Optimization On Manifolds

Pierre-Antoine Absil
Robert Mahony
Rodolphe Sepulchre

Based on ‘‘Optimization Algorithms on Matrix Manifolds’’, Princeton University Press, January 2008

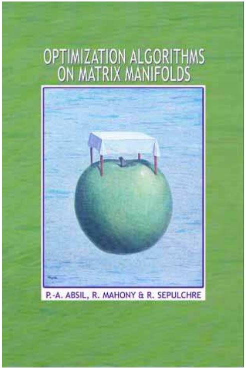
Compiled on February 12, 2011

Outline

Collaborations

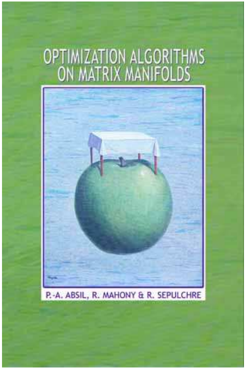
- Chris Baker (Oak Ridge National Laboratory)
- Kyle Gallivan (Florida State University)
- Paul Van Dooren (Université catholique de Louvain)
- Several other colleagues mentioned later on

Reference



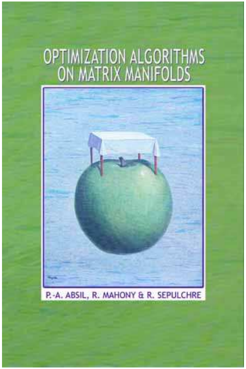
Optimization Algorithms on Matrix Manifolds
P.-A. Absil, R. Mahony, R. Sepulchre
Princeton University Press, January 2008

About the reference



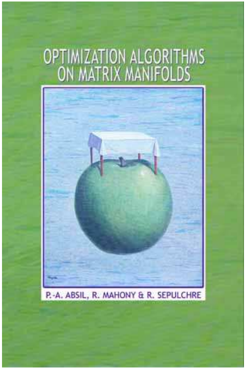
- The publisher, Princeton University Press, has been a non-profit company since 1910.
- PDF version of book chapters available on the publisher's web site.

Reference: contents



1. Introduction
2. Motivation and applications
3. Matrix manifolds: first-order geometry
4. Line-search algorithms
5. Matrix manifolds: second-order geometry
6. Newton's method
7. Trust-region methods
8. A constellation of superlinear algorithms

Matrix Manifolds: first-order geometry



Chap 3: Matrix Manifolds: first-order geometry

1. Charts, atlases, manifolds
2. Differentiable functions
3. Embedded submanifolds
4. Quotient manifolds
5. Tangent vectors and differential maps
6. Riemannian metric, distance, gradient

Smooth optimization in $\mathbb{R}^n$

General unconstrained optimization problem in $\mathbb{R}^n$:
Let

$$f : \mathbb{R}^n \to \mathbb{R},$$

The real-valued function f is termed the *cost function* or *objective function*.
Problem: find $x_* \in \mathbb{R}^n$ such that there exists $\epsilon > 0$ for which

$$f(x) \geq f(x_*) \text{ whenever } \|x - x_*\| < \epsilon.$$

Such a point x_* is called a *local minimizer* of f.

Smooth optimization in $\mathbb{R}^n$

General unconstrained optimization problem in $\mathbb{R}^n$:
Let

$$f : \mathbb{R}^n \to \mathbb{R},$$

The real-valued function f is termed the *cost function* or *objective function*.
Problem: find $x_* \in \mathbb{R}^n$ such that there exists a neighborhood $\mathcal{N}$ of x_* such that

$$f(x) \geq f(x_*) \text{ whenever } x \in \mathcal{N}.$$

Such a point x_* is called a *local minimizer* of f.

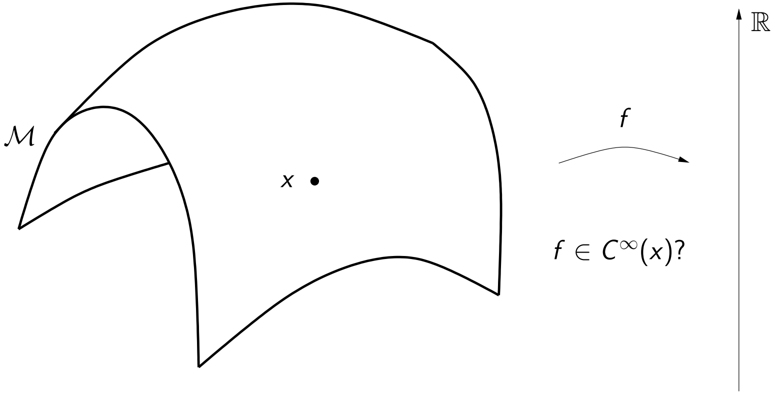
Smooth optimization *beyond* $\mathbb{R}^n$

$$? \ \arg\min_{x \in \mathbb{R}^n} f(x)$$

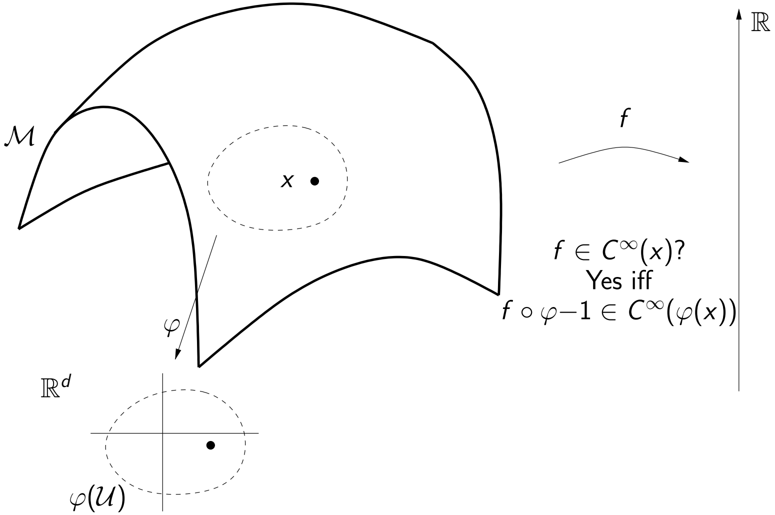
- Several optimization techniques require the cost function to be differentiable to some degree:
  - Steepest-descent at x requires $\mathrm{D}f(x)$.
  - Newton's method at x requires $\mathrm{D}^2 f(x)$.
- Can we go beyond $\mathbb{R}^n$ without losing the concept of differentiability?

$$\arg\min_{x \in \mathbb{R}^n} f(x) \qquad \rightsquigarrow \qquad \arg\min_{x \in \mathcal{M}} f(x)$$

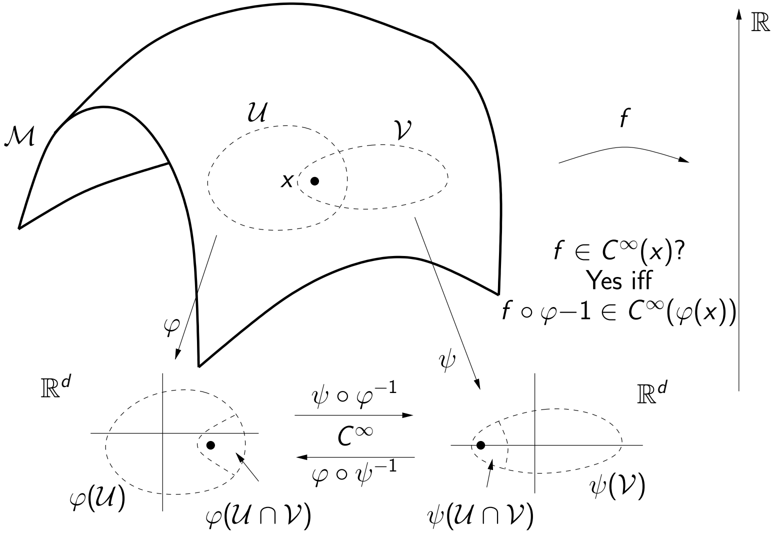
Smooth optimization on a manifold: what "smooth" means



Smooth optimization on a manifold: what "smooth" means



Smooth optimization on a manifold: what "smooth" means



Smooth optimization on a manifold: what "smooth" means

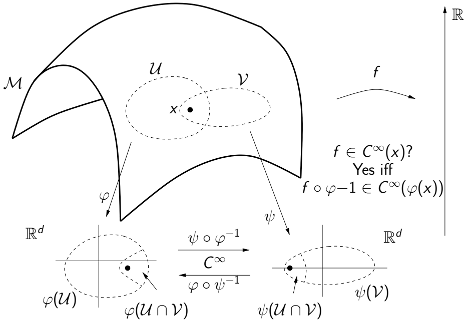
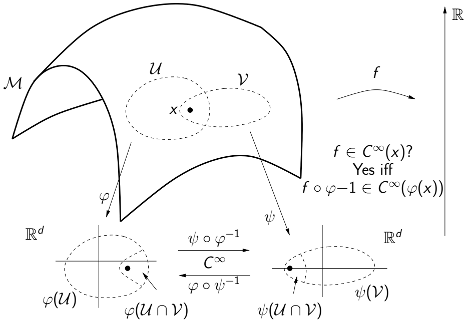


Chart: $\mathcal{U} \xrightarrow[\text{bij.}]{\varphi} \varphi(\mathcal{U})$
Atlas: Collection of "compatible chars" that cover $\mathcal{M}$
Manifold: Set with an atlas

Optimization on manifolds in its most abstract formulation

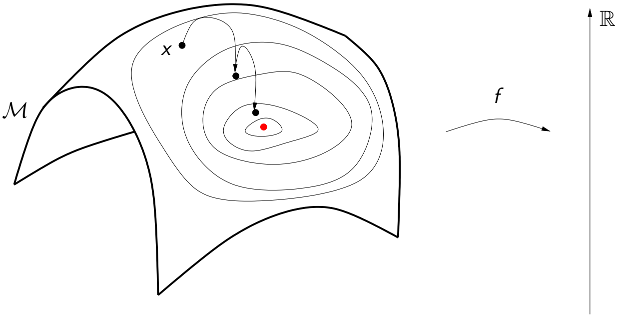


Given:

- A set $\mathcal{M}$ endowed (explicitly or implicitly) with a manifold structure (i.e., a collection of compatible charts).
- A function $f : \mathcal{M} \to \mathbb{R}$, smooth in the sense of the manifold structure.

Task: Compute a local minimizer of f.

Optimization on manifolds: algorithms

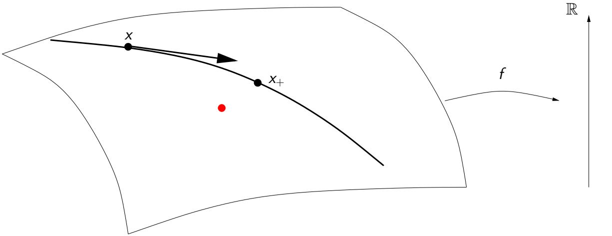


Given:

- A set $\mathcal{M}$ endowed (explicitly or implicitly) with a manifold structure (i.e., a collection of compatible charts).
- A function $f : \mathcal{M} \to \mathbb{R}$, smooth in the sense of the manifold structure.

Task: Compute a local minimizer of f.

Previous work on Optimization On Manifolds



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)".

The purely Riemannian era

Gabay (1982), *Minimizing a differentiable function over a differential manifold*. Stepest descent along geodesics; Newton's method along geodesics; Quasi-Newton methods along geodesics.

Smith (1994), *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.

The pragmatic era

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.

Looking ahead: Newton on abstract manifolds

Required: Riemannian manifold $\mathcal{M}$; retraction R on $\mathcal{M}$; affine connection ∇ on $\mathcal{M}$; real-valued function f on $\mathcal{M}$.
Iteration $x_k \in \mathcal{M} \mapsto x_{k+1} \in \mathcal{M}$ defined by

1. Solve the Newton equation

$$\mathrm{Hess}\, f(x_k)\eta_k = -\mathrm{grad}\, f(x_k)$$

   for the unknown $\eta_k \in T_{x_k}\mathcal{M}$, where

$$\mathrm{Hess}\, f(x_k)\eta_k := \nabla_{\eta_k}\mathrm{grad}\, f.$$

2. Set

$$x_{k+1} := R_{x_k}(\eta_k).$$

Looking ahead: Newton on submanifolds of $\mathbb{R}^n$

Required: Riemannian submanifold $\mathcal{M}$ of $\mathbb{R}^n$; retraction R on $\mathcal{M}$; real-valued function f on $\mathcal{M}$.
Iteration $x_k \in \mathcal{M} \mapsto x_{k+1} \in \mathcal{M}$ defined by

1. Solve the Newton equation

$$\mathrm{Hess}\, f(x_k)\eta_k = -\mathrm{grad}\, f(x_k)$$

   for the unknown $\eta_k \in T_{x_k}\mathcal{M}$, where

$$\mathrm{Hess}\, f(x_k)\eta_k := \mathrm{P}_{T_{x_k}\mathcal{M}} \mathrm{D}(\mathrm{grad}\, f)(x_k)[\eta_k].$$

2. Set

$$x_{k+1} := R_{x_k}(\eta_k).$$

Looking ahead: Newton on the unit sphere S^{n-1}

Required: real-valued function f on S^{n-1}.
Iteration $x_k \in S^{n-1} \mapsto x_{k+1} \in S^{n-1}$ defined by

1. Solve the Newton equation

$$\begin{cases} \mathrm{P}_{x_k} \mathrm{D}(\mathrm{grad}\, f)(x_k)[\eta_k] = -\mathrm{grad}\, f(x_k) \\ x^T \eta_k = 0, \end{cases}$$

   for the unknown $\eta_k \in \mathbb{R}^n$, where

$$\mathrm{P}_{x_k} = (I - x_k x_k^T).$$

2. Set

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

Looking ahead: Newton for Rayleigh quotient optimization on unit sphere

Iteration $x_k \in S^{n-1} \mapsto x_{k+1} \in S^{n-1}$ defined by

1. Solve the Newton equation

$$\begin{cases} \mathrm{P}_{x_k} A \mathrm{P}_{x_k} \eta_k - \eta_k x_k^T A x_k = -\mathrm{P}_{x_k} A x_k, \\ x_k^T \eta_k = 0, \end{cases}$$

   for the unknown $\eta_k \in \mathbb{R}^n$, where

$$\mathrm{P}_{x_k} = (I - x_k x_k^T).$$

2. Set

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

Programme

- Provide background in differential geometry instrumental for algorithmic development
- Present manifold versions of some classical optimization algorithms: steepest-descent, Newton, conjugate gradients, trust-region methods
- Show how to turn these abstract geometric algorithms into practical implementations
- Illustrate several problems that can be rephrased as optimization problems on manifolds.

Some important manifolds

- Stiefel manifold $\mathrm{St}(p, n)$: set of all orthonormal $n \times p$ matrices.
- Grassmann manifold $\mathrm{Grass}(p, n)$: set of all p-dimensional subspaces of $\mathbb{R}^n$
- Euclidean group $SE(3)$: set of all rotations-translations
- Flag manifold, shape manifold, oblique manifold...
- Several unnamed manifolds

A manifold-based approach to the symmetric eigenvalue problem

OPT

EVP

OPT

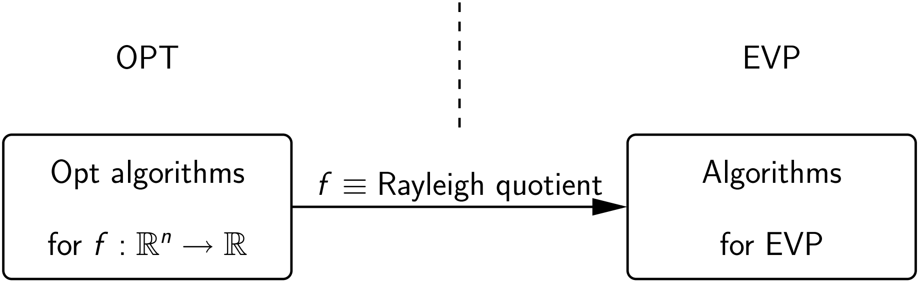
EVP

Opt algorithms

for $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

Algorithms

for EVP


OPT
EVP
Opt algorithms
for $f : \mathbb{R}^n \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP

Rayleigh quotient

Rayleigh quotient of (A, B):

$$f : \mathbb{R}^n_* \to \mathbb{R} : f(y) = \frac{y^T A y}{y^T B y}$$

Let A, B in $\mathbb{R}^{n \times n}$, $A = A^T$, $B = B^T \succ 0$,

$$Av_i = \lambda_i B v_i$$

with $\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n$.
Stationary points of f: αv_i, for all $\alpha \neq 0$.
Local (and global) minimizers of f: αv_1, for all $\alpha \neq 0$.

OPT

EVP

Opt algorithms

for $f : \mathbb{R}^n \to \mathbb{R}$

$\xrightarrow{f \equiv \text{Rayleigh quotient}}$

Algorithms

for EVP

"Block" Rayleigh quotient

Let $\mathbb{R}_*^{n\times p}$ denote the set of all full-column-rank $n \times p$ matrices.
Generalized ("block") Rayleigh quotient:

$$f : \mathbb{R}_*^{n\times p} \to \mathbb{R} : f(Y) = \operatorname{trace}\left((Y^T BY)^{-1} Y^T AY\right)$$

Stationary points of f:

$$\begin{bmatrix} v_{i_1} & \dots & v_{i_p} \end{bmatrix} M, \quad \text{for all } M \in \mathbb{R}_*^{p\times p}.$$

Minimizers of f:

$$\begin{bmatrix} v_1 & \dots v_p \end{bmatrix} M, \quad \text{for all } M \in \mathbb{R}_*^{p\times p}.$$

OPT

EVP

Opt algorithms

for $f : \mathbb{R}^n \to \mathbb{R}$

$\xrightarrow{f \,\equiv\, \text{Rayleigh quotient}}$

Algorithms

for EVP

OPT

EVP

Opt algorithms

for $f : \mathbb{R}^n \to \mathbb{R}$

$f \equiv$ Rayleigh quotient

Algorithms

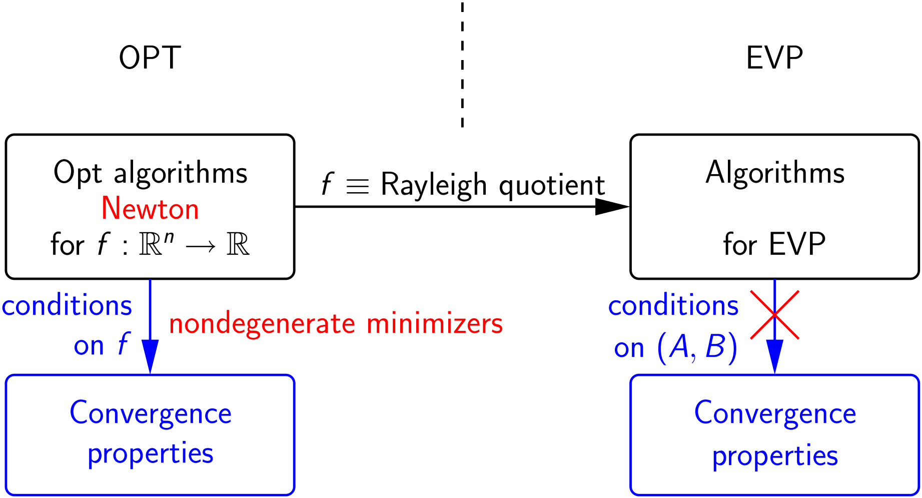
for EVP

conditions on f

Convergence properties

conditions on (A, B)

Convergence properties


OPT
EVP
Opt algorithms
Newton
for $f : \mathbb{R}^n \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
conditions
on (A, B)
Convergence
properties
Convergence
properties

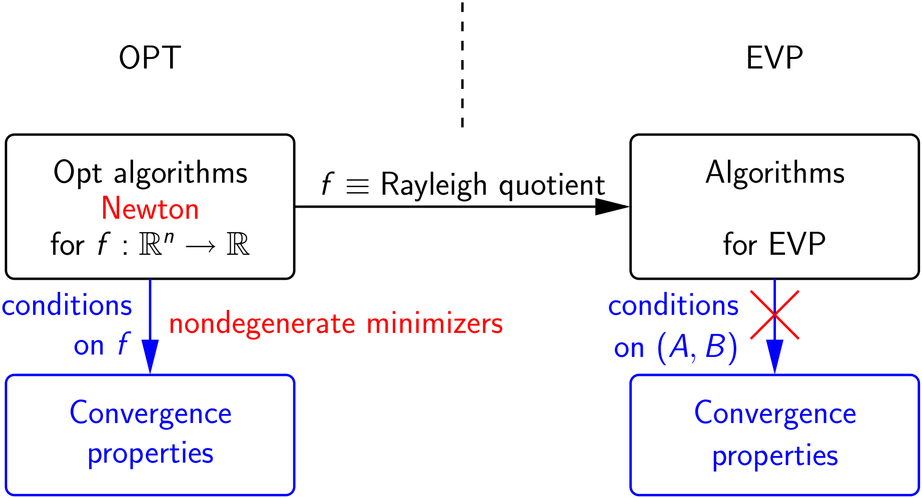
Newton for Rayleigh quotient in $\mathbb{R}^n_0$

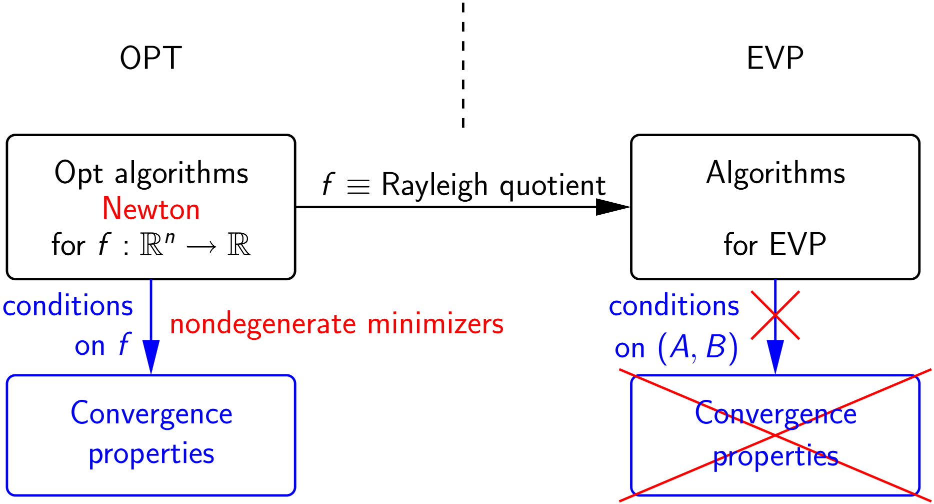
Let f denote the Rayleigh quotient of (A, B).
Let $x \in \mathbb{R}^n_0$ be any point such that $f(x) \notin \operatorname{spec}(B^{-1}A)$.
Then the Newton iteration

$$x \mapsto x - \left(\mathrm{D}^2 f(x)\right)^{-1} \cdot \operatorname{grad} f(x)$$

reduces to the iteration

$$x \mapsto 2x.$$


OPT
EVP
Opt algorithms
Newton
for $f : \mathbb{R}^n \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
conditions
on (A, B)
Convergence
properties
Convergence
properties


OPT
EVP
Opt algorithms
Newton
for $f : \mathbb{R}^n \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
conditions
on (A, B)
Convergence
properties
Convergence
properties

Invariance properties of the Rayleigh quotient

Rayleigh quotient of (A, B):

$$f : \mathbb{R}^n_* \to \mathbb{R} : f(y) = \frac{y^T A y}{y^T B y}$$

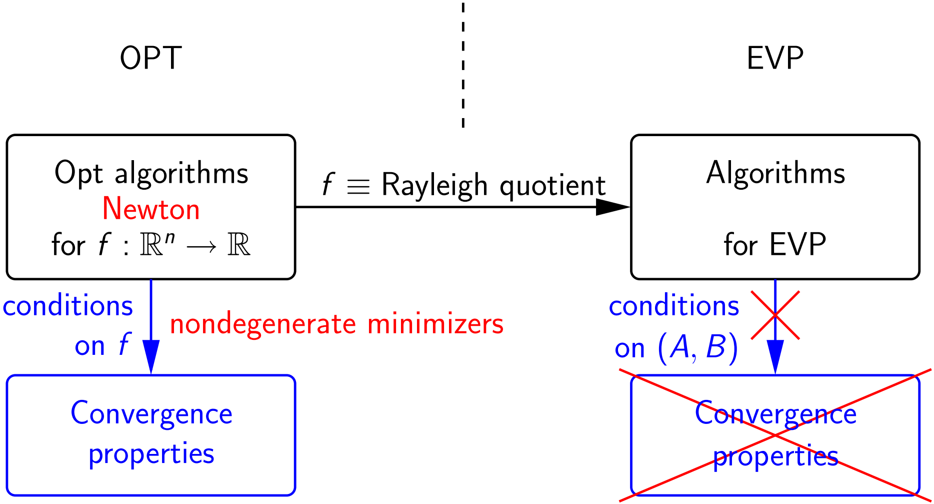
Invariance: $f(\alpha y) = f(y)$ for all $\alpha \in \mathbb{R}_0$.

