcosutti.net/research.html#talks

IV. Bonus material

Geodesics

- Generalization of straight lines to manifolds.
- Locally they are curves of shortest length, but globally they may not be.
- In general, they are defined as critical points of the length functional $L[\gamma]$, and may or may not be minima.



- The fundamental Hopf–Rinow theorem guarantees the existence of a length-minimizing geodesic connecting any two given points.

Hopf–Rinow Theorem

Theorem ([Hopf/Rinow]) Let $(\mathcal{M}, g)$ be a (connected) Riemannian manifold. Then the following conditions are equivalent:

1. Closed and bounded subsets of $\mathcal{M}$ are compact;
2. $(\mathcal{M}, g)$ is a complete metric space;
3. $(\mathcal{M}, g)$ is geodesically complete, i.e., for any $x \in \mathcal{M}$, the exponential map Exp_x is defined on the entire tangent space $\mathrm{T}_x\mathcal{M}$.

Any of the above implies that given any two points $x, y \in \mathcal{M}$, there exists a length-minimizing geodesic connecting these two points.

The Stiefel manifold is compact/complete/geodesically complete.

$\rightsquigarrow$ Length-minimizing geodesics exist.

Riemannian Geometry, Sakai 1992

The orthogonal group as a special case of $\mathrm{St}(n, p)$

- If $p = n$, then the Stiefel manifold reduces to the orthogonal group

$$\mathrm{O}(n) = \{X \in \mathbb{R}^{n \times n} \colon X^\top X = I_n\},$$

and the tangent space at X is given by

$$\mathrm{T}_X\mathrm{O}(n) = \{X\Omega \colon \Omega^\top = -\Omega\} = X\mathcal{S}_{\mathrm{skew}}(n).$$

- Furthermore, at $X = I_n$, we have $\mathrm{T}_{I_n}\mathrm{O}(n) = \mathcal{S}_{\mathrm{skew}}(n)$, i.e., the tangent space to $\mathrm{O}(n)$ at the identity matrix I_n is the set of skew-symmetric n-by-n matrices $\mathcal{S}_{\mathrm{skew}}(n)$. In the language of Lie groups, we say that $\mathcal{S}_{\mathrm{skew}}(n)$ is the Lie algebra of the Lie group $\mathrm{O}(n)$.

An analogy

Theory:	$\rightsquigarrow$	Algorithm:
Riemannian exponential	$\rightsquigarrow$	Retractions
Parallel transport	$\rightsquigarrow$	Transporters