# Riemannian Optimization: Proximal Gradient Methods

Speaker: Wen Huang

Xiamen University

May 28, 2021

# Outline

Research field: optimization on manifolds with its applications

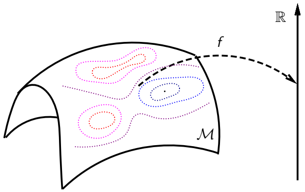
- Problem statement
- A non-exhaustive literature review
- Our work
- A Riemannian proximal gradient method
- Ongoing related research
- Summary

# Problem Statement

**Problem:** Given $f(x) : \mathcal{M} \rightarrow \mathbb{R}$,
solve

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

where $\mathcal{M}$ is a Riemannian manifold.

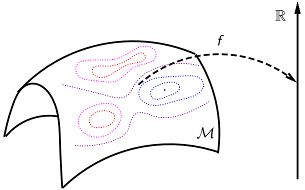


# Problem Statement

**Problem:** Given $f(x) : \mathcal{M} \to \mathbb{R}$, solve

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

where $\mathcal{M}$ is a Riemannian manifold.




---

**Manifolds:**

- Stiefel: $\mathrm{St}(p, n) = \{X \in \mathbb{R}^{n \times p} : X^T X = I_p\}$;
- Grassmann: the set of $p$ dimensional linear spaces in $\mathbb{R}^n$;
- Fixed rank: $\mathbb{R}_r^{m \times n} = \{X \in \mathbb{R}^{m \times n} : \mathrm{rank}(X) = r\}$ or tensor;
- Symmetric positive definite: $\mathcal{S}_{++}^n = \{X \in \mathbb{R}^{n \times n} : X \succ 0\}$;

# Review (non-exhaustive)

- Smooth unconstrained problems
  - Steepest descent: Smith 1994; Helmke-Moore 1994; Iannazzo-Porcelli 2019;
  - Conjugate gradient: Smith 1994; Gallivan-Absil 2010; Ring-Wirth 2012; Sato-Iwai 2015;
  - Quasi-Newton: Ring-Wirth 2012; Huang-Absil-Gallivan 2018;
  - Trust region Newton: Absil-Baker-Gallivan 2007;

- Nonsmooth unconstrained problems
  - Proximal point method: Ferreira-Oliveira 2002;
  - Optimality conditions: Yang-Zhang-Song 2014;
  - Gradient sampling: Huang 2013; Hosseini and Uschmajew 2017;
  - $\epsilon$-subgradient-based methods: Grohs and Hosseini 2015;

- Constrained problems:
  - Augmented Lagrangian methods: Boumal-Liu 2019;

# Review (non-exhaustive)

- Smooth unconstrained problems:
  - Stiefel manifold: Wen-Yin 2012; Jiang-Dai 2014; Xiao-Liu-Yuan 2020; Dai-Wang-Zhou 2020
  - Symmetric positive definite manifold: Bini-Iannazzo 2013; Zhang 2017; Yuan-Huang-Absil-Gallivan 2020;
  - Fixed rank manifold: Wen-Yin-Zhang 2012; Mishra 2014; Boumal-Absil 2014;
- Nonsmooth unconstrained problems:
  - Stiefel Manifold: Huang-Wei 2019; Chen-Ma-So-Zhang 2020; Xiao-Liu-Yuan 2020;
  - Fixed rank manifold: Cambier-Absil 2016;
- Constrained problems:
  - Stiefel + non-negativity: Jiang-Meng-Wen-Chen 2019;
  - Symmetric positive definite + zeros: Phan-Menickelly 2020;

# Review (non-exhaustive)

Riemannian optimization libraries for general problems:

- Boumal, Mishra, Absil, Sepulchre(2014)
  Manopt (Matlab library)
- Townsend, Koep, Weichwald (2016)
  Pymanopt (Python version of manopt)
- Bergmann (2019)
  manoptjl (Julia, nonsmooth methods)
- Huang, Absil, Gallivan, Hand (2018)
  ROPTLIB (C++ library, interfaces to Matlab and Julia)
- Martin, Raim, Huang, Adragni (2018)
  ManifoldOptim (R wrapper of ROPTLIB)
- Meghawanshi, Jawanpuria, Kunchukuttan, Kasai, Mishra (2018)
  McTorch (Python, GPU acceleration)