Invariance properties of the Rayleigh quotient

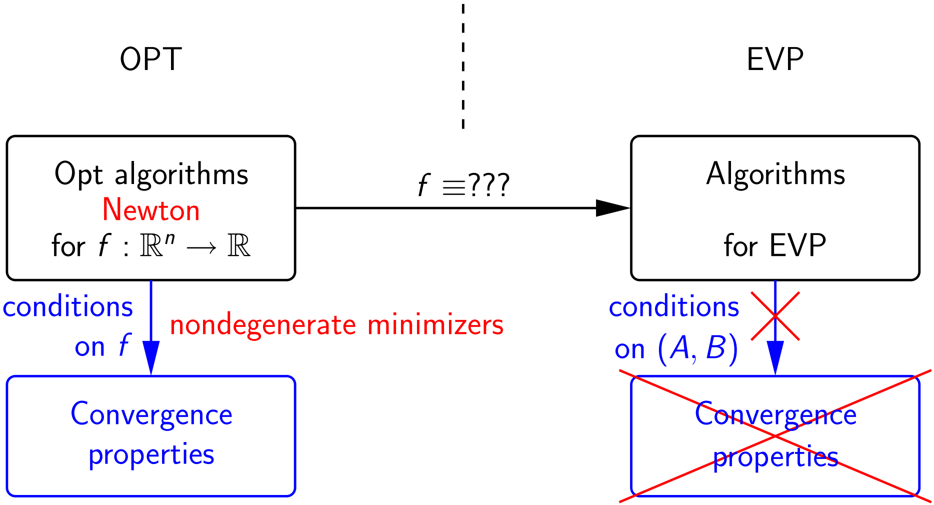
Generalized ("block") Rayleigh quotient:

$$f : \mathbb{R}_*^{n\times p} \to \mathbb{R} : f(Y) = \mathrm{trace}\left((Y^T BY)^{-1} Y^T AY\right)$$

Invariance: $f(YM) = f(Y)$ for all $M \in \mathbb{R}_*^{p\times p}$.


OPT
EVP
Opt algorithms
Newton
for $f : \mathbb{R}^n \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
Convergence
properties
conditions
on (A, B)
Convergence
properties

Remedy 1: modify f



Remedy 1: modify f

Consider

$$P_A : \mathbb{R}^n \to \mathbb{R} : x \mapsto P_A(x) := (x^T x)^2 - 2x^T Ax.$$

Theorem

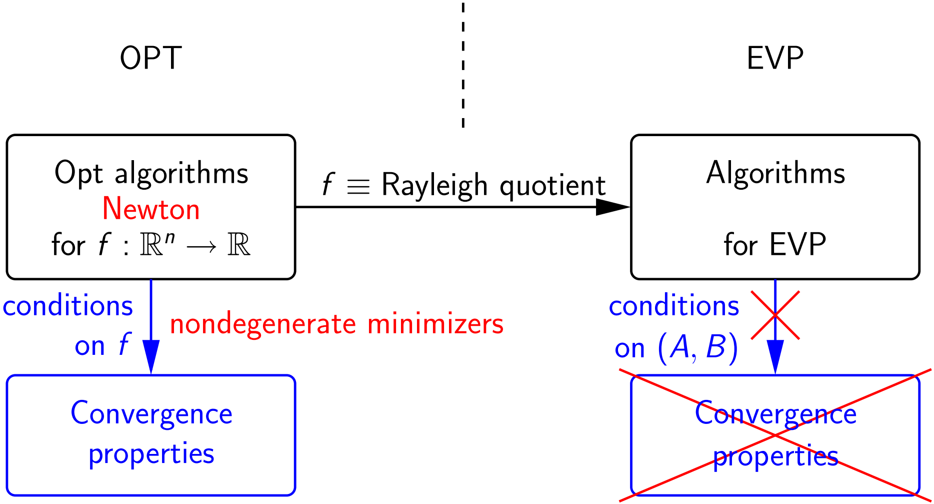
(i)

$$\min_{x \in \mathbb{R}^n} P_A(x) = -\lambda_n^2$$

The minimum is attained at any $\sqrt{\lambda_n} v_n$, where v_n is a unitary eigenvector related to λ_n.

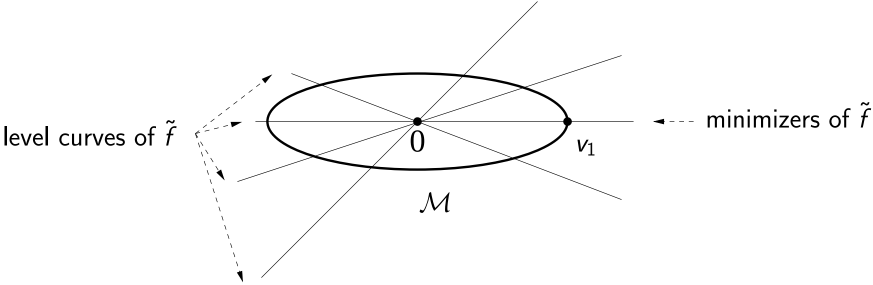
(ii) The set of critical points of P_A is $\{0\} \cup \{\sqrt{\lambda_k} v_k\}$.

References: Auchmuty (1989), Mongeau and Torki (2004).


OPT
EVP
Opt algorithms
Newton
for $f : \mathbb{R}^n \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
conditions
on (A, B)
Convergence
properties
~~Convergence properties~~

EVP: optimization on ellipsoid

$f(\alpha y) = f(y)$



Remedy 2: modify the search space

Instead of

$$f : \mathbb{R}^n_* \to \mathbb{R} : f(y) = \frac{y^T A y}{y^T B y},$$

minimize

$$f : \mathcal{M} \to \mathbb{R} : f(y) = \frac{y^T A y}{y^T B y},$$

where

$$\mathcal{M} = \{y \in \mathbb{R}^n : y^T B y = 1\}.$$

Stationary points of f: $\pm v_i$.
Local (and global) minimizers of f: $\pm v_1$.

Remedy 2: modify search space: block case

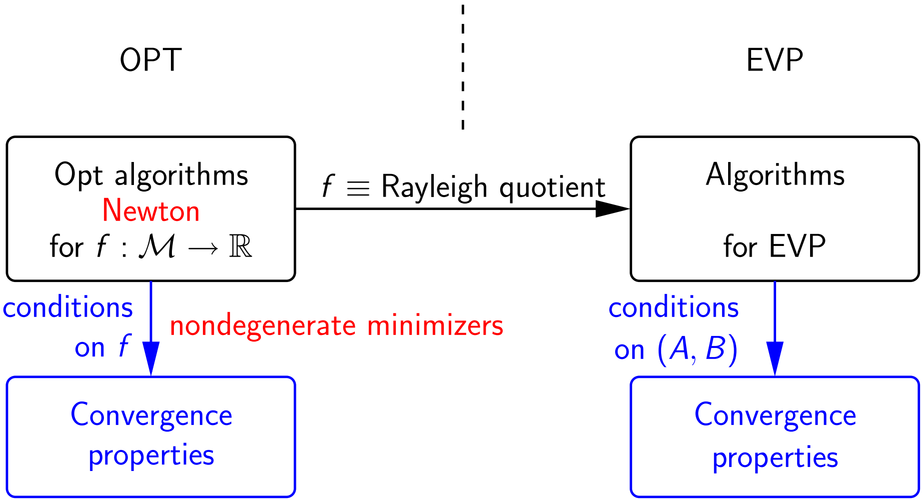
Instead of generalized ("block") Rayleigh quotient:

$$f : \mathbb{R}_*^{n\times p} \to \mathbb{R} : f(Y) = \operatorname{trace}\left((Y^T BY)^{-1} Y^T AY\right),$$

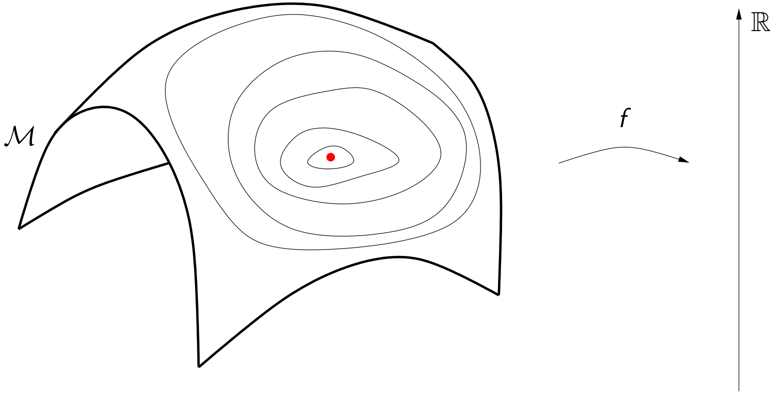
minimize

$$f : \mathrm{Grass}(p,n) \to \mathbb{R} : f(\mathrm{col}(Y)) = \operatorname{trace}\left((Y^T BY)^{-1} Y^T AY\right),$$

where $\mathrm{Grass}(p,n)$ denotes the set of all p-dimensional subspaces of $\mathbb{R}^n$, called the *Grassmann manifold*.
Stationary points of f: $\mathrm{col}(\begin{bmatrix} v_{i_1} & \dots & v_{i_p} \end{bmatrix})$.
Minimizer of f: $\mathrm{col}(\begin{bmatrix} v_1 & \dots v_p \end{bmatrix})$.

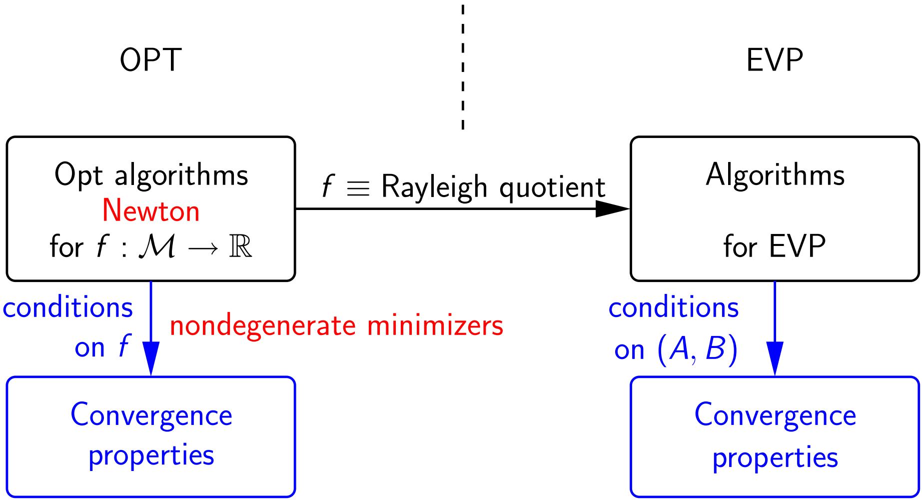

OPT
EVP
Opt algorithms
Newton
for $f : \mathcal{M} \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
conditions
on (A, B)
Convergence
properties
Convergence
properties

Smooth optimization on a manifold: big picture



Smooth optimization on a manifold: tools

	Purely Riemannian way	Pragmatic way
Search direction	Tangent vector	Tangent vector
Steepest descent dir.	$-\mathrm{grad}\, f(x)$	$-\mathrm{grad}\, f(x)$
Derivative of vector field	Levi-Civita connection $\overset{g}{\nabla}$	Any connection ∇
Update	Search along the geodesic tangent to the search direction	Search along any curve ta to the search direction scribed by a *retraction*)
Displacement of tgt vectors	Parallel translation induced by $\overset{g}{\nabla}$	Vector Transport


OPT
EVP
Opt algorithms
Newton
for $f : \mathcal{M} \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
conditions
on (A, B)
Convergence
properties
Convergence
properties

Newton's method on abstract manifolds

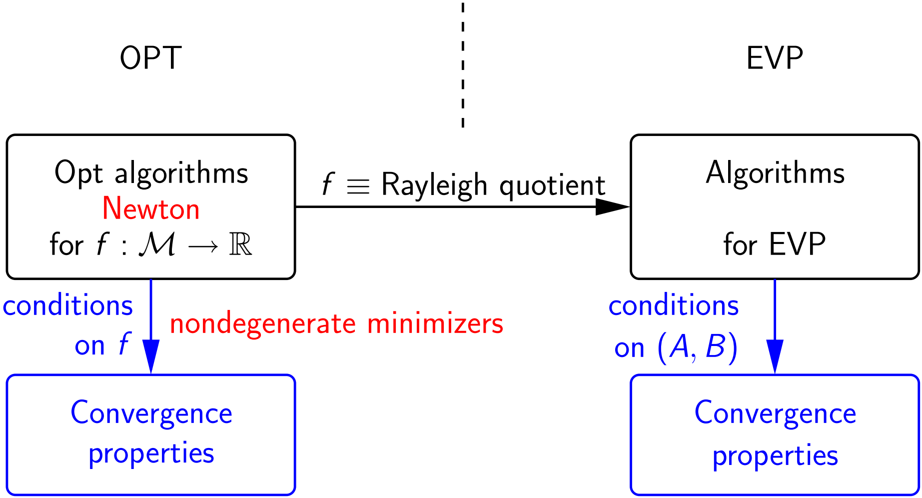
Required: Riemannian manifold $\mathcal{M}$; retraction R on $\mathcal{M}$; affine connection ∇ on $\mathcal{M}$; real-valued function f on $\mathcal{M}$.
Iteration $x_k \in \mathcal{M} \mapsto x_{k+1} \in \mathcal{M}$ defined by

1. Solve the Newton equation

$$\mathrm{Hess}\, f(x_k)\eta_k = -\mathrm{grad}\, f(x_k)$$

   for the unknown $\eta_k \in T_{x_k}\mathcal{M}$, where $\mathrm{Hess}\, f(x_k)\eta_k := \nabla_{\eta_k}\mathrm{grad}\, f$.
2. Set

$$x_{k+1} := R_{x_k}(\eta_k).$$

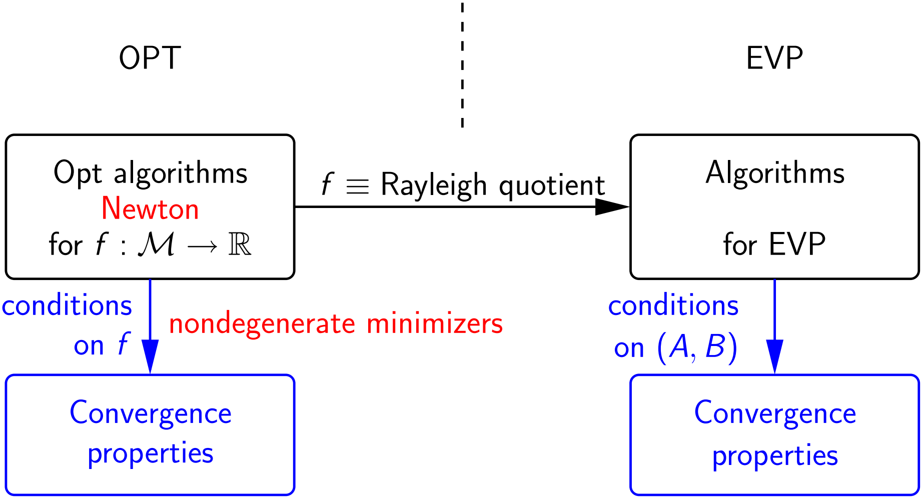

OPT
EVP
Opt algorithms
Newton
for $f : \mathcal{M} \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
conditions
on (A, B)
Convergence
properties
Convergence
properties

Convergence of Newton's method on abstract manifolds

Theorem

Let $x_* \in \mathcal{M}$ be a nondegenerate critical point of f, i.e., $\mathrm{grad}\, f(x_*) = 0$ and $\mathrm{Hess}\, f(x_*)$ invertible.

Then there exists a neighborhood $\mathcal{U}$ of x_* in $\mathcal{M}$ such that, for all $x_0 \in \mathcal{U}$, Newton's method generates an infinite sequence $(x_k)_{k=0,1,\dots}$ converging superlinearly (at least quadratically) to x_*.


OPT
EVP
Opt algorithms
Newton
for $f : \mathcal{M} \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
conditions
on (A, B)
Convergence
properties
Convergence
properties

Geometric Newton for Rayleigh quotient optimization

Iteration $x_k \in S^{n-1} \mapsto x_{k+1} \in S^{n-1}$ defined by

1. Solve the Newton equation

$$\begin{cases} \mathrm{P}_{x_k} A \mathrm{P}_{x_k} \eta_k - \eta_k x_k^T A x_k = -\mathrm{P}_{x_k} A x_k, \\ x_k^T \eta_k = 0, \end{cases}$$

   for the unknown $\eta_k \in \mathbb{R}^n$, where

$$P_{x_k} = (I - x_k x_k^T).$$

2. Set

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

Geometric Newton for Rayleigh quotient optimization: block case

Iteration $\mathrm{col}(Y_k) \in \mathrm{Grass}(p,n) \mapsto \mathrm{col}(Y_{k+1}) \in \mathrm{Grass}(p,n)$ defined by

1. Solve the linear system

$$\begin{cases} \mathrm{P}^h_{Y_k}\left(AZ_k - Z_k(Y_k^T Y_k)^{-1} Y_k^T A Y_k\right) = -\mathrm{P}^h_{Y_k}(AY_k) \\ Y_k^T Z_k = 0 \end{cases}$$

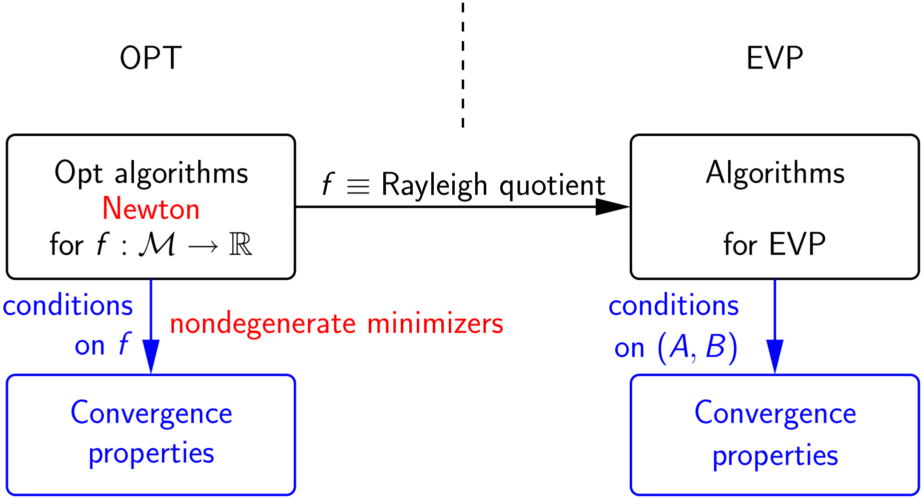
   for the unknown $Z_k \in \mathbb{R}^{n\times p}$, where

$$\mathrm{P}^h_{Y_k} = (I - Y_k(Y_k^T Y_k)^{-1} Y_k^T).$$

2. Set

$$Y_{k+1} = (Y_k + Z_k)N_k$$

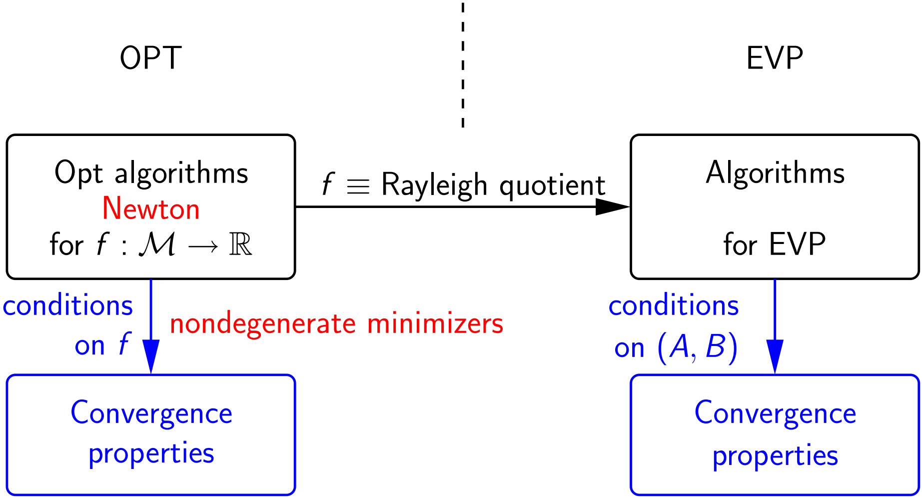
   where N_k is a nonsingular $p \times p$ matrix chosen for normalization.


OPT
EVP
Opt algorithms
Newton
for $f : \mathcal{M} \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
conditions
on (A, B)
Convergence
properties
Convergence
properties

Convergence of the EVP algorithm

Theorem

Let $Y_* \in \mathbb{R}^{n\times p}$ be such that $\mathrm{col}(Y_*)$ is a spectral invariant subspace of $B^{-1}A$. Then there exists a neighborhood $\mathcal{U}$ of $\mathrm{col}(Y_*)$ in $\mathrm{Grass}(p,n)$ such that, for all $Y_0 \in \mathbb{R}^{n\times p}$ with $\mathrm{col}(Y_0) \in \mathcal{U}$, Newton's method generates an infinite sequence $(Y_k)_{k=0,1,\ldots}$ such that $(\mathrm{col}(Y_k))_{k=0,1,\ldots}$ converges superlinearly (at least quadratically) to $\mathrm{col}(Y_*)$ on $\mathrm{Grass}(p,n)$.