# Our Work

- Smooth unconstrained problems
  - Broyden family including BFGS method [HGA15, HAG16, HAG18]
  - Trust-region symmetric rank-one method [HAG15]
  - Their limited-memory versions [HG21]
- Nonsmooth unconstrained problems
  - $\epsilon$-subgradient with quasi-Newton method [HHY18]
  - Proximal gradient methods [HW21]
- Applications:
  - Elastic shape analysis [HGSA15]
  - Blind deconvolution [HH18]
  - Phase retrieval [HGZ16]
  - Sparse principal component analysis [HW19]
- Library: ROPTLIB [HAGH18]

# Recent work: Riemannian Proximal Gradient Methods

$$\min_{x \in \mathcal{M}} F(x) = f(x) + g(x),$$

- $\mathcal{M}$ is a Riemannian manifold;
- $f$ is continuously differentiable and may be nonconvex; and
- $g$ is continuous, but may be not differentiable.

**Optimization with Structure:** $\mathcal{M} = \mathbb{R}^n$

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + g(x), \qquad (1)$$

---

1

**Optimization with Structure:** $\mathcal{M} = \mathbb{R}^n$

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + g(x), \tag{1}$$

---

A proximal gradient method[1]:

initial iterate: $x_0$,

$$\begin{cases} d_k = \arg\min_{p \in \mathbb{R}^n} \langle \nabla f(x_k), p \rangle + \frac{L}{2}\|p\|_F^2 + g(x_k + p), & \text{(Proximal mapping)} \\ x_{k+1} = x_k + d_k. & \text{(Update iterates)} \end{cases}$$

[1]The update rule: $x_{k+1} = \arg\min_x \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|^2 + g(x)$.

**Optimization with Structure:** $\mathcal{M} = \mathbb{R}^n$

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + g(x), \quad (1)$$

---

A proximal gradient method[1]:

initial iterate: $x_0$,

$$\begin{cases} d_k = \arg\min_{p \in \mathbb{R}^n} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p), & \text{(Proximal mapping)} \\ x_{k+1} = x_k + d_k. & \text{(Update iterates)} \end{cases}$$

- $g = 0$: reduce to steepest descent method;

[1] The update rule: $x_{k+1} = \arg\min_x \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|^2 + g(x)$.

**Optimization with Structure:** $\mathcal{M} = \mathbb{R}^n$

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + g(x), \qquad (1)$$

---

A proximal gradient method[1]:

initial iterate: $x_0$,

$$\begin{cases} d_k = \arg\min_{p \in \mathbb{R}^n} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p), & \text{(Proximal mapping)} \\ x_{k+1} = x_k + d_k. & \text{(Update iterates)} \end{cases}$$

- $g = 0$: reduce to steepest descent method;
- $L$: greater than the Lipschitz constant of $\nabla f$;

[1] The update rule: $x_{k+1} = \arg\min_x \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|^2 + g(x)$.

**Optimization with Structure:** $\mathcal{M} = \mathbb{R}^n$

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + g(x), \tag{1}$$

---

A proximal gradient method[1]:

$$\text{initial iterate:} x_0,$$

$$\begin{cases} d_k = \arg\min_{p \in \mathbb{R}^n} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p), & \text{(Proximal mapping)} \\ x_{k+1} = x_k + d_k. & \text{(Update iterates)} \end{cases}$$

- $g = 0$: reduce to steepest descent method;
- $L$: greater than the Lipschitz constant of $\nabla f$;
- Proximal mapping: easy to compute;

[1] The update rule: $x_{k+1} = \arg\min_x \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|^2 + g(x)$.

**Optimization with Structure:** $\mathcal{M} = \mathbb{R}^n$

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + g(x), \qquad (1)$$

---

A proximal gradient method[1]:

initial iterate: $x_0$,

$$\begin{cases} d_k = \arg\min_{p \in \mathbb{R}^n} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p), & \text{(Proximal mapping)} \\ x_{k+1} = x_k + d_k. & \text{(Update iterates)} \end{cases}$$

- $g = 0$: reduce to steepest descent method;
- $L$: greater than the Lipschitz constant of $\nabla f$;
- Proximal mapping: easy to compute;
- Any limit point is a critical point;

[1] The update rule: $x_{k+1} = \arg\min_x \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|^2 + g(x)$.

**Optimization with Structure:** $\mathcal{M} = \mathbb{R}^n$

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + g(x), \quad (1)$$

---

A proximal gradient method[1]:

initial iterate: $x_0$,

$$\begin{cases} d_k = \arg\min_{p \in \mathbb{R}^n} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p), & \text{(Proximal mapping)} \\ x_{k+1} = x_k + d_k. & \text{(Update iterates)} \end{cases}$$

- $g = 0$: reduce to steepest descent method;
- $L$: greater than the Lipschitz constant of $\nabla f$;
- Proximal mapping: easy to compute;
- Any limit point is a critical point;
- $O(1/k)$ sublinear convergence rate for convex $f$ and $g$;

[1] The update rule: $x_{k+1} = \arg\min_x \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|^2 + g(x)$.

**Optimization with Structure:** $\mathcal{M} = \mathbb{R}^n$

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + g(x), \qquad (1)$$

---

A proximal gradient method[1]:

initial iterate: $x_0$,

$$\begin{cases} d_k = \arg\min_{p \in \mathbb{R}^n} \langle \nabla f(x_k), p \rangle + \frac{L}{2}\|p\|_F^2 + g(x_k + p), & \text{(Proximal mapping)} \\ x_{k+1} = x_k + d_k. & \text{(Update iterates)} \end{cases}$$

- $g = 0$: reduce to steepest descent method;
- $L$: greater than the Lipschitz constant of $\nabla f$;
- Proximal mapping: easy to compute;
- Any limit point is a critical point;
- $O(1/k)$ sublinear convergence rate for convex $f$ and $g$;
- Local convergence rate by KL property;

[1] The update rule: $x_{k+1} = \arg\min_x \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|^2 + g(x)$.

**Assumption**

$\min_{x \in \mathbb{R}^{n \times m}} F(x) = f(x) + g(x)$, with $F$ satisfying the Kurdyka-Łojasiewicz (KL) property with exponent $\theta \in (0, 1]$;

Reference [BST14]:

- Only one accumulation point;
- if $\theta = 1$, then the proximal gradient method terminates in finite steps;
- if $\theta \in [0.5, 1)$, then $\|x_k - x_*\| < C_1 d^k$ for $C_1 > 0$ and $d \in (0, 1)$;
- if $\theta \in (0, 0.5)$, then $\|x_k - x_*\| < C_2 k^{\frac{-1}{1-2\theta}}$ for $C_2 > 0$;

### Euclidean proximal mapping

$$d_k = \arg \min_{p \in \mathbb{R}^{n \times m}} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p)$$

---

In the Riemannian setting:

- How to define the proximal mapping?
- Can be solved cheaply?
- Share the same convergence rate?

**Euclidean proximal mapping**

$$d_k = \arg \min_{p \in \mathbb{R}^{n \times m}} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p)$$

---

**A Riemannian proximal mapping [CMSZ20]**

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \nabla f(x_k), \eta \rangle + \frac{L}{2} \|\eta\|_F^2 + g(x_k + \eta)$;

- Only works for embedded submanifold;

[1][CMSZ18]: S. Chen, S. Ma, M. C. So, and T. Zhang, Proximal gradient method for nonsmooth optimization over the Stiefel manifold. SIAM Journal on Optimization, 30(1):210-239, 2020

# Euclidean proximal mapping

$$d_k = \arg \min_{p \in \mathbb{R}^{n \times m}} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p)$$

---

# A Riemannian proximal mapping [CMSZ20]

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \nabla f(x_k), \eta \rangle + \frac{L}{2} \|\eta\|_F^2 + g(x_k + \eta)$;

- Only works for embedded submanifold;
- Proximal mapping is defined in tangent space;