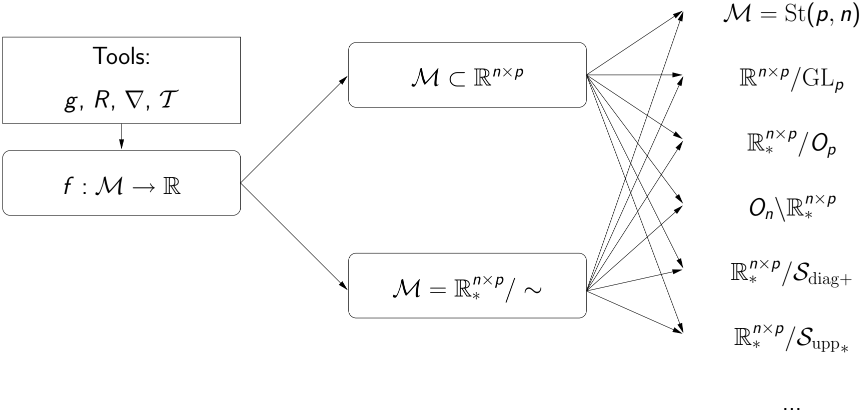

OPT
EVP
Opt algorithms
Newton
for $f : \mathcal{M} \to \mathbb{R}$
$f \equiv$ Rayleigh quotient
Algorithms
for EVP
conditions
on f
nondegenerate minimizers
conditions
on (A, B)
Convergence
properties
Convergence
properties

Other optimization methods

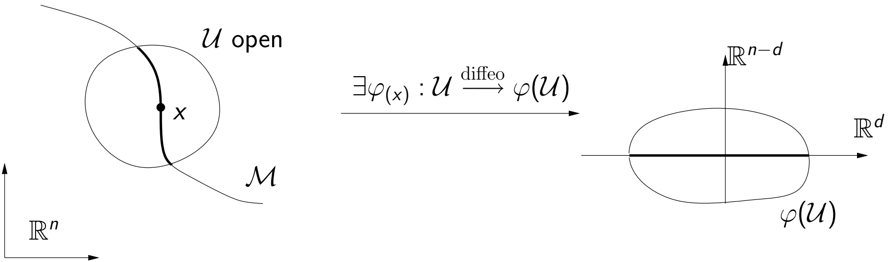
- Trust-region methods: PAA, C. G. Baker, K. A. Gallivan, *Trust-region methods on Riemannian manifolds*, Foundations of Computational Mathematics, 2007.
- "Implicit" trust-region methods: PAA, C. G. Baker, K. A. Gallivan, submitted.

Manifolds

Manifolds, submanifolds, quotient manifolds

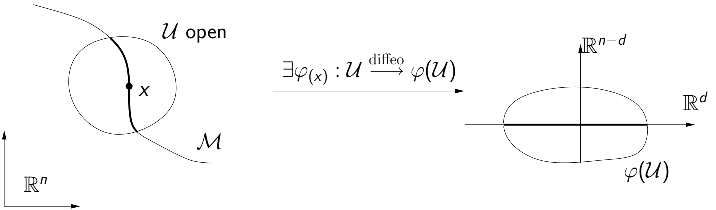


Submanifolds of $\mathbb{R}^n$



The *set* $\mathcal{M} \subset \mathbb{R}^n$ is termed a *submanifold* of $\mathbb{R}^n$ if the situation described above holds for all $x \in \mathcal{M}$.

Submanifolds of $\mathbb{R}^n$



The manifold structure on $\mathcal{M}$ is defined in a unique way as the manifold structure generated by the atlas $\left\{ \begin{bmatrix} e_1^T \\ \vdots \\ e_d^T \end{bmatrix} \varphi_{(x)}\big|_{\mathcal{M}} : x \in \mathcal{M} \right\}$.

Back to the basics: partial derivatives in $\mathbb{R}^n$

Let $F : \mathbb{R}^n \to \mathbb{R}^q$.
Define $\partial_i F : \mathbb{R}^n \to \mathbb{R}^q$ by

$$\partial_i F(x) = \lim_{t \to 0} \frac{F(x + te_i) - F(x)}{t}.$$

If $\partial_i F$ is defined and continuous on $\mathbb{R}^n$, then F is termed *continuously differentiable*, denoted by $F \in C^1$.

Back to the basics: (Fréchet) derivative in $\mathbb{R}^n$

If $F \in C^1$, then

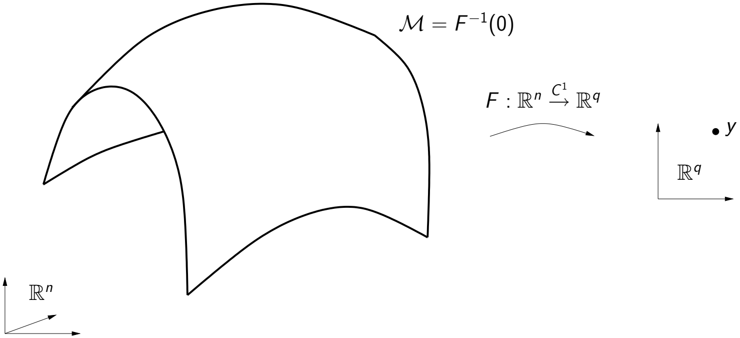
$$\mathrm{D}F(x) : \mathbb{R}^n \xrightarrow{\text{lin}} \mathbb{R}^q : z \mapsto \mathrm{D}F(x)[z] := \lim_{t \to 0} \frac{F(x + tz) - F(x)}{t}$$

is the *derivative* (or *differential*) of F at x.
We have $\mathrm{D}F(x)[z] = \mathrm{J}_F(x)z$, where the matrix

$$\mathrm{J}_F(x) = \begin{bmatrix} \partial_1(e_1^T F)(x) & \cdots & \partial_n(e_1^T F)(x) \\ \vdots & \ddots & \vdots \\ \partial_1(e_q^T F)(x) & \cdots & \partial_n(e_q^T F)(x) \end{bmatrix}$$

is the *Jacobian matrix* of F at x.

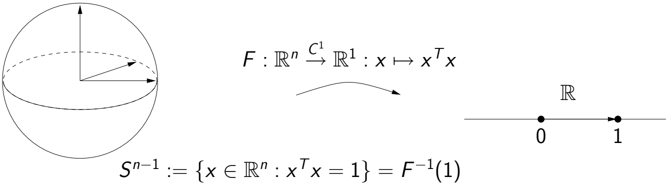
Submanifolds of $\mathbb{R}^n$: sufficient condition



$y \in \mathbb{R}^q$ is a *regular value* of F if, for all $x \in F^{-1}(y)$, $\mathrm{D}F(x)$ is an onto function (*surjection*).

Theorem (submersion theorem): If $y \in \mathbb{R}^q$ is a regular value of F, then $F^{-1}(y)$ is a submanifold of $\mathbb{R}^n$.

Submanifolds of $\mathbb{R}^n$: sufficient condition: application



The unit sphere

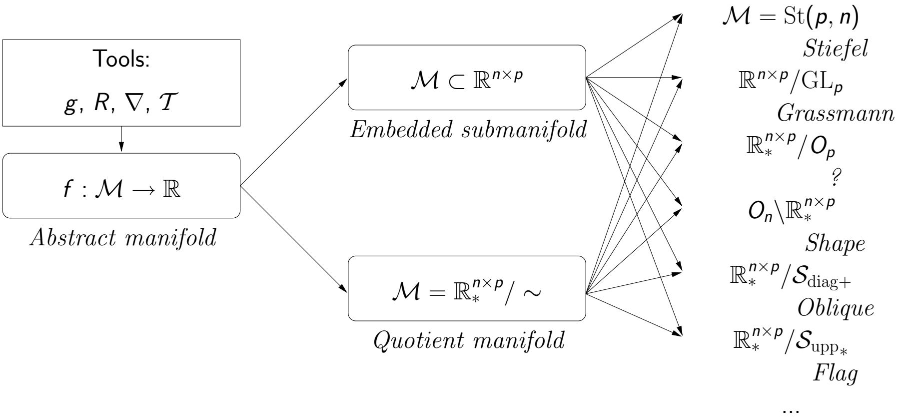
$$S^{n-1} := \{x \in \mathbb{R}^n : x^T x = 1\}$$

is a submanifold of $\mathbb{R}^n$.
Indeed, for all $x \in S^{n-1}$, we have that

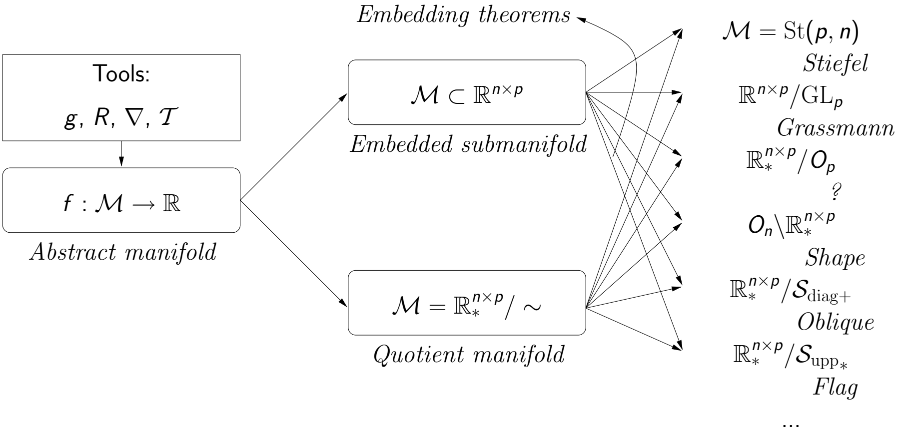
$$\mathrm{D}F(x) : \mathbb{R}^n \to \mathbb{R} : z \mapsto \mathrm{D}F(x)[z] = x^T z + z^T x$$

is an onto function.

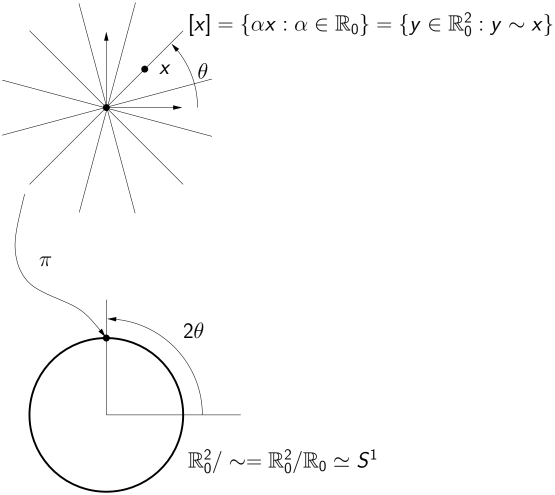
Manifolds, submanifolds, quotient manifolds



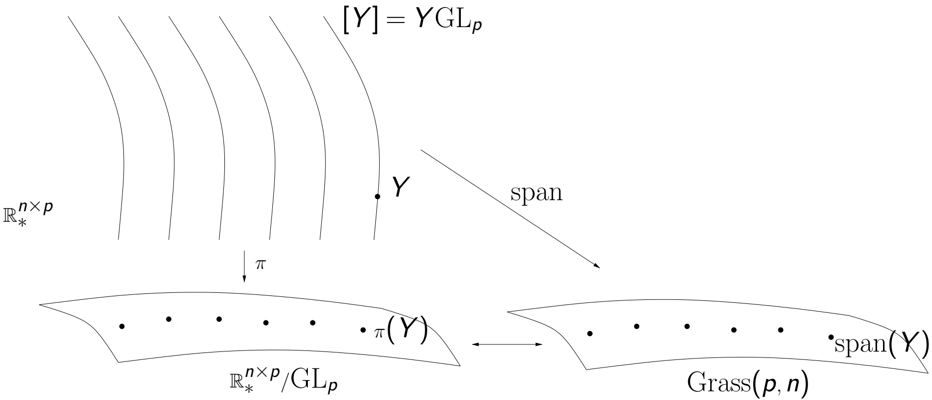
Manifolds, submanifolds, quotient manifolds



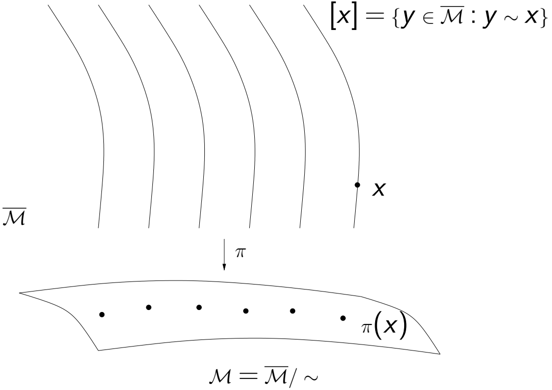
A simple quotient set: the projective space



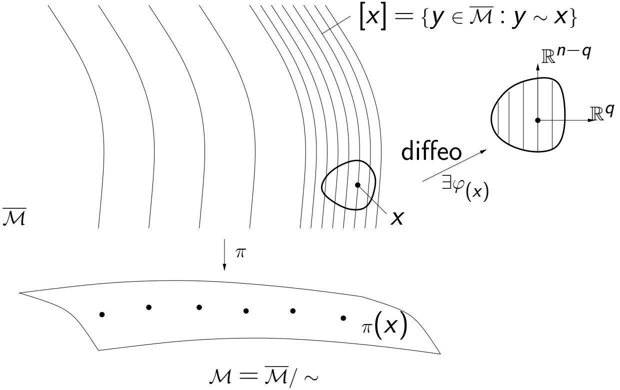
A slightly less simple quotient set: $\mathbb{R}_*^{n\times p}/\mathrm{GL}_p$



Abstract quotient set $\overline{\mathcal{M}}/\sim$

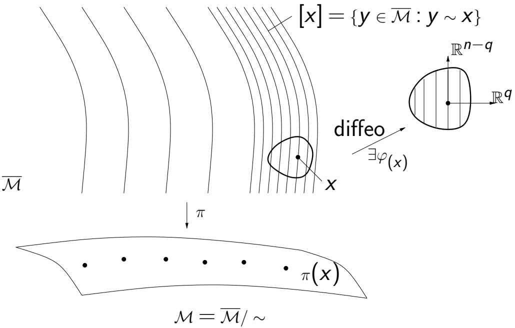


Abstract quotient manifold $\overline{\mathcal{M}}/\sim$



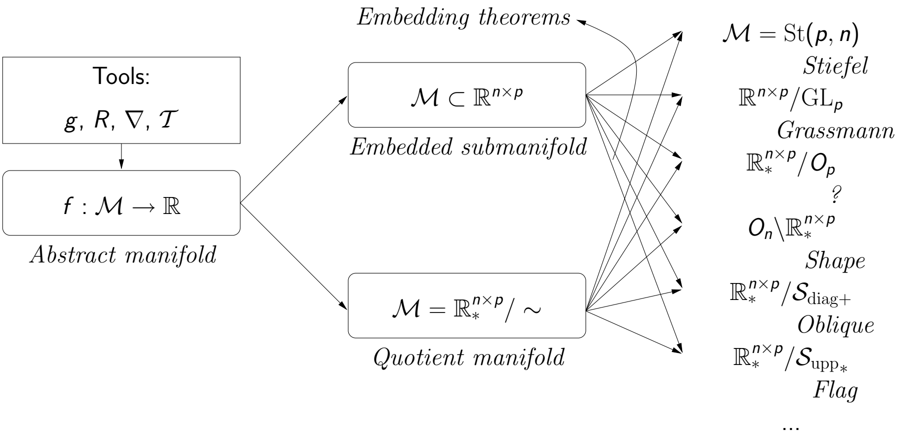
The set $\overline{\mathcal{M}}/\sim$ is termed a *quotient manifold* if the situation described above holds for all $x \in \overline{\mathcal{M}}$.

Abstract quotient manifold $\overline{\mathcal{M}}/\sim$



The manifold structure on $\overline{\mathcal{M}}/\sim$ is defined in a unique way as the manifold structure generated by the atlas $\left\{ \begin{bmatrix} e_1^T \\ \vdots \\ e_q^T \end{bmatrix} \varphi_{(x)} \circ \pi^{-1} : x \in \overline{\mathcal{M}} \right\}$.

Manifolds, submanifolds, quotient manifolds



Manifolds, and where they appear

- Stiefel manifold $\mathrm{St}(p, n)$ and orthogonal group $O_p = \mathrm{St}(n, n)$

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

  Applications: computer vision; principal component analysis; independent component analysis...
- Grassmann manifold $\mathrm{Grass}(p, n)$

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

  Applications: various dimension reduction problems...
- $\mathbb{R}_*^{n \times p} / O_p$

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

  Applications: Low-rank approximation of symmetric matrices; low-rank approximation of tensors...

Manifolds, and where they appear

- Shape manifold $O_n / \mathbb{R}_*^{n \times p}$

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

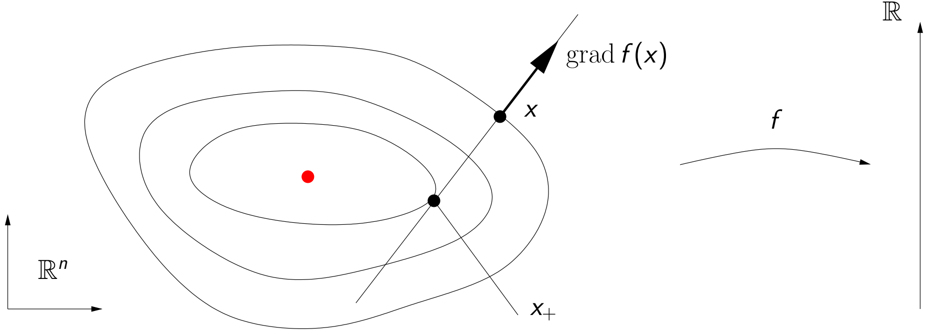
  Applications: shape analysis
- Oblique manifold $\mathbb{R}_*^{n \times p} / \mathcal{S}_{\mathrm{diag}+}$

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

  Applications: independent component analysis; factor analysis (oblique Procrustes problem)...
- 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...

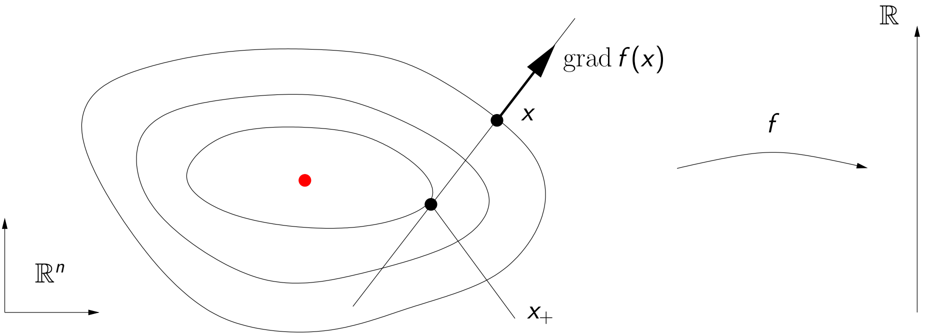
Steepest-descent methods on manifolds

Steepest-descent in $\mathbb{R}^n$



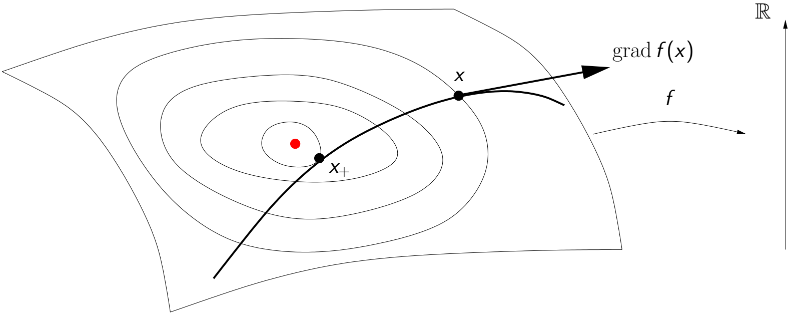
$$\operatorname{grad} f(x) = \begin{bmatrix} \partial_1 f(x) & \cdots & \partial_n f(x) \end{bmatrix}^T$$

Steepest-descent: from $\mathbb{R}^n$ to manifolds



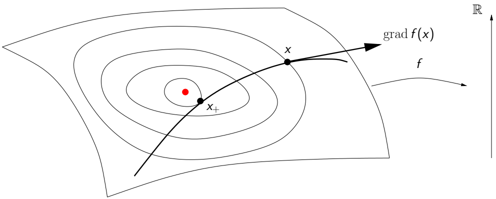
	$\mathbb{R}^n$	Manifold
Search direction	Vector at x	Tangent vector at x
Steepest-desc. dir.	$-\operatorname{grad} f(x)$	$-\operatorname{grad} f(x)$
Curve	$\gamma : t \mapsto x - t \operatorname{grad} f(x)$	γ s.t. $\gamma(0) = x$ and $\dot{\gamma}(0) = -\operatorname{grad} f(x)$

Steepest-descent: from $\mathbb{R}^n$ to manifolds



	$\mathbb{R}^n$	Manifold
Search direction	Vector at x	Tangent vector at x
Steepest-desc. dir.	$-\operatorname{grad} f(x)$	$-\operatorname{grad} f(x)$
Curve	$\gamma : t \mapsto x - t \operatorname{grad} f(x)$	γ s.t. $\gamma(0) = x$ and $\dot{\gamma}(0) = -\operatorname{grad} f(x)$

Update directions: tangent vectors



Let γ be a curve in the manifold $\mathcal{M}$ with $\gamma(0) = x$.

For an abstract manifold, the definition $\dot{\gamma}(0) = \frac{\mathrm{d}\gamma}{\mathrm{d}t}(0) = \lim_{t\to 0} \frac{\gamma(t)-\gamma(0)}{t}$ is meaningless.

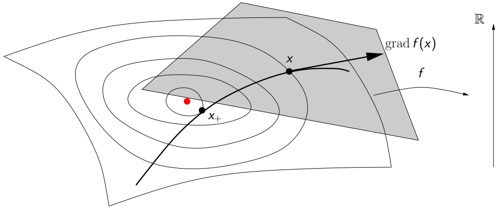
Instead, define: $\mathrm{D}f(x)[\dot{\gamma}(0)] := \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))\big|_{t=0}$

If $\mathcal{M} \subset \mathbb{R}^n$ and $f = \overline{f}|_{\mathcal{M}}$, then

$$\mathrm{D}f(x)[\dot{\gamma}(0)] = \mathrm{D}\overline{f}(x)\left[\frac{\mathrm{d}\gamma}{\mathrm{d}t}(0)\right].$$

The application $\dot{\gamma}(0) : f \mapsto \mathrm{D}f(x)[\dot{\gamma}(0)]$ is a *tangent vector* at x.

Update directions: tangent spaces



The set

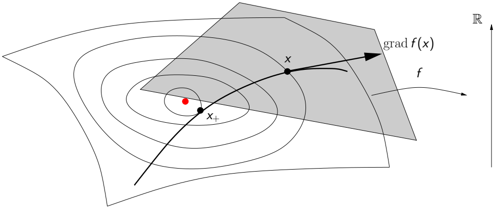
$$T_x\mathcal{M} = \{\dot\gamma(0) : \gamma \text{ curve in } \mathcal{M} \text{ through } x \text{ at } t = 0\}$$

is the *tangent space* to $\mathcal{M}$ at x.
With the definition

$$\alpha\dot\gamma_1(0) + \beta\dot\gamma_2(0) : f \mapsto \alpha \mathrm{D}f(x)[\dot\gamma_1(0)] + \beta \mathrm{D}f(x)[\dot\gamma_2(0)],$$

the tangent space $T_x\mathcal{M}$ becomes a linear space.
The *tangent bundle* $T\mathcal{M}$ is the set of all tangent vectors to $\mathcal{M}$.