[1][CMSZ18]: S. Chen, S. Ma, M. C. So, and T. Zhang, Proximal gradient method for nonsmooth optimization over the Stiefel manifold. SIAM Journal on Optimization, 30(1):210-239, 2020

## Euclidean proximal mapping

$$d_k = \arg \min_{p \in \mathbb{R}^{n \times m}} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p)$$

---

## A Riemannian proximal mapping [CMSZ20]

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \nabla f(x_k), \eta \rangle + \frac{L}{2} \|\eta\|_F^2 + g(x_k + \eta)$;

- Only works for embedded submanifold;
- Proximal mapping is defined in tangent space;
- Convex programming;

[1] [CMSZ18]: S. Chen, S. Ma, M. C. So, and T. Zhang, Proximal gradient method for nonsmooth optimization over the Stiefel manifold. SIAM Journal on Optimization, 30(1):210-239, 2020

## Euclidean proximal mapping

$$d_k = \arg \min_{p \in \mathbb{R}^{n \times m}} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p)$$

---

## ManPG [CMSZ20]

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \nabla f(x_k), \eta \rangle + \frac{L}{2} \|\eta\|_F^2 + g(x_k + \eta)$;

- Only works for embedded submanifold;
- Proximal mapping is defined in tangent space;
- Convex programming;
- Solved efficiently for the Stiefel manifold by a semi-Newton algorithm [XLWZ18];

## Euclidean proximal mapping

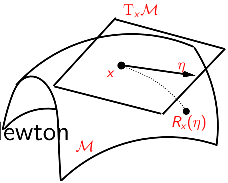
$$d_k = \arg \min_{p \in \mathbb{R}^{n \times m}} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p)$$

---

## ManPG [CMSZ20]

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \nabla f(x_k), \eta \rangle + \frac{L}{2} \|\eta\|_F^2 + g(x_k + \eta)$;
2. $x_{k+1} = R_{x_k}(\alpha_k \eta_k)$ with an appropriate step size $\alpha_k$;

- Only works for embedded submanifold;
- Proximal mapping is defined in tangent space;
- Convex programming;
- Solved efficiently for the Stiefel manifold by a semi-Newton algorithm [XLWZ18];
- Step size 1 is not necessary decreasing;



## Euclidean proximal mapping

$$d_k = \arg \min_{p \in \mathbb{R}^{n \times m}} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p)$$

---

## ManPG [CMSZ20]

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \nabla f(x_k), \eta \rangle + \frac{L}{2} \|\eta\|_F^2 + g(x_k + \eta)$;
2. $x_{k+1} = R_{x_k}(\alpha_k \eta_k)$ with an appropriate step size $\alpha_k$;

- Convergence to a stationary point;

## Euclidean proximal mapping

$$d_k = \arg \min_{p \in \mathbb{R}^{n \times m}} \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(x_k + p)$$

---

## ManPG [CMSZ20]

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \nabla f(x_k), \eta \rangle + \frac{L}{2} \|\eta\|_F^2 + g(x_k + \eta)$;
2. $x_{k+1} = R_{x_k}(\alpha_k \eta_k)$ with an appropriate step size $\alpha_k$;

- Convergence to a stationary point;
- No convergence rate analysis;

GOAL: Develop a Riemannian proximal gradient method with convergence rate analysis and good numerical performance for some instances

GOAL: Develop a Riemannian proximal gradient method with convergence rate analysis and good numerical performance for some instances

---

## A New Riemannian Proximal Gradient Method

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k}\mathcal{M}} \underbrace{\langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2}_{\text{Riemannian metric}} + g(\underbrace{R_{x_k}(\eta)}_{\text{replace } x_k + \eta});$
2. $x_{k+1} = R_{x_k}(\eta_k);$

- General framework for Riemannian optimization;

GOAL: Develop a Riemannian proximal gradient method with convergence rate analysis and good numerical performance for some instances

---

## A New Riemannian Proximal Gradient Method

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \underbrace{\langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2}_{\text{Riemannian metric}} + g(\underbrace{R_{x_k}(\eta)}_{\text{replace } x_k + \eta});$
2. $x_{k+1} = R_{x_k}(\eta_k);$