Tangent vectors: submanifolds of Euclidean spaces

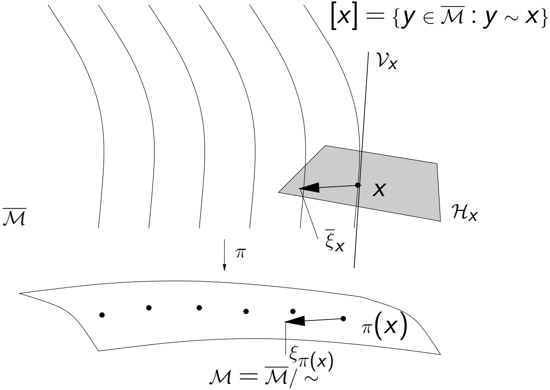


If $\mathcal{M}$ is a submanifold of $\mathbb{R}^n$ and $f = \overline{f}|_{\mathcal{M}}$, then

$$\mathrm{D}f(x)[\dot{\gamma}(0)] = \mathrm{D}\overline{f}(x)\left[\frac{\mathrm{d}\gamma}{\mathrm{d}t}(0)\right].$$

Proof: The left-hand side is equal to $\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))\big|_{t=0}$. This is equal to $\frac{\mathrm{d}}{\mathrm{d}t}\overline{f}(\gamma(t))\big|_{t=0}$ because $\gamma(t) \in \mathcal{M}$ for all t. The classical chain rule yields the right-hand side.

Tangent vectors: quotient manifolds



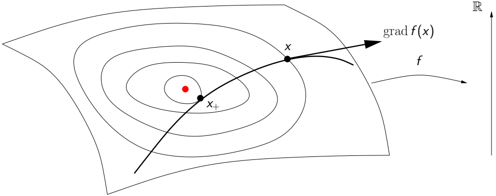
Let $\overline{\mathcal{M}}/\sim$ be a quotient manifold. Then $[x]$ is a submanifold of $\overline{\mathcal{M}}$. The tangent space $T_x[x]$ is the *vertical space* $\mathcal{V}_x$. A *horizontal space* is a subspace of $T_x\overline{\mathcal{M}}$ complementary to $\mathcal{V}_x$.

Let $\xi_{\pi(x)}$ be a tangent vector to $\overline{\mathcal{M}}/\sim$ at $\pi(x)$.

Theorem: In $\mathcal{H}_x$ there is one and only one $\bar{\xi}_x$ such that

$$\mathrm{D}\pi(x)[\bar{\xi}_x] = \xi_{\pi(x)}.$$

Steepest-descent: norm of tangent vectors



The steepest ascent direction is along

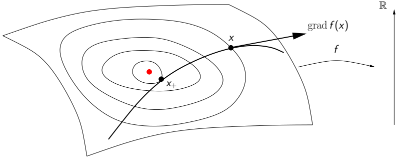
$$\underset{\substack{\xi \in T_x\mathcal{M} \\ \|\xi\|=1}}{\arg\max}\, \mathrm{D}f(x)[\xi].$$

To this end, we need a norm on $T_x\mathcal{M}$.
For all $x \in \mathcal{M}$, let g_x denote an inner product in $T_x\mathcal{M}$, and define

$$\|\xi_x\| := \sqrt{g_x(\xi_x, \xi_x)}.$$

When g_x "smoothly" depends on x, we say that $(\mathcal{M}, g)$ is a *Riemannian manifold*.

Steepest-descent: gradient



There is a unique $\operatorname{grad} f(x)$, called the *gradient* of f at x, such that

$$\begin{cases} \operatorname{grad} f(x) \in T_x\mathcal{M} \\ g_x(\operatorname{grad} f(x), \xi_x) = \mathrm{D}f(x)[\xi_x], \quad \forall \xi_x \in T_x\mathcal{M}. \end{cases}$$

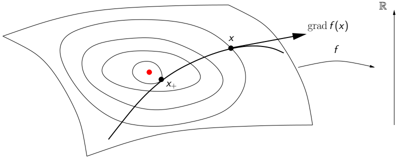
We have

$$\frac{\operatorname{grad} f(x)}{\|\operatorname{grad} f(x)\|} = \operatorname*{arg\,max}_{\substack{\xi \in T_x\mathcal{M} \\ \|\xi\|=1}} \mathrm{D}f(x)[\xi]$$

and

$$\|\operatorname{grad} f(x)\| = \mathrm{D}f(x)\left[\frac{\operatorname{grad} f(x)}{\|\operatorname{grad} f(x)\|}\right].$$

Steepest-descent: Riemannian submanifolds



Let $(\overline{\mathcal{M}}, \overline{g})$ be a Riemannian manifold and $\mathcal{M}$ be a submanifold of $\overline{\mathcal{M}}$. Then

$$g_x(\xi_x, \zeta_x) := \overline{g}_x(\xi_x, \eta_x), \ \forall \xi_x, \zeta_x \in T_x\mathcal{M}$$

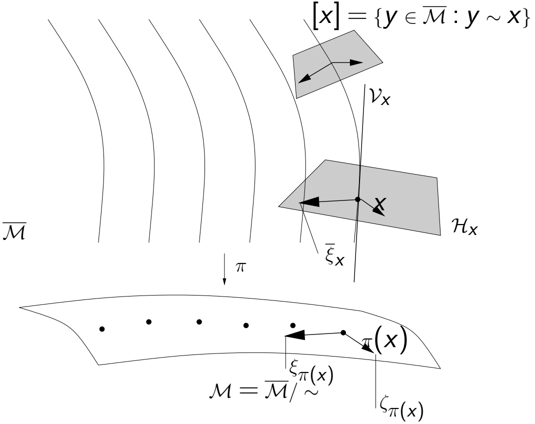
defines a Riemannian metric g on $\mathcal{M}$. With this Riemannian metric, $\mathcal{M}$ is a *Riemannian submanifold* of $\overline{\mathcal{M}}$.

Every $z \in T_x\overline{\mathcal{M}}$ admits a decomposition $z = \underbrace{\mathrm{P}_x z}_{\in T_x\mathcal{M}} + \underbrace{\mathrm{P}_x^{\perp} z}_{\in T_x^{\perp}\mathcal{M}}$.

If $\overline{f} : \overline{\mathcal{M}} \to \mathbb{R}$ and $f = \overline{f}|_{\mathcal{M}}$, then

$$\operatorname{grad} f(x) = \mathrm{P}_x \operatorname{grad} \overline{f}(x).$$

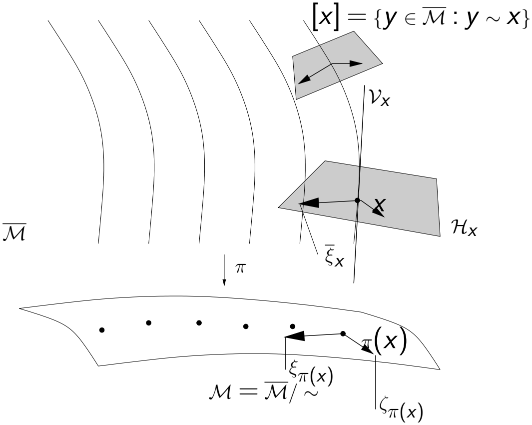
Steepest-descent: Riemannian quotient manifolds



Let $\tilde{g}$ be a Riemannian metric on $\overline{\mathcal{M}}$.
Suppose that, for all $\xi_{\pi(x)}$ and $\zeta_{\pi(x)}$ in $T_{\pi(x)}\overline{\mathcal{M}}/\sim$, and all $\tilde{x} \in \pi^{-1}(\pi(x))$, we have

$$\overline{g}_{\tilde{x}}(\overline{\xi}_{\tilde{x}}, \overline{\zeta}_{\tilde{x}}) = \overline{g}_x(\overline{\xi}_x, \overline{\zeta}_x).$$

Steepest-descent: Riemannian quotient manifolds

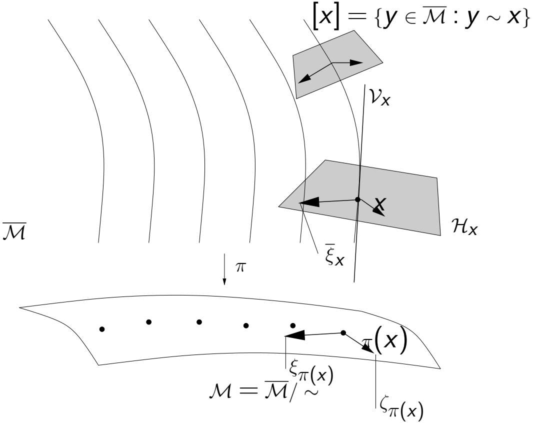


Then

$$g_{\pi(x)}(\xi_{\pi(x)}, \zeta_{\pi(x)}) := \overline{g}_x(\overline{\xi}_x, \overline{\zeta}_x).$$

defines a Riemannian metric on $\overline{\mathcal{M}}/\sim$. This turns $\overline{\mathcal{M}}/\sim$ into a Riemannian quotient manifold.

Steepest-descent: Riemannian quotient manifolds



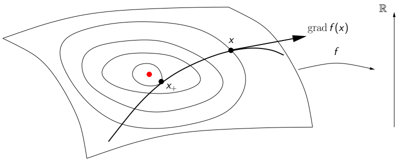
Let $f : \overline{\mathcal{M}}/\sim \to \mathbb{R}$. Let $\mathrm{P}_x^{h,\overline{g}}$ denote the orthogonal projection onto $\mathcal{H}_x$.

$$\overline{\mathrm{grad}\, f}_x = \mathrm{P}_x^{h,\overline{g}} \mathrm{grad}\,(f \circ \pi)(x).$$

If $\mathcal{H}_x$ is the orthogonal complement of $\mathcal{V}_x$ in the sense of $\overline{g}$ (π is a *Riemannian submersion*), then $\mathrm{grad}\,(f \circ \pi)(x)$ is already in $\mathcal{H}_x$, and thus

$$\overline{\mathrm{grad}\, f}_x = \mathrm{grad}\,(f \circ \pi)(x).$$

Steepest-descent: choosing the search curve



It remains to choose a curve γ through x at $t = 0$ such that

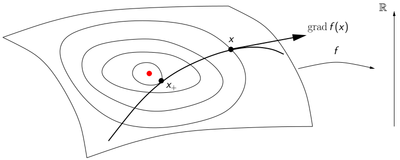
$$\dot{\gamma}(0) = -\mathrm{grad}\, f(x).$$

Let $R : T\mathcal{M} \to \mathcal{M}$ be a *retraction* on $\mathcal{M}$, that is

1. $R(0_x) = x$, where 0_x denotes the origin of $T_x\mathcal{M}$;
2. $\frac{\mathrm{d}}{\mathrm{d}t} R(t\xi_x) = \xi_x$.

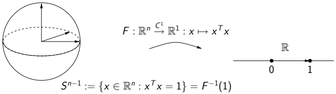
Then choose $\gamma : t \mapsto R(-t\mathrm{grad}\, f(x))$.

Steepest-descent: line-search procedure



Find t such that $f(\gamma(t))$ is "sufficiently smaller" than $f(\gamma(0))$. Since $t \mapsto f(\gamma(t))$ is just a function from $\mathbb{R}$ to $\mathbb{R}$, we can use the step selection techniques that are available for classical line-search methods.
For example: exact minimization, Armijo backtracking,...

Steepest-descent: Rayleigh quotient on unit sphere



Let the manifold be the unit sphere

$$S^{n-1} = \{x \in \mathbb{R}^n : x^T x = 1\} = F^{-1}(1),$$

where $F : \mathbb{R}^n \to \mathbb{R} : x \mapsto x^T x$.
Let $A = A^T \in \mathbb{R}^{n \times n}$ and let the cost function be the Rayleigh quotient

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

The tangent space to S^{n-1} at x is

$$T_x S^{n-1} = \ker(\mathrm{D}F(x)) = \{z \in \mathbb{R}^n : x^T z = 0\}.$$

Derivation formulas

If F is linear, then

$$\mathrm{D}F(x)[z] = F(z).$$

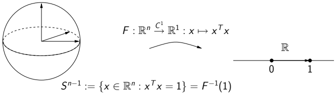
Chain rule: If $\mathrm{range}(F) \subseteq \mathrm{dom}(G)$, then

$$\mathrm{D}(G \circ F)(x)[z] = \mathrm{D}G(F(x))[\mathrm{D}F(x)[z]].$$

Product rule: If the ranges of F and G are in matrix spaces of compatible dimension, then

$$\mathrm{D}(FG)(x)[z] = \mathrm{D}F(x)[z]G(x) + F(x)\mathrm{D}G(x)[z].$$

Steepest-descent: Rayleigh quotient on unit sphere



Rayleigh quotient:

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

The tangent space to S^{n-1} at x is

$$T_x S^{n-1} = \ker(\mathrm{D}F(x)) = \{z \in \mathbb{R}^n : x^T z = 0\}.$$

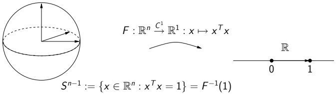
Product rule:

$$\mathrm{D}(FG)(x)[z] = \mathrm{D}F(x)[z]G(x) + F(x)\mathrm{D}G(x)[z].$$

Differential of f at $x \in S^{n-1}$:

$$\mathrm{D}f(x)[z] = x^T A z + z^T A x = 2z^T A x, \quad z \in T_x S^{n-1}.$$

Steepest-descent: Rayleigh quotient on unit sphere



"Natural" Riemannian metric on S^{n-1}:

$$g_x(z_1, z_2) = z_1^T z_2, \quad z_1, z_2 \in T_x S^{n-1}.$$

Differential of f at $x \in S^{n-1}$:

$$\mathrm{D}f(x)[z] = 2z^T Ax = 2g_x(z, Ax), \quad z \in T_x S^{n-1}.$$

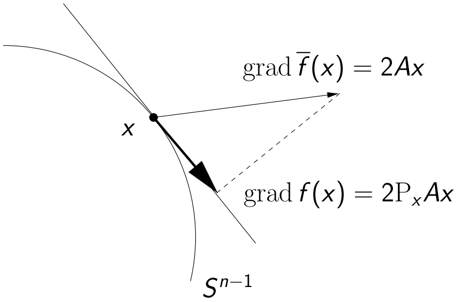
Gradient:

$$\operatorname{grad} f(x) = 2\mathrm{P}_x Ax = 2(I - xx^T)Ax.$$

Check:

$$\begin{cases} \operatorname{grad} f(x) \in T_x S^{n-1} \\ \mathrm{D}f(x)[z] = g_x(\operatorname{grad} f(x), z), \ \forall z \in T_x S^{n-1}. \end{cases}$$

Steepest-descent: Rayleigh quotient on unit sphere



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

$$\overline{f} : \mathbb{R}^n \to \mathbb{R} : x \mapsto x^T Ax$$

$$\operatorname{grad} \overline{f}(x) = 2Ax$$

$$\operatorname{grad} f(x) = 2\mathrm{P}_x Ax = 2(I - xx^T)Ax.$$

Newton's method on manifolds

Newton in $\mathbb{R}^n$

Let $f : \mathbb{R}^n \to \mathbb{R}$.
Recall $\operatorname{grad} f(x) = \begin{bmatrix} \partial_1 f(x) & \cdots & \partial_n f(x) \end{bmatrix}^T$.
Newton's iteration:

1. Solve, for the unknown $z \in \mathbb{R}^n$,

$$\mathrm{D}(\operatorname{grad} f)(x)[z] = -\operatorname{grad} f(x).$$

2. Set

$$x_+ = x + z.$$

Newton in $\mathbb{R}^n$: how it may fail

Let $f : \mathbb{R}^n_0 \to \mathbb{R} : x \mapsto \frac{x^T A x}{x^T x}$.
Newton's iteration:

1. Solve, for the unknown $z \in \mathbb{R}^n$,
$$\mathrm{D}(\operatorname{grad} f)(x)[z] = -\operatorname{grad} f(x).$$
2. Set
$$x_+ = x + z.$$

Proposition: For all x such that $f(x)$ is not an eigenvalue of A, we have

$$x_+ = 2x.$$

Newton: how to make it work for RQ

Let $f : S^{n-1} \to \mathbb{R} : x \mapsto \frac{x^T A x}{x^T x}$.
Newton's iteration:

1. Solve, for the unknown $z \in \mathbb{R}^n \rightsquigarrow \eta_x \in T_x S^{n-1}$

$$\mathrm{D}(\operatorname{grad} f)(x)[z] = -\operatorname{grad} f(x) \quad \rightsquigarrow \boxed{?}(\operatorname{grad} f)(x)[\eta_x] = -\operatorname{grad} f(x)$$

2. Set

$$x_+ = x + z \quad \rightsquigarrow x_+ = R(\eta_x)$$

Newton's equation on an abstract manifold

Let $\mathcal{M}$ be a manifold and let $f : \mathcal{M} \to \mathbb{R}$ be a cost function.
The mapping $x \in \mathcal{M} \;\mapsto\; \operatorname{grad} f(x) \in T_x\mathcal{M}$ is a *vector field*.

$$\mathrm{D}(\operatorname{grad} f)(x)[z] = -\operatorname{grad} f(x) \quad \rightsquigarrow \boxed{?}(\operatorname{grad} f)(x)[\eta_x] = -\operatorname{grad} f(x)$$

The new object has to be such that

- In $\mathbb{R}^n$, $\boxed{?}$ reduces to the classical derivative
- $\boxed{?}(\operatorname{grad} f)(x)[\eta_x]$ belongs to $T_x\mathcal{M}$
- $\boxed{?}$ has the same linearity properties and multiplication rule as the classical derivative.

Newton's equation on an abstract manifold

Let $\mathcal{M}$ be a manifold and let $f : \mathcal{M} \to \mathbb{R}$ be a cost function.
The mapping $x \in \mathcal{M} \;\mapsto\; \operatorname{grad} f(x) \in T_x\mathcal{M}$ is a *vector field*.

$$\mathrm{D}(\operatorname{grad} f)(x)[z] = -\operatorname{grad} f(x) \quad \rightsquigarrow \boxed{?}(\operatorname{grad} f)(x)[\eta_x] = -\operatorname{grad} f(x)$$

The new object has to be such that

- In $\mathbb{R}^n$, $\boxed{?}$ reduces to the classical derivative
- $\boxed{?}(\operatorname{grad} f)(x)[\eta_x]$ belongs to $T_x\mathcal{M}$
- $\boxed{?}$ has the same linearity properties and multiplication rule as the classical derivative.

Differential geometry offers a concept that matches these conditions: the concept of an *affine connection*.

Newton: affine connections

Let $\mathfrak{X}(\mathcal{M})$ denote the set of smooth vector fields on $\mathcal{M}$ and $\mathfrak{F}(\mathcal{M})$ the set of real-valued functions on $\mathcal{M}$.
An *affine connection* ∇ on a manifold $\mathcal{M}$ is a mapping

$$\nabla : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M}),$$

which is denoted by $(\eta, \xi) \xrightarrow{\nabla} \nabla_\eta \xi$ and satisfies the following properties:

i) $\mathfrak{F}(\mathcal{M})$-linearity in η: $\quad \nabla_{f\eta + g\chi}\xi = f\nabla_\eta \xi + g\nabla_\chi \xi$,
ii) $\mathbb{R}$-linearity in ξ: $\quad \nabla_\eta(a\xi + b\zeta) = a\nabla_\eta \xi + b\nabla_\eta \zeta$,
iii) Product rule (Leibniz' law): $\quad \nabla_\eta(f\xi) = (\eta f)\xi + f\nabla_\eta \xi$,

in which $\eta, \chi, \xi, \zeta \in \mathfrak{X}(\mathcal{M})$, $f, g \in \mathfrak{F}(\mathcal{M})$, and $a, b \in \mathbb{R}$.

Newton's method on abstract manifolds

Cost function: $f : \mathbb{R}^n \to \mathbb{R} \rightsquigarrow f : \mathcal{M} \to \mathbb{R}$.
Newton's iteration:

1. Solve, for the unknown $z \in \mathbb{R}^n \rightsquigarrow \eta_x \in T_x\mathcal{M}$

$$\mathrm{D}(\operatorname{grad} f)(x)[z] = -\operatorname{grad} f(x) \quad \rightsquigarrow \nabla(\operatorname{grad} f)(x)[\eta_x] = -\operatorname{grad} f(x)$$

2. Set

$$x_+ = x + z \quad \rightsquigarrow x_+ = R(\eta_x)$$

In the algorithm above, ∇ is an affine connection on $\mathcal{M}$ and R is a retraction on $\mathcal{M}$.

Newton's method on S^{n-1}

If $\mathcal{M}$ is a Riemannian submanifold of $\mathbb{R}^n$, then ∇ defined by

$$\nabla_{\eta_x}\xi = \mathrm{P}_x \mathrm{D}\xi(x)[\eta_x], \quad \eta_x \in T_x\mathcal{M},\ \xi \in \mathfrak{X}(\mathcal{M})$$

is a particular affine connection, called *Riemannian connection*.
For the unit sphere S^{n-1}, this yields

$$\nabla_{\eta_x}\xi = (I - xx^T)\mathrm{D}\xi(x)[\eta_x], \quad x^T\eta_x = 0.$$

Newton's method for Rayleigh quotient on S^{n-1}

Let $f : \begin{cases} \mathbb{R}^n \\ \mathcal{M} \\ S^{n-1} \end{cases} \rightarrow \mathbb{R} : x \mapsto \begin{cases} f(x) \\ f(x) \\ \frac{x^T A x}{x^T x} \end{cases}$.

Newton's iteration:

1. Solve, for the unknown $z \in \mathbb{R}^n \rightsquigarrow \eta_x \in T_x\mathcal{M} \rightsquigarrow x^T \eta_x = 0$

$$\mathrm{D}(\operatorname{grad} f)(x)[z] = -\operatorname{grad} f(x)$$
$$\rightsquigarrow \nabla(\operatorname{grad} f)(x)[\eta_x] = -\operatorname{grad} f(x)$$
$$\rightsquigarrow (I - xx^T)(A - f(x)I)\eta_x = -(I - xx^T)Ax$$

2. Set

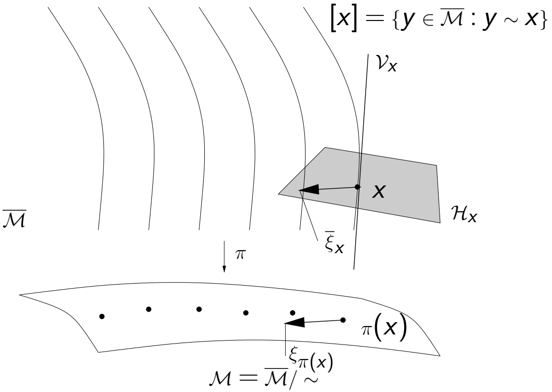
$$x_+ = x + z \quad \rightsquigarrow x_+ = R(\eta_x) \quad \rightsquigarrow x_+ = \frac{x + \eta_x}{\|x + \eta_x\|}$$

Newton for RQ on S^{n-1}: a closer look

$$
\begin{aligned}
(I - xx^T)(A - f(x)I)\eta_x &= -(I - xx^T)Ax \\
\Rightarrow (I - xx^T)(A - f(x)I)(x + \eta_x) &= 0 \\
\Rightarrow (A - f(x)I)(x + \eta_x) &= \alpha x
\end{aligned}
$$

Therefore, x_+ is collinear with $(A - f(x)I)^{-1}x$, which is the vector computed by the Rayleigh quotient iteration.

Newton method on quotient manifolds

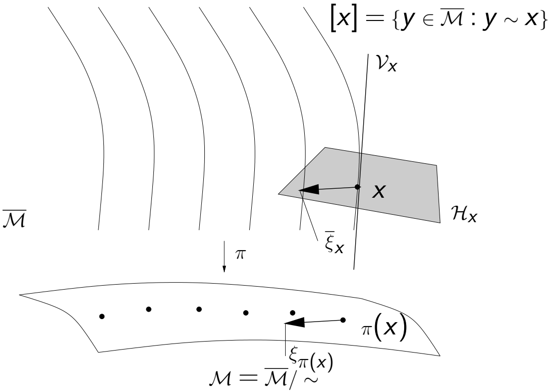


Affine connection: choose ∇ defined by

$$\overline{\nabla_\eta \xi}_x = \mathrm{P}^h_x \overline{\nabla}_{\overline{\eta}_x} \overline{\xi},$$

provided that this really defines a horizontal lift. This requires special choices of $\overline{\nabla}$.

Newton method on quotient manifolds



If $\pi : \overline{\mathcal{M}} \to \overline{\mathcal{M}}/\sim$ is a Riemannian submersion, then the Riemannian connection on $\overline{\mathcal{M}}/\sim$ is given by

$$\overline{\nabla_\eta \xi}_x = \mathrm{P}^h_x \overline{\nabla}_{\overline{\eta}_x} \overline{\xi},$$

where $\overline{\nabla}$ denotes the Riemannian connection on $\overline{\mathcal{M}}$.

A detailed exercise

Newton's method for the Rayleigh quotient on the Grassmann manifold

Manifold: Grassmann

The manifold is the Grassmann manifold of p-planes in $\mathbb{R}^n$:

$$\mathrm{Grass}(p, n) \simeq \mathrm{ST}(p, n)/\mathrm{GL}_p.$$

The one-to-one correspondence is

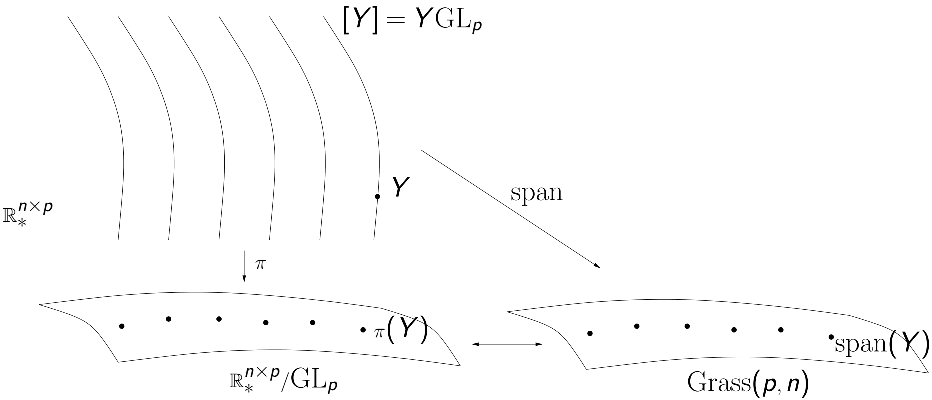
$$\mathrm{Grass}(p, n) \ni \mathcal{Y} \;\leftrightarrow\; Y\,\mathrm{GL}_p \in \mathrm{ST}(p, n)/\mathrm{GL}_p$$

such that $\mathcal{Y}$ is the column space of Y.
The quotient map

$$\pi : \mathrm{ST}(p, n) \to \mathrm{Grass}(p, n)$$

is the "column space" or "span" operation.

Grassmann and its quotient representation



Total space: the noncompact Stiefel manifold

The total space of the quotient is

$$\mathrm{ST}(p, n) = \{Y \in \mathbb{R}^{n \times p} : \mathrm{rank}(Y) = p\}.$$

This is an open submanifold of the Euclidean space $\mathbb{R}^{n \times p}$.
Tangent spaces: $T_Y \mathrm{ST}(p, n) \simeq \mathbb{R}^{n \times p}$.

Riemannian metric on the total space

Define a Riemannian metric $\overline{g}$ on $\mathrm{ST}(p, n)$ by

$$\overline{g}_Y(Z_1, Z_2) = \mathrm{trace}\left((Y^T Y)^{-1} Z_1^T Z_2\right).$$

This is not the canonical Riemannian metric, but it will allow us to turn the quotient map $\pi : \mathrm{ST}(p, n) \to \mathrm{Grass}(p, n)$ into a Riemannian submersion.

Vertical and horizontal spaces

The vertical spaces are the tangent spaces to the equivalence classes:

$$\mathcal{V}_Y := T_Y(Y\mathrm{GL}_p) = Y\ T_Y\mathrm{GL}_p = Y\mathbb{R}^{p\times p}.$$

Choice of horizontal space:

$$\begin{aligned}\mathcal{H}_Y &:= (\mathcal{V}_Y)^\perp \\ &= \{Z \in T_Y\mathrm{ST}(p,n) : \bar{g}_Y(Z,V) = 0, \forall V \in \mathcal{V}_Y\} \\ &= \{Z \in \mathbb{R}^{n\times p} : Y^T Z = 0\}.\end{aligned}$$

Horizontal projection:

$$\mathrm{P}^h_Y = (I - Y(Y^TY)^{-1}Y^T).$$

Compatibility equation for horizontal lifts

Given $\xi \in T_\pi(Y)\mathrm{Grass}(p, n)$, we have

$$\overline{\xi}_{YM} = \overline{\xi}_Y M.$$

To see this, observe that $\overline{\xi}_Y M$ is in $\mathcal{H}_{YM}$; moreover, since $YM + t\overline{\xi}_Y M$ and $Y + t\overline{\xi}_Y$ have the same column space for all t, one has

$$\mathrm{D}\pi(YM)[\overline{\xi}_Y M] = \mathrm{D}\pi(Y)[\overline{\xi}_Y] = \xi_{\pi(Y)}.$$

Thus $\overline{\xi}_Y M$ satisfies the conditions to be $\overline{\xi}_{YM}$.

Riemannian metric on the quotient

On $\mathrm{Grass}(p, n) \simeq \mathrm{ST}(p, n)/\mathrm{GL}_p$, define the Riemannian metric g by

$$g_{\pi(Y)}(\xi_{\pi(Y)}, \zeta_{\pi(Y)}) = \overline{g}_Y(\overline{\xi}_Y, \overline{\zeta}_Y).$$

This is well defined, because for all $\tilde{Y} \in \pi^{-1}(\pi(Y)) = Y\mathrm{GL}_p$, we have $\tilde{Y} = YM$ for some invertible M, and

$$\overline{g}_{YM}(\overline{\xi}_{YM}, \overline{\zeta}_{YM}) = \overline{g}_Y(\overline{\xi}_Y, \overline{\zeta}_Y).$$

This definition of g turns

$$\pi : (\mathrm{ST}(p, n), \overline{g}) \to (\mathrm{Grass}(p, n), g)$$

into a Riemannian submersion.

Cost function: Rayleigh quotient

Consider the cost function

$$f : \mathrm{Grass}(p, n) \to \mathbb{R} : \mathrm{span}(Y) \mapsto \mathrm{trace}\left((Y^T Y)^{-1} Y^T A Y\right).$$

This is the *projection* of

$$\overline{f} : \mathrm{ST}(p, n) \to \mathbb{R} : Y \mapsto \mathrm{trace}\left((Y^T Y)^{-1} Y^T A Y\right).$$

That is, $\overline{f} = f \circ \pi$.

Gradient of the cost function

For all $Z \in \mathbb{R}^{n \times p}$,

$$\mathrm{D}\overline{f}(Y)[Z] = 2\,\mathrm{trace}\left((Y^T Y)^{-1} Z^T (AY - Y(Y^T Y)^{-1} Y^T AY)\right).$$

Hence

$$\mathrm{grad}\,\overline{f}(Y) = 2\left(AY - Y(Y^T Y)^{-1} Y^T AY\right),$$

and

$$\overline{\mathrm{grad}\,f}_Y = 2\left(AY - Y(Y^T Y)^{-1} Y^T AY\right).$$

Riemannian connection

The quotient map is a Riemannian submersion. Therefore

$$\overline{\nabla_\eta \xi} = \mathrm{P}^h_Y \left(\overline{\nabla}_{\overline{\eta}_Y} \overline{\xi} \right)$$

It turns out that

$$\overline{\nabla_\eta \xi} = \mathrm{P}^h_Y \left(\mathrm{D}\overline{\xi}\,(Y)\,[\overline{\eta}_Y] \right).$$

(This is because the Riemanian metric $\overline{g}$ is "horizontally invariant".)
For the Rayleigh quotient f, this yields

$$\begin{aligned} \overline{\nabla_\eta \mathrm{grad}\, f} &= \mathrm{P}^h_Y \left(\mathrm{D}\overline{\mathrm{grad}\, f}\,(Y)\,[\overline{\eta}_Y] \right) \\ &= 2\,\mathrm{P}^h_Y \left(A\overline{\eta}_Y - \overline{\eta}_Y (Y^T Y)^{-1} Y^T A Y \right). \end{aligned}$$

Newton's equation

Newton's equation at $\pi(Y)$ is

$$\nabla_{\eta_{\pi(Y)}} \mathrm{grad}\, f = -\mathrm{grad}\, f(\pi(Y))$$

for the unknown $\eta_{\pi(Y)} \in T_{\pi(Y)}\mathrm{Grass}(p, n)$.
To turn this equation into a matrix equation, we take its horizontal lift. This yields

$$\mathrm{P}_Y^h \left(A\overline{\eta}_Y - \overline{\eta}_Y (Y^T Y)^{-1} Y^T AY \right) = -\mathrm{P}_Y^h AY, \qquad \overline{\eta}_Y \in \mathcal{H}_Y,$$

whose solution $\overline{\eta}_Y$ in the horizontal space $\mathcal{H}_Y$ is the horizontal lift of the solution η of the Newton equation.

Retraction

Newton's method sends $\pi(Y)$ to $\mathcal{Y}_+$ according to

$$\nabla_{\eta_{\pi(Y)}} \operatorname{grad} f = -\operatorname{grad} f(\pi(Y))$$
$$\mathcal{Y}_+ = R_{\pi(Y)}(\eta_{\pi(Y)}).$$

It remains to pick the retraction R.
Choice: R defined by

$$R_{\pi(Y)}\xi_{\pi(Y)} = \pi(Y + \overline{\xi}_Y).$$

(This is a well-defined retraction.)

Newton's iteration for RQ on Grassmann

Require: Symmetric matrix A.
Input: Initial iterate $Y_0 \in \mathrm{ST}(p, n)$.
Output: Sequence of iterates $\{Y_k\}$ in $\mathrm{ST}(p, n)$.

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: Solve the linear system

$$\begin{cases} \mathrm{P}^h_{Y_k}\left(AZ_k - Z_k(Y_k^T Y_k)^{-1} Y_k^T A Y_k\right) = -\mathrm{P}^h_{Y_k}(AY_k) \\ Y_k^T Z_k = 0 \end{cases}$$

for the unknown Z_k, where P^h_Y is the orthogonal projector onto $\mathcal{H}_Y$. (The condition $Y_k^T Z_k$ expresses that Z_k belongs to the horizontal space $\mathcal{H}_{Y_k}$.)

3: Set

$$Y_{k+1} = (Y_k + Z_k)N_k$$

where N_k is a nonsingular $p \times p$ matrix chosen for normalization purposes.

4: **end for**

Trust-region methods on Riemannian manifolds

Motivating application: Mechanical vibrations

Mass matrix M, stiffness matrix K.
Equation of vibrations (for undamped discretized linear structures):

$$Kx = \omega^2 Mx$$

were

- ω is an angular frequency of vibration
- x is the corresponding mode of vibration

Task: find lowest modes of vibration.

Generalized eigenvalue problem

Given $n \times n$ matrices $A = A^T$ and $B = B^T \succ 0$, there exist $v_1, \ldots, v_n$ in $\mathbb{R}^n$ and $\lambda_1 \leq \ldots \leq \lambda_n$ in $\mathbb{R}$ such that

$$Av_i = \lambda_i Bv_i$$
$$v_i^T Bv_j = \delta_{ij}.$$

Task: find $\lambda_1, \ldots, \lambda_p$ and $v_1, \ldots, v_p$.
We assume throughout that $\lambda_p < \lambda_{p+1}$.

Case $p = 1$: optimization in $\mathbb{R}^n$

$$Av_i = \lambda_i B v_i$$

Consider the Rayleigh quotient

$$\tilde{f} : \mathbb{R}^n_* \to \mathbb{R} : f(y) = \frac{y^T A y}{y^T B y}$$

Invariance: $\tilde{f}(\alpha y) = \tilde{f}(y)$.
Stationary points of $\tilde{f}$: αv_i, for all $\alpha \neq 0$.
Minimizers of $\tilde{f}$: αv_1, for all $\alpha \neq 0$.
Difficulty: the minimizers are not isolated.
Remedy: optimization on manifold.

Case $p = 1$: optimization on ellipsoid

$$\tilde{f} : \mathbb{R}^n_* \to \mathbb{R} : f(y) = \frac{y^T A y}{y^T B y}$$

Invariance: $\tilde{f}(\alpha y) = \tilde{f}(y)$.
Remedy 1:

- $\mathcal{M} := \{y \in \mathbb{R}^n : y^T B y = 1\}$, *submanifold* of $\mathbb{R}^n$.
- $f : \mathcal{M} \to \mathbb{R} : f(y) = y^T A y$.

Stationary points of f: $\pm v_1, \ldots, \pm v_n$.
Minimizers of f: $\pm v_1$.

Case $p = 1$: optimization on projective space

$$\tilde{f} : \mathbb{R}^n_* \to \mathbb{R} : f(y) = \frac{y^T A y}{y^T B y}$$

Invariance: $\tilde{f}(\alpha y) = \tilde{f}(y)$.

Remedy 2:

- $[y] := y\mathbb{R} := \{y\alpha : \alpha \in \mathbb{R}\}$
- $\mathcal{M} := \mathbb{R}^n_* / \mathbb{R} = \{[y]\}$
- $f : \mathcal{M} \to \mathbb{R} : f([y]) := \tilde{f}(y)$

Stationary points of f: $[v_1], \ldots, [v_n]$.

Minimizer of f: $[v_1]$.

Case $p \geq 1$: optimization on the Grassmann manifold

$$\tilde{f} : \mathbb{R}_*^{n \times p} \to \mathbb{R} : \tilde{f}(Y) = \operatorname{trace}\left((Y^T B Y)^{-1} Y^T A Y\right)$$

Invariance: $\tilde{f}(YR) = \tilde{f}(Y)$.
Define:

- $[Y] := \{YR : R \in \mathbb{R}_*^{p \times p}\}, \quad Y \in \mathbb{R}_*^{n \times p}$
- $\mathcal{M} := \mathrm{Grass}(p, n) := \{[Y]\}$
- $f : \mathcal{M} \to \mathbb{R} : f([Y]) := \tilde{f}(Y)$

Stationary points of f: $\mathrm{span}\{v_{i_1}, \ldots, v_{i_p}\}$.
Minimizer of f: $[Y] = \mathrm{span}\{v_1, \ldots, v_p\}$.

Optimization on Manifolds

- Luenberger [Lue73], Gabay [Gab82]: optimization on submanifolds of $\mathbb{R}^n$.
- Smith [Smi93, Smi94] and Udriște [Udr94]: optimization on general Riemannian manifolds (steepest descent, Newton, CG).
- ...
- PAA, Baker and Gallivan [ABG07]: trust-region methods on Riemannian manifolds.
- PAA, Mahony, Sepulchre [AMS08]:*Optimization Algorithms on Matrix Manifolds*, textbook.

The Problem : Leftmost Eigenpairs of Matrix Pencil

Given $n \times n$ matrix pencil (A, B), $A = A^T$, $B = B^T \succ 0$ with (unknown) eigen-decomposition

$$A\,[v_1|\ldots|v_n] = B\,[v_1|\ldots|v_n]\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$$

$[v_1|\ldots|v_n]^T\,B\,[v_1|\ldots|v_n] = I, \quad \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n.$
The problem is to compute the minor eigenvector $\pm v_1$.

The ideal algorithm

Given (A, B), $A = A^T$, $B = B^T \succ 0$ with (unknown) eigenvalues $0 < \lambda_1 \leq \ldots \lambda_n$ and associated eigenvectors $v_1, \ldots, v_n$.

1. Global convergence:
   - Convergence to some eigenvector for all initial conditions.
   - Stable convergence to the "leftmost" eigenvector $\pm v_1$ only.
2. Superlinear (cubic) local convergence to $\pm v_1$.

The ideal algorithm

Given (A, B), $A = A^T$, $B = B^T \succ 0$ with (unknown) eigenvalues $0 < \lambda_1 \leq \ldots \lambda_n$ and associated eigenvectors $v_1, \ldots, v_n$.

1. Global convergence:
   - Convergence to some eigenvector for all initial conditions.
   - Stable convergence to the "leftmost" eigenvector $\pm v_1$ only.
2. Superlinear (cubic) local convergence to $\pm v_1$.
3. "Matrix-free" (no factorization of A, B)
   but possible use of preconditioner.

The ideal algorithm

Given (A, B), $A = A^T$, $B = B^T \succ 0$ with (unknown) eigenvalues $0 < \lambda_1 \leq \ldots \lambda_n$ and associated eigenvectors $v_1, \ldots, v_n$.

1. Global convergence:
   - Convergence to some eigenvector for all initial conditions.
   - Stable convergence to the "leftmost" eigenvector $\pm v_1$ only.
2. Superlinear (cubic) local convergence to $\pm v_1$.
3. "Matrix-free" (no factorization of A, B)
   but possible use of preconditioner.
4. Minimal storage space required.

Strategy

- Rewrite computation of leftmost eigenpair as an optimization problem (on a manifold).

Strategy

- Rewrite computation of leftmost eigenpair as an optimization problem (on a manifold).
- Use a model-trust-region scheme to solve the problem.
  $\rightsquigarrow$ Global convergence.

Strategy

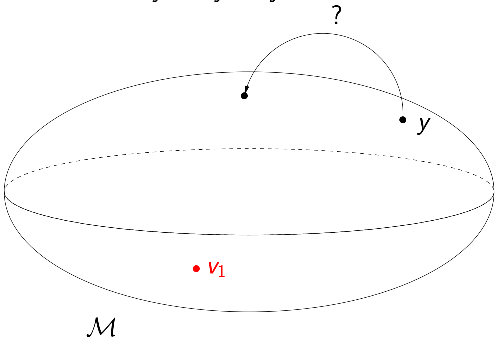
- Rewrite computation of leftmost eigenpair as an optimization problem (on a manifold).
- Use a model-trust-region scheme to solve the problem.
  $\rightsquigarrow$ Global convergence.
- Take the exact quadratic model (at least, close to the solution).
  $\rightsquigarrow$ Superlinear convergence.

Strategy

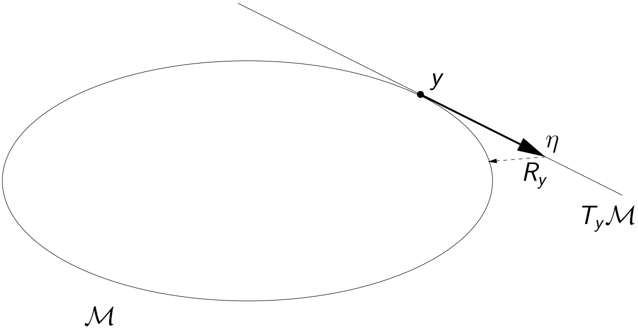
- Rewrite computation of leftmost eigenpair as an optimization problem (on a manifold).
- Use a model-trust-region scheme to solve the problem.
  $\rightsquigarrow$ Global convergence.
- Take the exact quadratic model (at least, close to the solution).
  $\rightsquigarrow$ Superlinear convergence.
- Solve the trust-region subproblems using the (Steihaug-Toint) truncated CG (tCG) algorithm.
  $\rightsquigarrow$ "Matrix-free", preconditioned iteration.
  $\rightsquigarrow$ Minimal storage of iteration vectors.

Iteration on the manifold

Manifold: ellipsoid $\mathcal{M} = \{y \in \mathbb{R}^n : y^T B y = 1\}$.
Cost function: $f : \mathcal{M} \to \mathbb{R} : y \mapsto y^T A y$



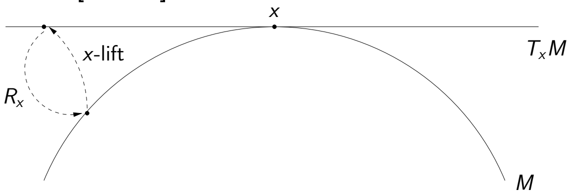
Tangent space and retraction (2D picture)



Tangent space: $T_y\mathcal{M} := \{\eta \in \mathbb{R}^n : y^T B\eta = 0\}$.
Retraction: $R_y\eta := (y+\eta)/\|y+\eta\|_B$.
Lifted cost function: $\hat{f}_y(\eta) := f(R_y\eta) = \frac{(y+\eta)^T A(y+\eta)}{(y+\eta)^T B(y+\eta)}$.

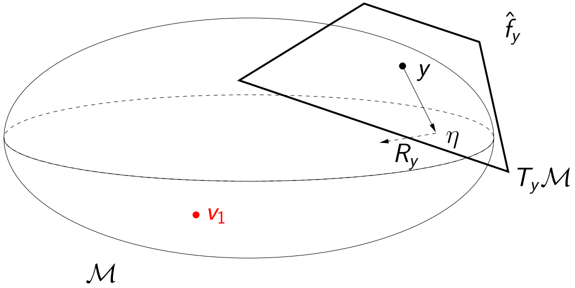
Concept of retraction

Introduced by Shub [Shu86].



1. R_x is defined and one-to-one in a neighbourhood of 0_x in T_xM.
2. $R_x(0_x) = x$.
3. $\mathrm{D}R_x(0_x) = \mathrm{id}_{T_xM}$, the identity mapping on T_xM, with the canonical identification $T_{0_x}T_xM \simeq T_xM$.

Tangent space and retraction



Tangent space: $T_y\mathcal{M} := \{\eta \in \mathbb{R}^n : y^T B\eta = 0\}$.
Retraction: $R_y\eta := (y+\eta)/\|y+\eta\|_B$.
Lifted cost function: $\hat{f}_y(\eta) := f(R_y\eta) = \frac{(y+\eta)^T A(y+\eta)}{(y+\eta)^T B(y+\eta)}$.