- General framework for Riemannian optimization;
- Step size can be fixed to be 1;

Assumption:

1. The function $F$ is bounded from below and the sublevel set $\Omega_{x_0} = \{x \in \mathcal{M} \mid F(x) \leq F(x_0)\}$ is compact;

---

This assumption hold if, for example, $F$ is continuous and $\mathcal{M}$ is compact.

$$\text{Sparse PCA: } \min_{X \in \mathrm{St}(p,n)} -\mathrm{trace}(X^T A^T A X) + \lambda \|X\|_1,$$

Assumption:

1. The function $F$ is bounded from below and the sublevel set $\Omega_{x_0} = \{x \in \mathcal{M} \mid F(x) \leq F(x_0)\}$ is compact;
2. The function $f$ is $L$-retraction-smooth with respect to the retraction $R$ in the sublevel set $\Omega_{x_0}$.

---

**Definition**

A function $h : \mathcal{M} \to \mathbb{R}$ is called $L$-retraction-smooth with respect to a retraction $R$ in $\mathcal{N} \subseteq \mathcal{M}$ if for any $x \in \mathcal{N}$ and any $\mathcal{S}_x \subseteq \mathrm{T}_x \mathcal{M}$ such that $R_x(\mathcal{S}_x) \subseteq \mathcal{N}$, we have that

$$h(R_x(\eta)) \leq h(x) + \langle \mathrm{grad}\, h(x), \eta \rangle_x + \frac{L}{2} \|\eta\|_x^2, \quad \forall \eta \in \mathcal{S}_x.$$

Assumption:

1. The function $F$ is bounded from below and the sublevel set $\Omega_{x_0} = \{x \in \mathcal{M} \mid F(x) \leq F(x_0)\}$ is compact;
2. The function $f$ is $L$-retraction-smooth with respect to the retraction $R$ in the sublevel set $\Omega_{x_0}$.

---

if the following conditions hold, then $f$ is $L$-retraction-smooth with respect to the retraction $R$ in the manifold $\mathcal{M}$ [BAC18, Lemma 2.7]

- $\mathcal{M}$ is a compact Riemannian submanifold of a Euclidean space $\mathbb{R}^n$;
- the retraction $R$ is globally defined;
- $f : \mathbb{R}^n \to \mathbb{R}$ is $L$-smooth in the convex hull of $\mathcal{M}$;

$$\text{Sparse PCA:} \quad \min_{X \in \mathrm{St}(p,n)} -\mathrm{trace}(X^T A^T A X) + \lambda \|X\|_1,$$

Assumption:

1. The function $F$ is bounded from below and the sublevel set $\Omega_{x_0} = \{x \in \mathcal{M} \mid F(x) \leq F(x_0)\}$ is compact;
2. The function $f$ is $L$-retraction-smooth with respect to the retraction $R$ in the sublevel set $\Omega_{x_0}$.

---

Theoretical results:

- For any accumulation point $x_*$ of $\{x_k\}$, $x_*$ is a stationary point, i.e., $0 \in \partial F(x_*)$.

Additional Assumptions:

- $f$ and $g$ are retraction-convex in $\Omega \supseteq \Omega_{x_0}$;
- Retraction approximately satisfies the triangle relation in $\Omega$: for all $x, y, z \in \Omega$,

$$\left| \|\xi_x - \eta_x\|_x^2 - \|\zeta_y\|_y^2 \right| \leq \kappa \|\eta_x\|_x^2, \text{ for a constant } \kappa$$

where $\eta_x = R_x^{-1}(y)$, $\xi_x = R_x^{-1}(z)$, $\zeta_y = R_y^{-1}(z)$.