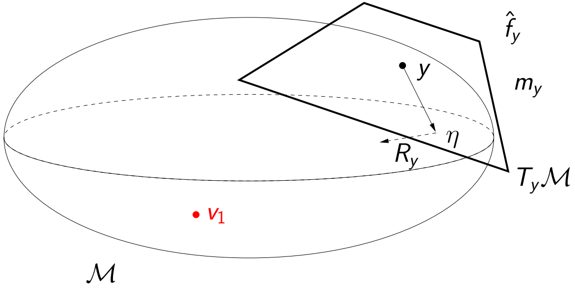
Quadratic model

$$\begin{aligned}\hat{f}_y(\eta) &= \frac{y^T A y}{y^T B y} + 2\frac{y^T A \eta}{y^T B y} + \frac{1}{y^T B y}\left(\eta^T A \eta - \frac{y^T A y}{y^T B y}\eta^T B \eta\right) + \dots \\ &= f(y) + 2\langle PAy, \eta\rangle + \frac{1}{2}\langle 2P(A - f(y)B)P\eta, \eta\rangle + \dots\end{aligned}$$

where $\langle u, v\rangle = u^T v$ and $P = I - By(y^T B^2 y)^{-1} y^T B$.
Model:

$$m_y(\eta) = f(y) + 2\langle PAy, \eta\rangle + \frac{1}{2}\langle P(A - f(y)B)P\eta, \eta\rangle, \quad y^T B\eta = 0.$$

Quadratic model



$$m_y(\eta) = f(y) + 2\langle PAy, \eta\rangle + \frac{1}{2}\langle P(A - f(y)B)P\eta, \eta\rangle, \quad y^T B\eta = 0.$$

Newton vs Trust-Region

Model:

$$m_y(\eta) = f(y) + 2\langle PAy, \eta\rangle + \frac{1}{2}\langle P(A - f(y)B)P\eta, \eta\rangle, \quad y^T B\eta = 0. \quad (1)$$

Newton vs Trust-Region

Model:

$$m_y(\eta) = f(y) + 2\langle PAy, \eta\rangle + \frac{1}{2}\langle P(A - f(y)B)P\eta, \eta\rangle, \quad y^T B\eta = 0. \quad (1)$$

Newton method: Compute the stationary point of the model, i.e., solve

$$P(A - f(y)B)P\,\eta = -PAy.$$

Newton vs Trust-Region

Model:

$$m_y(\eta) = f(y) + 2\langle PAy, \eta\rangle + \frac{1}{2}\langle P(A - f(y)B)P\eta, \eta\rangle, \quad y^T B\eta = 0. \quad (1)$$

Newton method: Compute the stationary point of the model, i.e., solve

$$P(A - f(y)B)P\,\eta = -PAy.$$

Instead, compute (approximately) the minimizer of m_y within a trust-region

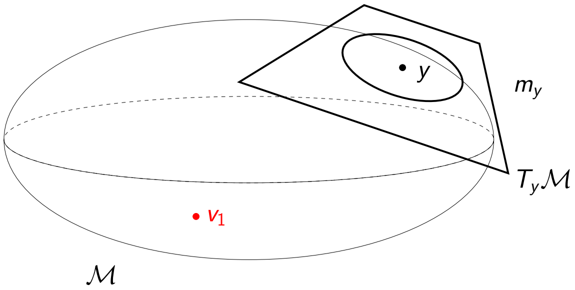
$$\{\eta \in T_x\mathcal{M} : \eta^T\eta \leq \Delta^2\}.$$

Trust-region subproblem

Minimize

$$m_y(\eta) = f(y) + 2\langle PAy, \eta\rangle + \frac{1}{2}\langle P(A - f(y)B)P\eta, \eta\rangle, \quad y^T B\eta = 0.$$

subject to $\eta^T\eta \leq \Delta^2$.



Truncated CG method for the TR subproblem (1)

Let $\langle \cdot, \cdot \rangle$ denote the standard inner product and let $\mathcal{H}_{x_k} := P(A - f(x_k)B)P$ denote the Hessian operator.

Initializations:

Set $\eta_0 = 0$, $r_0 = P_{x_k} A x_k = A x_k - B x_k (x_k^T B^2 x_k)^{-1} x_k^T B A x_k$, $\delta_0 = -r_0$;

Then repeat the following loop on j:

Check for negative curvature

if $\langle \delta_j, \mathcal{H}_{x_k} \delta_j \rangle \leq 0$

Compute τ such that $\eta = \eta_j + \tau \delta_j$ minimizes $m(\eta)$ in (1) and satisfies $\|\eta\| = \Delta$;

return η;

Truncated CG method for the TR subproblem (2)

Generate next inner iterate

Set $\alpha_j = \langle r_j, r_j \rangle / \langle \delta_j, \mathcal{H}_{x_k} \delta_j \rangle$;

Set $\eta_{j+1} = \eta_j + \alpha_j \delta_j$;

Check trust-region

if $\|\eta_{j+1}\| \geq \Delta$

Compute $\tau \geq 0$ such that $\eta = \eta_j + \tau \delta_j$ satisfies $\|\eta\| = \Delta$;

return η;

Truncated CG method for the TR subproblem (3)

Update residual and search direction

Set $r_{j+1} = r_j + \alpha_j \mathcal{H}_{x_k} \delta_j$;

Set $\beta_{j+1} = \langle r_{j+1}, r_{j+1} \rangle / \langle r_j, r_j \rangle$;

Set $\delta_{j+1} = -r_{j+1} + \beta_{j+1} \delta_j$;

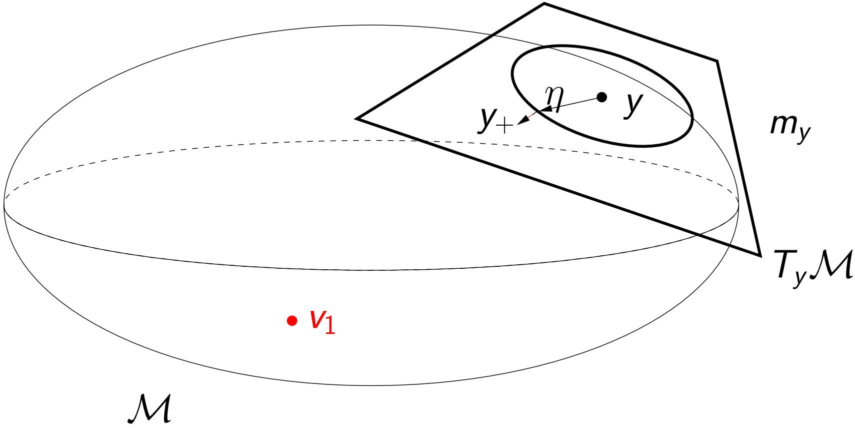
$j \leftarrow j + 1$;

Check residual

If $\|r_j\| \leq \|r_0\| \min \left(\|r_0\|^\theta, \kappa \right)$ for some prescribed θ and κ

return η_j;

Overall iteration



The outer iteration – manifold trust-region (1)

Data: symmetric $n \times n$ matrices A and B, with B positive definite.
Parameters: $\bar{\Delta} > 0$, $\Delta_0 \in (0, \bar{\Delta})$, and $\rho' \in (0, \frac{1}{4})$.
Input: initial iterate $x_0 \in \{y : y^T By = 1\}$.
Output: sequence of iterates $\{x_k\}$ in $\{y : y^T By = 1\}$.
Initialization: $k = 0$
Repeat the following:

The outer iteration – manifold trust-region (2)

- Obtain η_k using the Steihaug-Toint truncated conjugate-gradient method to approximately solve the trust-region subproblem

$$\min_{x_k^T B\eta=0} m_{x_k}(\eta) \quad \text{s.t. } \|\eta\| \leq \Delta_k, \tag{2}$$

where m is defined in (1).

The outer iteration – manifold trust-region (3)

- Evaluate
$$\rho_k = \frac{\hat{f}_{x_k}(0) - \hat{f}_{x_k}(\eta_k)}{m_{x_k}(0) - m_{x_k}(\eta_k)} \tag{3}$$
where $\hat{f}_{x_k}(\eta) = \frac{(x_k+\eta)^T A (x_k+\eta)}{(x_k+\eta)^T B (x_k+\eta)}$.
- Update the trust-region radius:
  if $\rho_k < \frac{1}{4}$
  $\quad \Delta_{k+1} = \frac{1}{4}\Delta_k$
  else if $\rho_k > \frac{3}{4}$ **and** $\|\eta_k\| = \Delta_k$
  $\quad \Delta_{k+1} = \min(2\Delta_k, \bar{\Delta})$
  else
  $\quad \Delta_{k+1} = \Delta_k$;

The outer iteration – manifold trust-region (4)

- Update the iterate:
  if $\rho_k > \rho'$
  $$x_{k+1} = (x_k + \eta_k)/\|x_k + \eta_k\|_B; \tag{4}$$
  else
  $x_{k+1} = x_k;$
  $k \leftarrow k + 1$

Strategy

- Rewrite computation of leftmost eigenpair as an optimization problem (on a manifold).
- Use a model-trust-region scheme to solve the problem.
  $\rightsquigarrow$ Global convergence.
- Take the exact quadratic model (at least, close to the solution).
  $\rightsquigarrow$ Superlinear convergence.
- Solve the trust-region subproblems using the (Steihaug-Toint) truncated CG (tCG) algorithm.
  $\rightsquigarrow$ "Matrix-free", preconditioned iteration.
  $\rightsquigarrow$ Minimal storage of iteration vectors.

Summary

We have obtained a trust-region algorithm for minimizing the Rayleigh quotient over an ellipsoid.

Summary

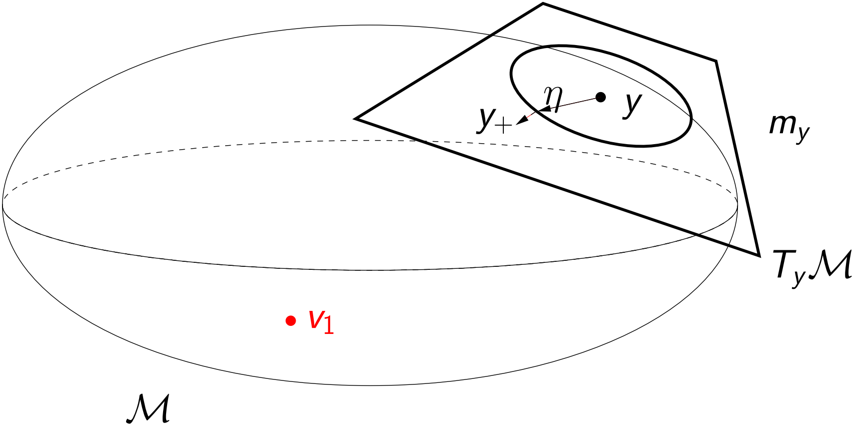
We have obtained a trust-region algorithm for minimizing the Rayleigh quotient over an ellipsoid.

Summary

We have obtained a trust-region algorithm for minimizing the Rayleigh quotient over an ellipsoid.

Generalization to trust-region algorithms for minimizing functions on manifolds: the Riemannian Trust-Region (RTR) method [ABG07].

Convergence analysis



Global convergence of Riemannian Trust-Region algorithms

Let $\{x_k\}$ be a sequence of iterates generated by the RTR algorithm with $\rho' \in (0, \frac{1}{4})$. Suppose that f is C^2 and bounded below on the level set $\{x \in M : f(x) < f(x_0)\}$. Suppose that $\|\mathrm{grad}\, f(x)\| \leq \beta_g$ and $\|\mathrm{Hess}\, f(x)\| \leq \beta_H$ for some constants β_g, β_H, and all $x \in M$. Moreover suppose that

$$\|\frac{D}{dt}\frac{d}{dt}Rt\xi\| \leq \beta_D \tag{5}$$

for some constant β_D, for all $\xi \in TM$ with $\|\xi\| = 1$ and all $t < \delta_D$, where $\frac{D}{dt}$ denotes the covariant derivative along the curve $t \mapsto Rt\xi$. Further suppose that all approximate solutions η_k of the trust-region subproblems produce a decrease of the model that is at least a fixed fraction of the Cauchy decrease.

Global convergence (cont'd)

It then follows that

$$\lim_{k\to\infty} \operatorname{grad} f(x_k) = 0.$$

And only the local minima are stable (the saddle points and local maxima are unstable).

Local convergence of Riemannian Trust-Region algorithms

Consider the RTR-tCG algorithm. Suppose that f is a C^2 cost function on M and that

$$\|\mathcal{H}_k - \mathrm{Hess}\,\hat{f}_{x_k}(0_k)\| \leq \beta_{\mathcal{H}} \|\mathrm{grad}\, f(x_k)\|. \tag{6}$$

Let $v \in M$ be a nondegenerate local minimum of f, (i.e., $\mathrm{grad}\, f(v) = 0$ and $\mathrm{Hess}\, f(v)$ is positive definite). Further assume that $\mathrm{Hess}\,\hat{f}_{x_k}$ is Lipschitz-continuous at 0_x uniformly in x in a neighborhood of v, i.e., there exist $\beta_1 > 0$, $\delta_1 > 0$ and $\delta_2 > 0$ such that, for all $x \in B_{\delta_1}(v)$ and all $\xi \in B_{\delta_2}(0_x)$, it holds

$$\|\mathrm{Hess}\,\hat{f}_{x_k}(\xi) - \mathrm{Hess}\,\hat{f}_{x_k}(0_{x_k})\| \leq \beta_{L2} \|\xi\|. \tag{7}$$

Local convergence (cont'd)

Then there exists $c > 0$ such that, for all sequences $\{x_k\}$ generated by the RTR-tCG algorithm converging to v, there exists $K > 0$ such that for all $k > K$,

$$\mathrm{dist}(x_{k+1}, v) \leq c\,(\mathrm{dist}(x_k, v))^{\min\{\theta+1,2\}}, \tag{8}$$

where θ governs the stopping criterion of the tCG inner iteration.

Convergence of trust-region-based eigensolver

Theorem:

Let (A, B) be an $n \times n$ symmetric/positive-definite matrix pencil with eigenvalues $\lambda_1 < \lambda_2 \leq \ldots \leq \lambda_{n-1} \leq \lambda_n$ and an associated B-orthonormal basis of eigenvectors $(v_1, \ldots, v_n)$.

Let $\mathcal{S}_i = \{y : Ay = \lambda_i By,\ y^T By = 1\}$ denote the intersection of the eigenspace of (A, B) associated to λ_i with the set $\{y : y^T By = 1\}$.

...

Convergence (global)

(i) Let $\{x_k\}$ be a sequence of iterates generated by the Algorithm. Then $\{x_k\}$ converges to the eigenspace of (A, B) associated to one of its eigenvalues. That is, there exists i such that $\lim_{k\to\infty} \mathrm{dist}(x_k, \mathcal{S}_i) = 0$.

(ii) Only the set $\mathcal{S}_1 = \{\pm v_1\}$ is stable.

Convergence (local)

(iii) There exists $c > 0$ such that, for all sequences $\{x_k\}$ generated by the Algorithm converging to $\mathcal{S}_1$, there exists $K > 0$ such that for all $k > K$,

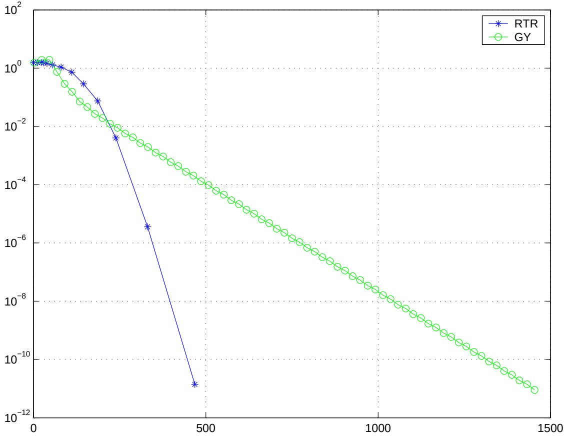
$$\mathrm{dist}(x_{k+1}, \mathcal{S}_1) \leq c\,(\mathrm{dist}(x_k, \mathcal{S}_1))^{\min\{\theta+1,2\}} \tag{9}$$

with $\theta > 0$.

Strategy

- Rewrite computation of leftmost eigenpair as an optimization problem (on a manifold).
- Use a model-trust-region scheme to solve the problem.
  $\rightsquigarrow$ Global convergence.
- Take the exact quadratic model (at least, close to the solution).
  $\rightsquigarrow$ Superlinear convergence.
- Solve the trust-region subproblems using the (Steihaug-Toint) truncated CG (tCG) algorithm.
  $\rightsquigarrow$ "Matrix-free", preconditioned iteration.
  $\rightsquigarrow$ Minimal storage of iteration vectors.

Numerical experiments: RTR vs Krylov [GY02]



Distance to target versus matrix-vector multiplications.
Symmetric/positive-definite generalized eigenvalue problem.

A new tool for Optimization On Manifolds:

Vector Transport

Filling a gap

	Purely Riemannian way	Pragmatic way
Update	Search along the geodesic tangent to the search direction	Search along any curve tangent to the search direction (prescribed by a *retraction*)
Displacement of tgt vectors	Parallel translation induced by $\overset{g}{\nabla}$	??

Where do we use parallel translation?

In CG. Quoting (approximately) Smith (1994):

1. Select $x_0 \in \mathcal{M}$, compute $\eta_0 = -\operatorname{grad} f(x_0)$, and set $k = 0$
2. Compute t_k such that $f(\operatorname{Exp}_{x_k}(t_k \eta_k)) \leq f(\operatorname{Exp}_{x_k}(t \eta_k))$ for all $t \geq 0$.
3. Set $x_{k+1} = \operatorname{Exp}_{x_k}(t_k \eta_k)$.
4. Set $\eta_{k+1} = -\operatorname{grad} f(x_{k+1}) + \beta_{k+1} \tau \eta_k$, where τ is the parallel translation along the geodesic from x_k to x_{k+1}. Increment k and go to step 2.

Where do we use parallel translation?

In BFGS. Quoting (approximately) Gabay (1982):
$x_{k+1} = \mathrm{Exp}_{x_k}(t_k \xi_k)$ (update along geodesic)
$\mathrm{grad}\, f(x_{k+1}) - \tau_0^{t_k} \mathrm{grad}\, f(x_k) = B_{k+1} \tau_0^{t_k}(t_k \xi_k)$ (requirement on approximate Jacobian B)
This leads to the a *generalized BFGS update formula* involving parallel translation.

Where else could we use parallel translation?

In finite-difference quasi-Newton.

Let ξ be a vector field on a Riemannian manifold $\mathcal{M}$. Exact Jacobian of ξ at $x \in \mathcal{M}$: $J_\xi(x)[\eta] = \nabla_\eta \xi$.

Finite difference approximation to J_ξ: choose a basis $(E_1, \cdots, E_d)$ of $T_x\mathcal{M}$ and define $\tilde{J}(x)$ as the linear operator that satisfies

$$\tilde{J}(x)[E_i] = \frac{\tau_h^0 \xi_{\mathrm{Exp}_x(hE_i)} - \xi_x}{h}.$$

Filling a gap

	Purely Riemannian way	Pragmatic way
Update	Search along the geodesic tangent to the search direction	Search along any prescribed curve tangent to the search direction
Displacement of tgt vectors	Parallel translation induced by $\overset{g}{\nabla}$	??

Parallel translation can be tough

Edelman et al (1998): We are unaware of any closed form expression for the parallel translation on the Stiefel manifold (defined with respect to the Riemannian connection induced by the embedding in $\mathbb{R}^{n\times p}$).
Parallel transport along geodesics on Grassmannians:

$$\overline{\xi(t)}_{Y(t)} = -Y_0 V \sin(\Sigma t) U^T \overline{\xi(0)}_{Y_0} + U \cos(\Sigma t) U^T \overline{\xi(0)}_{Y_0} + (I - UU^T) \overline{\xi(0)}_{Y_0}.$$

where $\overline{\dot{\mathcal{Y}}(0)}_{Y_0} = U \Sigma V^T$ is a thin SVD.

Alternatives found in the literature

Edelman et al (1998): "extrinsic" CG algorithm. "Tangency of the search direction at the new point is imposed via the projection $I - YY^T$" (instead of via parallel translation).

Brace & Manton (2006), *An improved BFGS-on-manifold algorithm for computing weighted low rank approximation*. "The second change is that parallel translation is not defined with respect to the Levi-Civita connection, but rather is all but ignored."

Filling a gap

	Purely Riemannian way	Pragmatic way
Update	Search along the geodesic tangent to the search direction	Search along any curve tangent to the search direction (prescribed by a *retraction*)
Displacement of tgt vectors	Parallel translation induced by $\overset{g}{\nabla}$	??

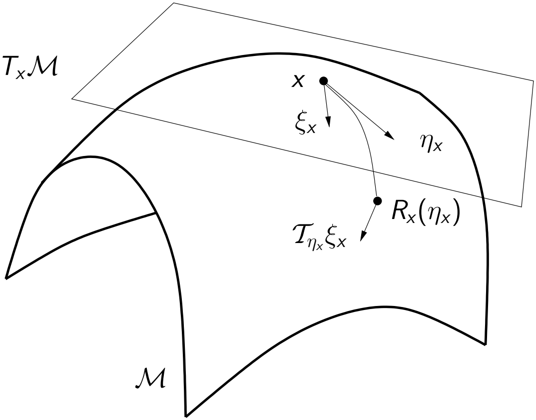
Filling a gap: Vector Transport

	Purely Riemannian way	Pragmatic way
Update	Search along the geodesic tangent to the search direction	Search along any curve tangent to the search direction (prescribed by a *retraction*)
Displacement of tgt vectors	Parallel translation induced by $\overset{g}{\nabla}$	Vector Transport

Still to come

- Vector transport in one picture
- Formal definition
- Particular vector transports
- Applications: finite-difference Newton, BFGS, CG.

The concept of vector transport



Retraction

A *retraction* on a manifold $\mathcal{M}$ is a smooth mapping

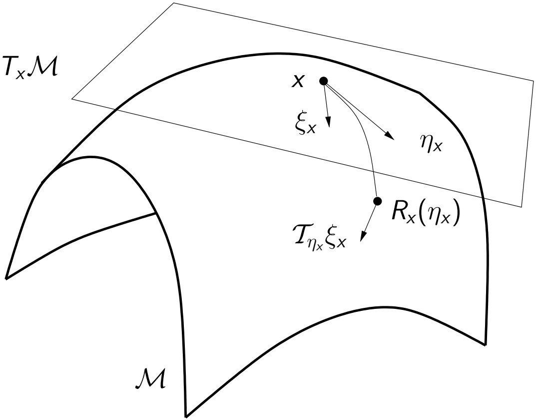
$$R : T\mathcal{M} \to \mathcal{M}$$

such that

1. $R(0_x) = x$ for all $x \in \mathcal{M}$, where 0_x denotes the origin of $T_x\mathcal{M}$;
2. $\frac{d}{dt}R(t\xi_x)\big|_{t=0} = \xi_x$ for all $\xi_x \in T_x\mathcal{M}$.

Consequently, the curve $t \mapsto R(t\xi_x)$ is a curve on $\mathcal{M}$ tangent to ξ_x.

The concept of vector transport – Whitney sum



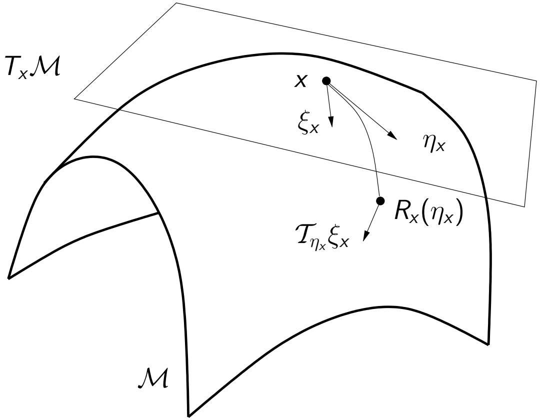
Whitney sum

Let $T\mathcal{M} \oplus T\mathcal{M}$ denote the set

$$T\mathcal{M} \oplus T\mathcal{M} = \{(\eta_x, \xi_x) : \eta_x, \xi_x \in T_x\mathcal{M},\ x \in \mathcal{M}\}.$$

This set admits a natural manifold structure.

The concept of vector transport – definition



Vector transport: definition

A *vector transport* on a manifold $\mathcal{M}$ on top of a retraction R is a smooth map

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

satisfying the following properties for all $x \in \mathcal{M}$:

1. (Underlying retraction) $\mathcal{T}_{\eta_x}\xi_x$ belongs to $T_{R_x(\eta_x)}\mathcal{M}$.
2. (Consistency) $\mathcal{T}_{0_x}\xi_x = \xi_x$ for all $\xi_x \in T_x\mathcal{M}$;
3. (Linearity) $\mathcal{T}_{\eta_x}(a\xi_x + b\zeta_x) = a\mathcal{T}_{\eta_x}(\xi_x) + b\mathcal{T}_{\eta_x}(\zeta_x)$.

Inverse vector transport

When it exists, $(\mathcal{T}_{\eta_x})^{-1}(\xi_{R_x(\eta_x)})$ belongs to $T_x\mathcal{M}$. If η and ξ are two vector fields on $\mathcal{M}$, then $(\mathcal{T}_{\eta})^{-1}\xi$ is naturally defined as the vector field satisfying

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

Still to come

- Vector transport in one picture
- Formal definition
- Particular vector transports
- Applications: finite-difference Newton, BFGS, CG.

Parallel translation is a vector transport

Proposition

If ∇ is an affine connection and R is a retraction on a manifold $\mathcal{M}$, then

$$\mathcal{T}_{\eta_x}(\xi_x) := P_{\gamma}^{1 \leftarrow 0} \xi_x \tag{10}$$

is a vector transport with associated retraction R, where P_γ denotes the parallel translation induced by ∇ along the curve $t \mapsto \gamma(t) = R_x(t\eta_x)$.

Vector transport on Riemannian submanifolds

If $\mathcal{M}$ is an embedded submanifold of a Euclidean space $\mathcal{E}$ and $\mathcal{M}$ is endowed with a retraction R, then we can rely on the natural inclusion $T_y\mathcal{M} \subset \mathcal{E}$ for all $y \in \mathcal{N}$ to simply define the vector transport by

$$\mathcal{T}_{\eta_x}\xi_x := \mathrm{P}_{R_x(\eta_x)}\xi_x, \tag{11}$$

where P_x denotes the orthogonal projector onto $T_x\mathcal{N}$.

Still to come

- Vector transport in one picture
- Formal definition
- Particular vector transports
- Applications: finite-difference Newton, BFGS, CG.

Vector transport in finite differences

Let $\mathcal{M}$ be a manifold endowed with a vector transport $\mathcal{T}$ on top of a retraction R. Let $x \in \mathcal{M}$ and let $(E_1, \ldots, E_d)$ be a basis of $T_x\mathcal{M}$. Given a smooth vector field ξ and a real constant $h > 0$, let $\tilde{J}_\xi(x) : T_x\mathcal{M} \to T_x\mathcal{M}$ be the linear operator that satisfies, for $i = 1, \ldots, d$,

$$\tilde{J}_\xi(x)[E_i] = \frac{(\mathcal{T}_{hE_i})^{-1}\xi_{R(hE_i)} - \xi_x}{h}. \tag{12}$$

Lemma (finite differences)

Let x_* be a nondegenerate zero of ξ. Then there is $c > 0$ such that, for all x sufficiently close to x_* and all h sufficiently small, it holds that

$$\|\tilde{J}_\xi(x)[E_i] - J(x)[E_i]\| \leq c(h + \|\xi_x\|). \tag{13}$$

Convergence of Newton's method with finite differences

Proposition

Consider the geometric Newton method where the exact Jacobian $J(x_k)$ is replaced by the operator $\tilde{J}_\xi(x_k)$ with $h := h_k$. If

$$\lim_{k \to \infty} h_k = 0,$$

then the convergence to nondegenerate zeros of ξ is superlinear. If, moreover, there exists some constant c such that

$$h_k \leq c \|\xi_{x_k}\|$$

for all k, then the convergence is (at least) quadratic.

Vector transport in BFGS

With the notation

$$
\begin{aligned}
s_k &:= \mathcal{T}_{\eta_k}\eta_k \in T_{x_{k+1}}\mathcal{M}, \\
y_k &:= \operatorname{grad} f(x_{k+1}) - \mathcal{T}_{\eta_k}(\operatorname{grad} f(x_k)) \in T_{x_{k+1}}\mathcal{M},
\end{aligned}
$$

we define the operator $A_{k+1} : T_{x_{k+1}}\mathcal{M} \mapsto T_{x_{k+1}}\mathcal{M}$ by

$$
A_{k+1}\eta = \tilde{A}_k\eta - \frac{\langle s_k, \tilde{A}_k\eta\rangle}{\langle s_k, \tilde{A}_k s_k\rangle}\tilde{A}_k s_k + \frac{\langle y_k, \eta\rangle}{\langle y_k, s_k\rangle} y_k \quad \text{for all } \eta \in T_{x_{k+1}}\mathcal{M},
$$

with

$$
\tilde{A}_k = \mathcal{T}_{\eta_k} \circ A_k \circ (\mathcal{T}_{\eta_k})^{-1}.
$$

Vector transport in CG

Compute a step size α_k and set

$$x_{k+1} = R_{x_k}(\alpha_k \eta_k). \tag{14}$$

Compute β_{k+1} and set

$$\eta_{k+1} = -\mathrm{grad}\, f(x_{k+1}) + \beta_{k+1} \mathcal{T}_{\alpha_k \eta_k}(\eta_k). \tag{15}$$

Filling a gap: Vector Transport

	Purely Riemannian way	Pragmatic way
Update	Search along the geodesic tangent to the search direction	Search along any curve tangent to the search direction (prescribed by a *retraction*)
Displacement of tgt vectors	Parallel translation induced by $\overset{g}{\nabla}$	Vector Transport

Ongoing work

- Use vector transport wherever we can.
- Extend convergence analyses.
- Develop recipies for building efficient vector transports.

BFGS Algorithm on Manifolds

Source: *Riemannian BFGS algorithm with applications*. Chunhong Qi, Kyle A. Gallivan, P.-A. Absil. Recent Advances in Optimization and its Applications in Engineering, Springer-Verlag, pp. 183-192, 2010. URL: `http://www.inma.ucl.ac.be/~absil/Publi/Qi_RBFGS.htm`

A (questionable) historical overview

	In $\mathbb{R}^n$	On Riemannian manifolds	
		using classical objects	using novel objects
Steepest descent	1966 (Armijo backtracking)	1972 (Luenberger)	1986–2008 ?
Newton	1740 (Simpson)	1993 (Smith)	2002 (Adler et al.)
Conjugate Grad	1964 (Fletcher–Reeves)	1993 (Smith)	2008 (PAA, Mahony, Sepulchre) ?
Trust regions	1985 (name created by Celis, Dennis, Tapia)	2007 (PAA, Baker, Gallivan)	2007 (PAA, Baker, Gallivan)
BFGS	1970 (B-F-G-S)	1982 (Gabay)	2010 (Qi, Gallivan, PAA)

Background on classical BFGS

- BFGS stands for Broyden–Fletcher–Goldfarb–Shanno.
- BFGS is a *quasi-Newton method*, where the Hessian found in the pure Newton is replaced by an approximation $\mathcal{B}_k$.
- The approximation $\mathcal{B}_k$ undergoes a rank-two update at each iteration and satisfies the *secant condition*:

$$\mathcal{B}_{k+1}(x_{k+1} - x_k) = \operatorname{grad} f(x_{k+1}) - \operatorname{grad} f(x_k).$$

Symmetric secant update (PSB)

- Let $s_k = x_{k+1} - x_k$ and $y_k = \mathrm{grad}\, f(x_{k+1}) - \mathrm{grad}\, f(x_k)$. Then the secant condition becomes

$$\mathcal{B}_{k+1} s_k = y_k.$$

- What is $\mathcal{B}_{k+1}$ that minimizes $\|\mathcal{B}_{k+1} - \mathcal{B}_k\|_F$ subject to $\mathcal{B}_{k+1} s_k = y_k$ and $\mathcal{B}_{k+1} - \mathcal{B}_k$ symmetric?
  Answer given by the *symmetric secant update*, also called *Powell-symmetric-Broyden* (PSB) update:

$$\mathcal{B}_{k+1} = \mathcal{B}_k + \frac{(y_k - \mathcal{B}_k s_k) s_k^T + s_k^T (y_k - \mathcal{B}_k s_k)^T}{s_k^T s_k} - \frac{\langle y_k - \mathcal{B}_k s_k, s_k \rangle s_k s_k^T}{(s_k^T s_k)^2}$$

- Drawback: $\mathcal{B}_{k+1}$ is not necessarily positive-definite. Hence the next search direction $\eta_k = -\mathcal{B}_k^{-1} \mathrm{grad}\, f(\mathbf{x}_k)$ may not be a descent direction.

Positive-definite secant update (BFGS)

- Let $s_k = x_{k+1} - x_k$ and $y_k = \mathrm{grad}\, f(x_{k+1}) - \mathrm{grad}\, f(x_k)$. Then the secant condition becomes

$$\mathcal{B}_{k+1} s_k = y_k.$$

- Let also $\mathcal{B}_k = LL^T$ be the Cholesky factorization.
- What is $\mathcal{B}_{k+1} = JJ^T$ with J nonsingular (guaranties $\mathcal{B}_{k+1}$ symmetric positive definite) such that $\mathcal{B}_{k+1} s_k = y_k$ and $\|J - L\|_F$ as small as possible?
  Answer given by the *positive definite secant update*, discovered independently by Broyden, Fletcher, Goldfarb and Shanno (BFGS) in 1970:

$$\mathcal{B}_{k+1} = \mathcal{B}_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{\mathcal{B}_k s_k (\mathcal{B}_k s_k)^T}{s_k^T \mathcal{B}_k s_k},$$

  iff $s_k^T y_k > 0$. Otherwise, no solution.

Formulation of classical BFGS (in $\mathbb{R}^n$)

Algorithm 1 The classical BFGS algorithm (in $\mathbb{R}^n$)

1: Given: real-valued function f on $\mathbb{R}^n$; initial iterate $x_1 \in \mathbb{R}^n$; initial Hessian approximation $\mathcal{B}_1$;
2: **for** k = 1, 2,. . . **do**
3: Obtain $\eta_k \in \mathbb{R}^n$ by solving: $\eta_k = -\mathcal{B}_k^{-1} \mathrm{grad}\, f(\mathbf{x}_k)$.
4: Perform a line search to obtain a step size α_k and set $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \eta_k$.
5: Set $s_k := \alpha_k \eta_k$
6: Set $y_k := \mathrm{grad}\, f(\mathbf{x}_{k+1}) - \mathrm{grad}\, f(\mathbf{x}_k)$
7: $\mathcal{B}_{k+1} = \mathcal{B}_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{\mathcal{B}_k s_k (\mathcal{B}_k s_k)^T}{s_k^T \mathcal{B}_k s_k}$.
8: **end for**

Significant Riemannian Manifolds

Sphere S^{n-1}

The manifold of unit sphere:

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

Compact Stiefel Manifold

The manifold of orthonormal bases:

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

Grassmann manifold

Manifold of linear subspaces:

$$\mathrm{Grass}((,k),n) = \{k\text{-dimensional subspaces of } \mathbb{R}^n\}$$

Applications

- computing the leftmost eigenvector of A (S^{n-1})

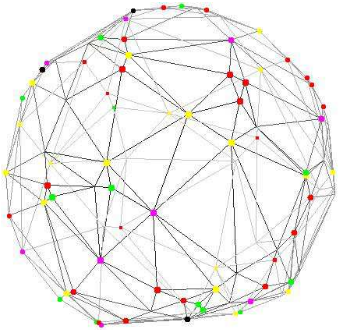
$$f : S^{n-1} \to \mathbb{R} : x \mapsto x^T Ax, A = A^T$$

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

$$f : \mathrm{St}(p,n) \to \mathbb{R} : Q \to \|AQ - QB\|_F, A : n \times n, B : p \times p$$

Application

- Thomson Problem
  $(S^{n-1} \times \cdots \times S^{n-1})$
- $f : [x_1, x_2, \cdots, x_N] \longmapsto \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{1}{\|x_i - x_j\|^2}$
- Optimally arrange N repulsive particles on a sphere
- Determining the minimum energy configuration of these particles



Applet: `http://thomson.phy.syr.edu/thomsonapplet.htm`

- The weighted low rank approximation problem on Grass(n, k):

$$\min_{\substack{R \in \mathbb{R}^{p\times n} \\ \mathrm{rank}\{R\} \leq r}} \|X - R\|_Q^2 \tag{16}$$

$X \in \mathbb{R}^{p\times n}$: a given data matrix, $Q \in \mathbb{R}^{pn\times pn}$: a weighted matrix,
$\|X - R\|_Q^2 = \mathrm{vec}\{X - R\}^T Q \mathrm{vec}\{X - R\}$., rewrite (16) as

$$\min_{\substack{N \in \mathbb{R}^{n\times(n-r)} \\ N^T N = 1}} \min_{\substack{R \in \mathbb{R}^{p\times n} \\ RN = 0}} \|X - R\|_Q^2$$

The inner minimization has a closed form solution, call it $f(N)$:

$$f(N) = \mathrm{vec}\{X\}^T (N \otimes I_p) \left[(N \otimes I_p)^T Q^{-1} (N \otimes I_p) \right]^{-1} (N \otimes I_p)^T \mathrm{vec}\{X\}$$

Riemannian BFGS: past and future

Previous work on BFGS on manifolds

- Gabay [Gab82] discussed a version using parallel translation
- Brace and Manton restrict themselves to a version on the Grassmann manifold and the problem of weighted low-rank approximations [BM06].
- Savas and Lim apply a version to the more complicated problem of best multilinear approximations with tensors on a product of Grassmann manifolds [SL10].

Our goals

- Make the algorithm faster.
- Understand its convergence better.

Riemannian BFGS: a glimpse of the 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: $\eta_k = -\mathcal{B}_k^{-1} \operatorname{grad} f(\mathbf{x}_k)$.

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 $y_k = \operatorname{grad} f(\mathbf{x}_{k+1}) - \mathcal{T}_{\alpha\eta_k} \operatorname{grad} f(\mathbf{x}_k)$

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**

Vector transport

Manifold algorithms

- Conjugate gradients
- Secant methods
- BFGS

where parallel translation is used to combine two or more tangent vectors from distinct tangent spaces.

Vector transport

We define a **vector transport** on a manifold $\mathcal{M}$ to be a smooth mapping

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

satisfying three properties for all $x \in \mathcal{M}$.

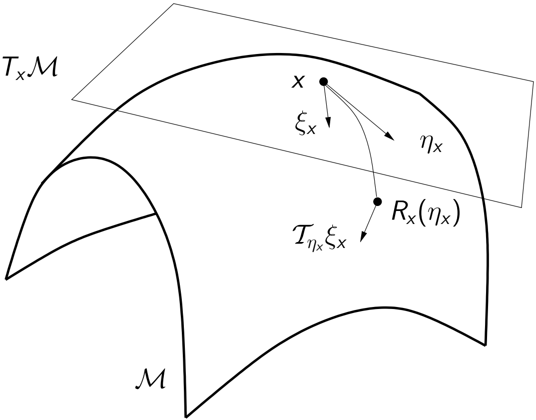


Figure: Vector transport.

Vector Transport

- (Associated retraction) There exists a retraction R, called the *retraction associated with $\mathcal{T}$*, such that the following diagram commutes

$$\begin{tikzcd} (\eta_x, \xi_x) \arrow[r, "\mathcal{T}"] \arrow[d] & \mathcal{T}_{\eta_x}(\xi_x) \arrow[d, "\pi"] \\ \eta_x \arrow[r, "R"'] & \pi\left(\mathcal{T}_{\eta_x}(\xi_x)\right) \end{tikzcd}$$

  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 $\xi_x \in T_x\mathcal{M}$;
- (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 by differentiated retraction

Let M be a manifold endowed with retraction R, a particular vector transport is given by

$$\mathcal{T}_{\eta_x}\xi_x := \mathrm{D}R_x(\eta_x)[\xi_x]; \text{ i.e.,}$$
$$\mathcal{T}_{\eta_x}\xi_x := \frac{\mathrm{d}}{\mathrm{dt}} R_x(\eta_x + t\xi_x)\Big|_{t=0};$$

Vector transport by projection [AMS08, §8.1.2] (submanifolds only)

If M is an embedded submanifold of a Euclidean space ε and M is endowed with a retraction R, then

$$\mathcal{T}_{\eta_x}\xi_x := \mathrm{P}_{R_x(\eta_x)}\xi_x,$$

where P_x denotes the orthgonal projector onto T_xM, is a vector transport.

Vector transport on quotient manifold

$\mathcal{M} = \overline{\mathcal{M}} / \sim$: a quotient manifold, where $\mathcal{M}$ is an open subset of a Euclidean space ε.

$$\overline{(\mathcal{T}_{\eta_x} \xi_x)}_{\overline{x}+\overline{\eta}_{\overline{x}}} := \mathrm{P}^{h}_{\overline{x}+\overline{\eta}_{\overline{x}}} \overline{\xi}_{\overline{x}},$$

where $\mathrm{P}^{h}_{\overline{x}} Z : T_{\overline{x}} \overline{\mathcal{M}} \to \mathcal{H}_{\overline{x}}$ denotes the projection parallel to the vertical space $\mathcal{V}_{\overline{x}}$ onto the horizontal space $\mathcal{H}_{\overline{x}}$ at $\overline{x}$.

Algorithm 2 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: $\eta_k = -\mathcal{B}_k^{-1} \mathrm{grad}\, f(\mathbf{x}_k)$.

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 $y_k = \mathrm{grad}\, f(\mathbf{x}_{k+1}) - \mathcal{T}_{\alpha \eta_k} \mathrm{grad}\, f(\mathbf{x}_k)$

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**

Sherman-Morrison formula

Let A is an invertible matrix. The for all vectors u, v such that $1 + v^T A^{-1} u \neq 0$, one has

$$(A + uv^T)^{-1} = A^{-1} + \frac{A^{-1} uv^T A^{-1}}{1 + v^T A^{-1} u}.$$

Another version of the RBFGS algorithm

Works with the inverse Hessian $\mathcal{H}_k = {\mathcal{B}_k}^{-1}$ approximation rather than the Hessian approximation B_k. In this case the step 4 in algorithm 2 will be replaced by:
$\mathcal{H}_{k+1} = \tilde{\mathcal{H}}_k p - \frac{g(y_k, \tilde{\mathcal{H}}_k p)}{g(y_k, s_k)} s_k - \frac{g(s_k, p_k)}{g(y_k, s_k)} \tilde{\mathcal{H}}_k y_k + \frac{g(s_k, p) g(y_k, \tilde{\mathcal{H}}_k y_k)}{g(y_k, s_k)^2} s_k + \frac{g(s_k, s_k)}{g(y_k, s_k)} p$
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.

Implementation of RBFGS in submanifolds of $\mathbb{R}^n$

Let $x \in M, \xi_x, \eta_x \in T_x M$, define the inclusions:

$$\mathrm{i}: M \to \mathbb{R}^n; x \mapsto \mathrm{i}(x)$$
$$\mathrm{i}_x : T_x M \to \mathbb{R}^n; \xi_x \mapsto \mathrm{i}_x(\xi_x)$$

use the matrix B_k to represent the linear operator $\mathcal{B}_k : T_{x_k} M \to T_{x_k} M$.

$$B_k \leftarrow \mathcal{B}_k$$

we have

$$\mathrm{i}_x(\mathcal{B}_k \xi_x) = B_k(\mathrm{i}_x(\xi_x))$$

$$g_x(\xi_x, \eta_x) = \langle \mathrm{i}_x(\xi_x), \mathrm{i}_x(\eta_x) \rangle$$

Compute $\eta_k = -\mathcal{B}_k^{-1} \mathrm{grad}\, f(x_k)$ for Submanifolds.

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

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

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

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

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

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

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

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

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

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 2 converge to the minimizer of f.

Superlinear convergence of quasi-Newton: 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 $T_x M$ 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 $\lim\limits_{k\to\infty} x_k = x^*$.

Then $\{x_k\}$ converges superlinearly to x^* and $F(x^*) = 0$ if and only if

$$\lim_{k\to\infty} \frac{\|[\mathcal{B}_k - \mathcal{T}_{\xi_k} \mathbb{D}F(x^*) \mathcal{T}_{\xi_k}^{-1}]\eta_k\|}{\|\eta_k\|} = 0 \tag{17}$$

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.

On the Unit Sphere $\mathbb{R}^n$

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:

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

Retraction:

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

Transport on the Unit Sphere $\mathbb{R}^n$

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:

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

On the Unit Sphere

Let $T_{\eta_k}^{(n)}$ be the representation of $\mathcal{T}_{\eta_k}$

$$T_{\eta_k}^{(n)} = \left(I - \frac{(x+\eta)(x+\eta)^T}{\|x+\eta\|^2} \right)$$

Approach 1: Realize $\mathcal{B}_k$ by an n-by-n matrix

1) $\tilde{B}_k^{(n)} = T_{\eta_k}^{(n)} B_k^{(n)} ((T_{\eta_k})^{(n)})^{-1}$;

2) $B_{k+1}^{(n)} = \tilde{B}_k^n - \frac{\tilde{B}_k^{(n)} s_k s_k^T \tilde{B}_k^n}{\langle s_k, \tilde{B}_k^{(n)} s_k \rangle} + \frac{y_k y_k^T}{\langle y_k, s_k \rangle}$,

Approach 2: Use bases

1) Calculate $\tilde{B}_k^d$ though $B_k^{(d)}$:

$$\tilde{B}_k^d = \underline{E}_{k+1}^+ \tilde{B}_k^{(n)} \underline{E}_{k+1};$$

$$= \underline{E}_{k+1}^+ T_{\eta_k}^{(n)} B_k^{(n)} (T_{\eta_k}^{(n)})^{-1} \underline{E}_{k+1}$$

$$= \underline{E}_{k+1}^+ T_{\eta_k}^{(n)} \underline{E}_k B_k^{(d)} \underline{E}_k^+ (T_{\eta_k}^{(n)})^{-1} \underline{E}_{k+1}$$

2) $B_{k+1}^{(d)} = \tilde{B}_k^{(d)} - \frac{\tilde{B}_k^{(d)} s_k^{(d)} (s_k^{(d)})^T \tilde{B}_k^{(d)}}{\langle s_k^{(d)}, \tilde{B}_k^{(d)} s_k^{(d)} \rangle} + \frac{y_k^{(d)} (y_k^{(d)})^T}{\langle y_k^{(d)}, s_k^{(d)} \rangle}$,

Rayleigh quotient minimization on S^{n-1}

Cost function on S^{n-1}

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

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

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

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

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

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

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

Methods Numerical Experiment

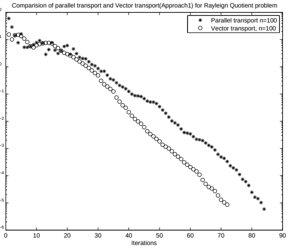
1. Vector transport (approach 1), update $H = B^{-1}, \eta = -H \operatorname{grad} f(x)$
2. Vector transport (approach 2), update $H = B^{-1}, \eta = -H \operatorname{grad} f(x)$
3. Parallel transport, update $H = B^{-1}, \eta = -H \operatorname{grad} f(x)$
4. Vector transport (approach 1), Update L, solve $L_+ L_+^T \eta = -\operatorname{grad} f(x)$ (QR factorization)
5. Riemannian Line Search Newton-CG
6. Riemannian Trust Region with Truncated-CG

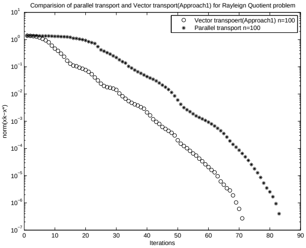
Numerical Result for Rayleigh Quotient on S^{n-1}

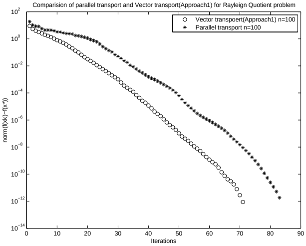
- Problem sizes $n = 100$ and $n = 300$ with many different initial points.
- All versions of RBFGS converge superlinearly to local minimizer.
- Updating L and B^{-1} combined with Vector transport display similar convergence rates.
- Vector transport Approach 1 and Approach 2 display the same convergence 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 faster than Parallel transport. On S^{n-1}, they have similar computational cost.
- Newton−CG version converges slightly more quickly than the Vector transport versions.

Rayleigh quotient on S^{n-1}

Vector transport has better convergence rate than Parallel transport







Rayleigh quotient on S^{n-1}

Table: Comparison of Vector transport vs. Parallel translation for Rayleigh quotient Problem

Case	Vector trans. (n=100)	Vector trans. (n=300)	Parallel trans. (n=100)	Parallel trans. (n=300)
Time	0.22	4.06	0.46	5.49
Iteration	71	97	84	95

Table: Vector transport approach1 vs. approach2 for Rayleigh quotient problem

Case	approach 1 (n=100)	approach 1 (n=300)	approach 2 (n=100)	approach 2 (n=300)
Time	0.22	4.06	2.2	33.6
Iteration	71	97	71	97

Other 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

On the Manifold $S^{n-1} \times \cdots \times S^{n-1}$

$$\begin{aligned} X &= [x_1, x_2, \cdots, x_N] \in S^{n-1} \times \cdots \times S^{n-1} \\ x_i^T x_i &= 1, \text{ for } i = 1 \text{ to } N \end{aligned}$$

Riemannian metric:

$$\ll Z, W \gg_X = \langle z_1, w_1 \rangle_{x_1} + \cdots + \langle z_N, w_N \rangle_{x_N} = \mathrm{tr}(Z^T W), Z, W \in T_X \mathcal{M}$$

Tangent space at x:

$$T_x \mathcal{M} = \left\{ Z = [z_1, \cdots, z_N] \in \mathbb{R}^{n \times N} \middle| x_1^T z_1 = x_2^T z_2 = \cdots = x_N^T z_N = 0 \right\}$$

Orthogonal projection to tangent space:

$\mathrm{P}_X W = [(I - x_1 x_1^T) w_1, \cdots, (I - x_N x_N^T) w_N]$ projects $W \in \mathbb{R}^{n \times N}$ to $T_x \mathcal{M}$

Retraction:

$$R_X(Z) = \left[\frac{x_1 + z_1}{\|x_1 + z_1\|}, \cdots, \frac{x_N + z_N}{\|x_N + z_N\|} \right]$$

Transport on $S^{n-1} \times \cdots \times S^{n-1}$

Parallel and vector transport (and their inverses) of

$$\xi_X = [\xi_1, \xi_2, \cdots, \xi_N] \in T_x\mathcal{M}$$

defined by directions

$$\eta_X = [\eta_1, \eta_2, \cdots, \eta_N] \in T_x\mathcal{M}$$

simply apply the corresponding transport mechanisms from S^{n-1} componentwise.

Thomson Problem on $S^{n-1} \times \cdots S^{n-1}$

$$X = [x_1, x_2, \cdots, x_N] \in \mathcal{M}, x_i^T x_i = 1, \text{ for } i = 1 \text{ to } N$$

$$f : [x_1, x_2, \cdots, x_N] \longmapsto \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{1}{\|x_i - x_j\|^2}$$

$$\operatorname{grad} f(X) = \left[(I - x_1 x_1^T) \sum_{j=2}^{N} \frac{1}{(1 - x_1^T x_j)^2} x_j, \cdots, (I - x_N x_N^T) \sum_{j=1}^{N-1} \frac{1}{(1 - x_N^T x_j)^2} x_j \right]$$

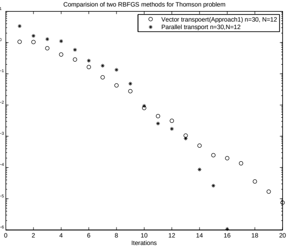
Methods Numerical Experiment

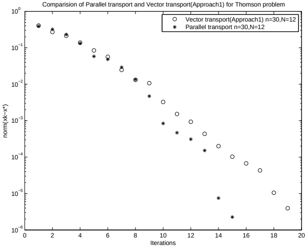
1. Vector transport (approach 1), update $H = B^{-1}, \eta = -H \operatorname{grad} f(x)$
2. Vector transport (approach 2), update $H = B^{-1}, \eta = -H \operatorname{grad} f(x)$
3. Parallel transport (approach 1), update $H = B^{-1}, \eta = -H \operatorname{grad} f(x)$
4. Vector transport (approach 1), Update L, solve $L_+ L_+^T \eta = -\operatorname{grad} f(x)$ (QR factorization)
5. Riemannian Trust Region with Truncated-CG

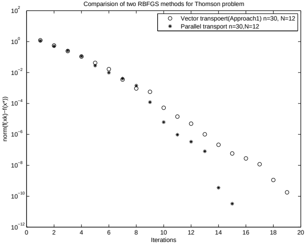
Numerical Result for Thomson Problem

- Problem sizes $(n, N) = (30, 12)$ and $(n, N) = (50, 20)$ 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.
- Parallel transport converge slightly faster than Vector transport versions .

Update of B^{-1}, Parallel and Vector Transport







Update of B^{-1}, Parallel and Vector Transport

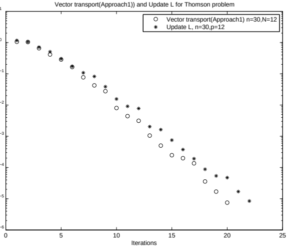
Table: Vector transport (approach 1) vs. Parallel transport for Thomson problem

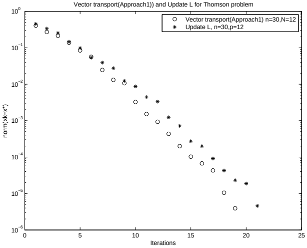
Case	Vector trans. (n=30, N=12)	Vector trans. (n=50, N=20)	Parallel trans. (n=30, N=12)	Parallel trans. (n=50, N=20)
Time	3.9	60	3.4	47.6
Iteration	20	24	16	19

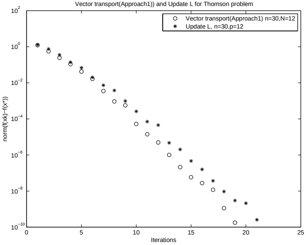
Table: Vector transport (approach 1) vs. Parallel transport (approach 1) for Thomson problem

Case	approach 1 (n=30, N=12)	approach 1 (n=50, N=20)	approach 2 (n=30, N=12)	approach 2 (n=50, N=20)
Time	3.9	60	13	252
Iteration	20	24	20	24

Update L and Update of B^{-1} for Thomson Problem







Update of B^{-1} and Riemannian Trust Region Method

- Total inner iteration count of RTR is larger than iteration count of R BFGS
- RTR inner iteration and RBFGS iteration similar complexity

Update of B^{-1} and Riemannian Trust Region Method

Table: RBFGS (Vector transport, approach 1) vs. RTR for Rayleigh Quotient problem

Case	RBFGS (n=30,N=12)	RBFGS (n=50,N=20)	RTR (n=30,N=12)	RTR (n=50,N=20
Iteration	20	24	30	36

Compact Stiefel Manifold $St(p, n)$

View $St(p, n)$ as a Riemannian submanifold of the Euclidean space $\mathbb{R}^{n \times p}$
Riemannian metric: $g(\xi, \eta) = \mathrm{tr}(\xi^T \eta)$
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 :

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

Retraction:

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

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

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

The parallel transport of H along the geodesic from Y in direction H:

$$P_{\gamma_H}^{t \leftarrow 0} H = HM(t) - YR^T N(t)$$

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 $St(p, n)$ Approach 1

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

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

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

Vector Transport on $St(p, n)$ Approach2

- Find d independent tangent vectors $E_{k,1}, E_{k,2}, \cdots E_{k,d} \in T_{X_k}$;
- Vector transport each $E_{ki}, i = 1, 2, \cdots d$ to $T_{X_{k+1}}$,
  $\underline{E}_{k+1} = [T_{\eta_k}^{(np)} E_{k,1} \quad T_{\eta_k}^{(np)} E_{k,2} \quad \cdots \quad T_{\eta_k}^{(np)} E_{k,d}]$
- Calculate $\tilde{B}_k^{(np)} = T_{\eta_k}^{(np)} B_k^{(np)} (T_{\eta_k}^{(np)})^{-1}$:

$$\tilde{B}_k^{(np)} \underline{E}_{k+1} = \left[T_{\eta_k}^{(np)}(B_k^{(np)} E_{k,1}) \quad T_{\eta_k}^{(np)}(B_k^{(np)} E_{k,2}) \quad \cdots \quad T_{\eta_k}^{(np)}(B_k^{(np)} E_{k,d}) \right],$$

$$\tilde{B}_k^{(np)} = \left[T_{\eta_k}^{(np)}(B_k^{(np)} E_{k,1}) \quad T_{\eta_k}^{(np)}(B_k^{(np)} E_{k,2}) \quad \cdots \quad T_{\eta_k}^{(np)}(B_k^{(np)} E_{k,d}) \right] \underline{E}_{k+1}^{+}.$$

- Compute the RBFGS update

$$B_{k+1}^{(np)} = \tilde{B}_k^{(np)} - \frac{\tilde{B}_k^{(np)} s_k^{(np)} s_k^{(np)\,T} \tilde{B}_k^{(np)}}{\langle s_k^{(np)}, \tilde{B}_k^{(np)} s_k^{(np)} \rangle} + \frac{y_k^{(np)} y_k^{(np)\,T}}{\langle y_k^{(np)}, s_k^{(np)} \rangle}, \text{ and set}$$

$$\eta_{k+1} = \mathrm{unvec}\{(-B_{k+1}^{(np)})^{-1} \mathrm{vec}\{\mathrm{grad}\, f(X_k)\}\}.$$

A Procrustes Problem on St(p, n)

Cost function on St(p, n)

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

where A: $n \times n$ matix, B : $p \times p$ matix, $X^T X = I_p$
Cost function embedded in $\mathbb{R}^{n \times p}$

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

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

$$\mathrm{D}\bar{f}(X)[Z] = \frac{\mathrm{tr}(Z^T Q)}{\bar{f}(X)}, \text{ where } Q = A^T AX - A^T XB - B^T AX + B^T XB,$$

Projection onto $T_x \mathbb{R}^n$:

$$\mathrm{P}_X Z \;\; = (I - XX^T)Z + X\mathrm{skew}(X^T Z)$$

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

Methods Numerical Experiment

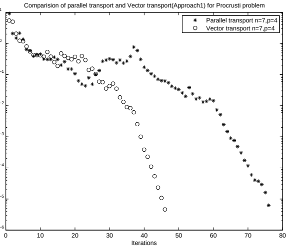
1. Vector transport (approach 1), update $H = B^{-1}, \eta = -H\mathrm{grad}\, f(x)$
2. Vector transport (approach 2), update $H = B^{-1}, \eta = -H\mathrm{grad}\, f(x)$
3. Parallel transport, update $H = B^{-1}, \eta = -H\mathrm{grad}\, f(x)$
4. Vector transport (approach 1), Update L, solve $L_+ L_+^T \eta = -\mathrm{grad}\, f(x)$ (QR factorization)
5. Riemannian Line Search Newton-CG
6. Riemannian Trust Region with Truncated-CG

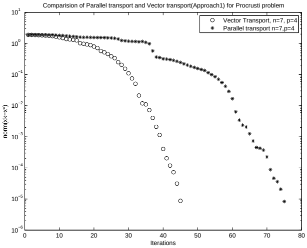
Numerical Result for Procrustes on $\mathrm{St}(p, n)$

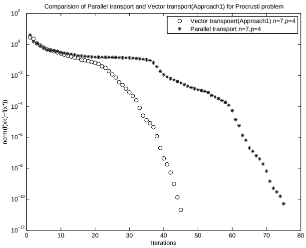
- Problem sizes $(n, p) = (7, 4)$ and $(n, p) = (12, 7)$ with many different initial points.
- All versions of RBFGS converge superlinearly to local minimizer.
- Updating L and B^{-1} combined with Vector transport display B^{-1} is slightly faster converging.
- Vector transport Approach 1 and Approach 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.
- Newton−CG version has convergence problems compared to the Vector transport RBFGS versions.

Procrustes Problem on $St(p, n)$

Vector transport has better convergence rate than Parallel transport







Procrustes Problem on St(p, n)

Table: B^{-1} update w/ Vector transport (approach 1) vs. Parallel transport

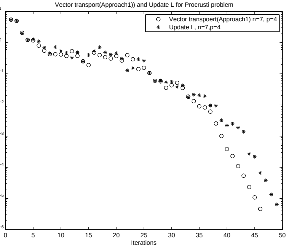
Case	Vector trans. (n=7, p=4)	Vector trans. (n=12, p=7)	Parallel trans. (n=7, p=4)	Parallel trans. (n=12, p=7)
Time	4.1	45	81	781
Iteration	46	82	67	174

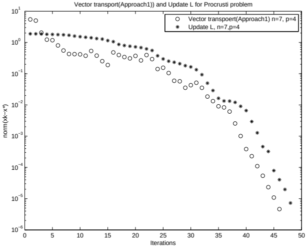
Table: Vector transport approach1 vs. approach2 for Procrustes problem

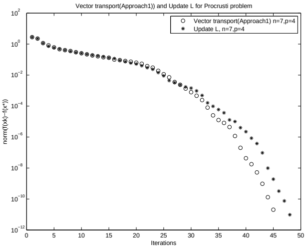
Case	approach 1 (n=7, p=4)	approach 1 (n=12, p=7)	approach 2 (n=7, p=4)	approach 2 (n=12, p=7)
Time	4.1	46	7.5	95
Iteration	46	82	48	86

Update of L and Update of B^{-1}

- Both $O(n^2)$ operations per step and use Vector transport with Approach 1.
- Similar convergence behavior

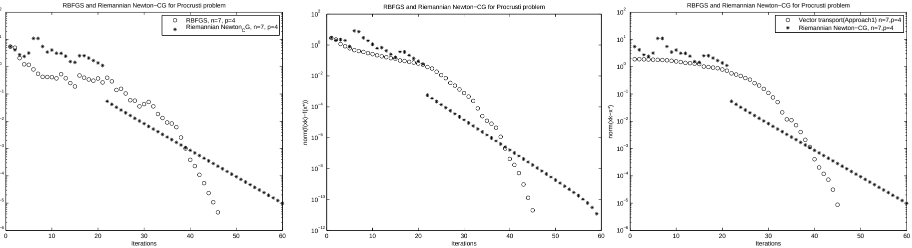






Update of B^{-1} and Riemannian Line Search Newton−CG

- The Convergence of RBFGS is superlinear, while Newton−CG is linear since no forcing function used in CG convergence check.



Update of B^{-1} and Riemannian Trust Region Method

- Total inner iteration count of RTR is larger than iteration count of RBFGS
- RTR inner iteration and RBFGS iteration similar complexity

Comparision of RBFGS with Riemannian Trust Region Method

Table: RBFGS (Vector transport, approach 1) vs. RTR for Procrustes problem

Case	RBFGS (n=7, p=4)	RBFGS (n=12, p=7)	RTR (n=7, p=4)	RTR (n=12, p=7)
Iteration	47	86	115	357

A (questionable) historical overview

	In $\mathbb{R}^n$	On Riemannian manifolds	
		using classical objects	using novel objects
Steepest descent	1966 (Armijo backtracking)	1972 (Luenberger)	1986–2008 ?
Newton	1740 (Simpson)	1993 (Smith)	2002 (Adler et al.)
Conjugate Grad	1964 (Fletcher–Reeves)	1993 (Smith)	2008 (PAA, Mahony, Sepulchre) ?
Trust regions	1985 (name created by Celis, Dennis, Tapia)	2007 (PAA, Baker, Gallivan)	2007 (PAA, Baker, Gallivan)
BFGS	1970 (B-F-G-S)	1982 (Gabay)	Now!

Conclusion: A Three-Step Approach

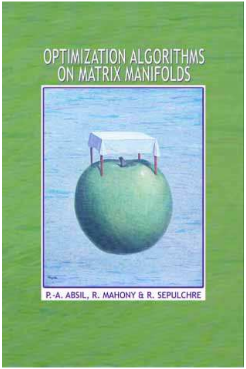
- Formulation of the computational problem as a geometric optimization problem.
- Generalization of optimization algorithms on abstract manifolds.
- Exploit flexibility and additional structure to build numerically efficient algorithms.

A few pointers

- Optimization on manifolds: Luenberger [Lue73], Gabay [Gab82], Smith [Smi93, Smi94], Udriște [Udr94], Manton [Man02], Mahony and Manton [MM02], PAA *et al.* [ABG04, ABG07]...
- Trust-region methods: Powell [Pow70], Moré and Sorensen [MS83], Moré [Mor83], Conn *et al.* [CGT00].
- Truncated CG: Steihaug [Ste83], Toint [Toi81], Conn *et al.* [CGT00]...
- Retractions: Shub [Shu86], Adler *et al.* [ADM+02]...

THE END

Optimization Algorithms on Matrix Manifolds
P.-A. Absil, R. Mahony, R. Sepulchre
Princeton University Press, January 2008



1. Introduction
2. Motivation and applications
3. Matrix manifolds: first-order geometry
4. Line-search algorithms
5. Matrix manifolds: second-order geometry
6. Newton's method
7. Trust-region methods
8. A constellation of superlinear algorithms

P.-A. Absil, C. G. Baker, and K. A. Gallivan, *Trust-region methods on Riemannian manifolds with applications in numerical linear algebra*, Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS2004), Leuven, Belgium, 5–9 July 2004, 2004.

———, *Trust-region methods on Riemannian manifolds*, Found. Comput. Math. **7** (2007), no. 3, 303–330.

Roy L. Adler, Jean-Pierre Dedieu, Joseph Y. Margulies, Marco Martens, and Mike Shub, *Newton's method on Riemannian manifolds and a geometric model for the human spine*, IMA J. Numer. Anal. **22** (2002), no. 3, 359–390.

P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, Princeton, NJ, 2008.

Ian Brace and Jonathan H. Manton, *An improved BFGS-on-manifold algorithm for computing weighted low rank approximations*, Proceedings of the 17h International Symposium on Mathematical Theory of Networks and Systems, 2006, pp. 1735–1738.

Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint, *Trust-region methods*, MPS/SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. MR MR1774899 (2003e:90002)

D. Gabay, *Minimizing a differentiable function over a differential manifold*, J. Optim. Theory Appl. **37** (1982), no. 2, 177–219. MR MR663521 (84h:49071)

Gene H. Golub and Qiang Ye, *An inverse free preconditioned Krylov subspace method for symmetric generalized eigenvalue problems*, SIAM J. Sci. Comput. **24** (2002), no. 1, 312–334.

- Magnus R. Hestenes and William Karush, *A method of gradients for the calculation of the characteristic roots and vectors of a real symmetric matrix*, J. Research Nat. Bur. Standards **47** (1951), 45–61.
- Uwe Helmke and John B. Moore, *Optimization and dynamical systems*, Communications and Control Engineering Series, Springer-Verlag London Ltd., London, 1994, With a foreword by R. Brockett. MR MR1299725 (95j:49001)
- David G. Luenberger, *Introduction to linear and nonlinear programming*, Addison-Wesley, Reading, MA, 1973.
- Jonathan H. Manton, *Optimization algorithms exploiting unitary constraints*, IEEE Trans. Signal Process. **50** (2002), no. 3, 635–650. MR MR1895067 (2003i:90078)

Robert Mahony and Jonathan H. Manton, *The geometry of the Newton method on non-compact Lie groups*, J. Global Optim. **23** (2002), no. 3-4, 309–327, Nonconvex optimization in control. MR MR1923049 (2003g:90114)

J. J. Moré, *Recent developments in algorithms and software for trust region methods*, Mathematical programming: the state of the art (Bonn, 1982) (Berlin), Springer, 1983, pp. 258–287.

Jorge J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput. **4** (1983), no. 3, 553–572. MR MR723110 (86b:65063)

M. Mongeau and M. Torki, *Computing eigenelements of real symmetric matrices via optimization*, Comput. Optim. Appl. **29** (2004), no. 3, 263–287. MR MR2101850 (2005h:65061)

M. J. D. Powell, *A new algorithm for unconstrained optimization*, Nonlinear Programming (Proc. Sympos., Univ. of Wisconsin, Madison, Wis., 1970), Academic Press, New York, 1970, pp. 31–65.

Michael Shub, *Some remarks on dynamical systems and numerical analysis*, Proc. VII ELAM. (L. Lara-Carrero and J. Lewowicz, eds.), Equinoccio, U. Simón Bolívar, Caracas, 1986, pp. 69–92.

B. Savas and L.-H. Lim, *Quasi-newton methods on grassmannians and multilinear approximations of tensors*, SIAM J. Sci. Comput. **32** (2010), no. 6, 3352–3393.

Steven Thomas Smith, *Geometric optimization methods for adaptive filtering*, Ph.D. thesis, Division of Applied Sciences, Harvard University, Cambridge, MA, May 1993.

Steven T. Smith, *Optimization techniques on Riemannian manifolds*, Hamiltonian and gradient flows, algorithms and control (Anthony Bloch, ed.), Fields Inst. Commun., vol. 3, Amer. Math. Soc., Providence, RI, 1994, pp. 113–136. MR MR1297990 (95g:58062)

Trond Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal. **20** (1983), no. 3, 626–637. MR MR701102 (84g:49047)

Ph. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, Sparse Matrices and Their Uses (I. S. Duff, ed.), Academic Press, London, 1981, pp. 57–88.

Constantin Udrişte, *Convex functions and optimization methods on Riemannian manifolds*, Mathematics and its Applications, vol. 297, Kluwer Academic Publishers Group, Dordrecht, 1994. MR MR1326607 (97a:49038)