---

Theoretical results:

- Convergence rate $O(1/k)$:

$$F(x_k) - F(x_*) \leq \frac{1}{k} \left( \frac{L}{2} \|R_{x_0}^{-1}(x_*)\|_{x_0}^2 + \frac{L\kappa C}{2}(F(x_0) - F(x_*)) \right).$$

Assumption:

1. Assumptions for the global convergence

---

1. The function $F$ is bounded from below and the sublevel set $\Omega_{x_0} = \{x \in \mathcal{M} \mid F(x) \leq F(x_0)\}$ is compact;
2. The function $f$ is $L$-retraction-smooth with respect to the retraction $R$ in the sublevel set $\Omega_{x_0}$.

$$\min_{X \in \mathrm{St}(p,n)} -\mathrm{trace}(X^T A^T A X) + \lambda \|X\|_1,$$

Assumption:

1. Assumptions for the global convergence
2. $f$ is locally Lipschitz continuously differentiable

---

**Definition ( [AMS08, 7.4.3])**

A function $f$ on $\mathcal{M}$ is Lipschitz continuously differentiable if it is differentiable and if there exists $\beta_1$ such that, for all $x, y$ in $\mathcal{M}$ with $\mathrm{dist}(x, y) < i(\mathcal{M})$, it holds that

$$\|\mathcal{P}_{\gamma}^{0 \leftarrow 1} \operatorname{grad} f(y) - \operatorname{grad} f(x)\|_x \leq \beta_1 \operatorname{dist}(x, y),$$

where $\gamma$ is the unique minimizing geodesic with $\gamma(0) = x$ and $\gamma(1) = y$.

Assumption:

1. Assumptions for the global convergence
2. $f$ is locally Lipschitz continuously differentiable

---

If $f$ is smooth and the manifold $\mathcal{M}$ is compact, then the function $f$ is Lipschitz continuously differentiable. [AMS08, Proposition 7.4.5 and Corollary 7.4.6].

$$\min_{X \in \mathrm{St}(p,n)} -\mathrm{trace}(X^T A^T A X) + \lambda \|X\|_1,$$

Assumption:

1. Assumptions for the global convergence
2. $f$ is locally Lipschitz continuously differentiable
3. $F$ satisfies the Riemannian KL property defined in [BdCNO11]

---

**Definition**

A continuous function $f : \mathcal{M} \to \mathbb{R}$ is said to have the Riemannian KL property at $x \in \mathcal{M}$ if and only if there exists $\varepsilon \in (0, \infty]$, a neighborhood $U \subset \mathcal{M}$ of $x$, and a continuous concave function $\varsigma : [0, \varepsilon] \to [0, \infty)$ such that

- $\varsigma(0) = 0$, $\varsigma$ is $C^1$ on $(0, \varepsilon)$, and $\varsigma' > 0$ on $(0, \eta)$,
- For every $y \in U$ with $f(x) < f(y) < f(x) + \varepsilon$, we have

$$\varsigma'(f(y) - f(x)) \operatorname{dist}(0, \partial f(y)) \geq 1,$$

where $\operatorname{dist}(0, \partial f(y)) = \inf\{\|v\|_y : v \in \partial f(y)\}$ and $\partial$ denotes the Riemannian generalized subdifferential. The function $\varsigma$ is called the desingularising function.

The desingularising function $\varsigma$ quantifies the relationship between $f(x_k) - f(x_*)$ and $\operatorname{dist}(0, \partial f(x_k))$.

Assumption:

1. Assumptions for the global convergence
2. $f$ is locally Lipschitz continuously differentiable
3. $F$ satisfies the Riemannian KL property defined in [BdCNO11]

---

Theoretical results:

- it holds that

$$\sum_{k=0}^{\infty} \mathrm{dist}(x_k, x_{k+1}) < \infty.$$

Therefore, there exists only a unique accumulation point.

Assumption:

1. Assumptions for the global convergence
2. $f$ is locally Lipschitz continuously differentiable
3. $F$ satisfies the Riemannian KL property defined in [BdCNO11]

---

Theoretical results:

- If the desingularising function has the form $\varsigma(t) = \frac{C}{\theta} t^{\theta}$ for $C > 0$ and $\theta \in (0, 1]$ for all $x \in \Omega_{x_0}$, then
    - if $\theta = 1$, then the Riemannian proximal gradient method terminates in finite steps;
    - if $\theta \in [0.5, 1)$, then $\|x_k - x_*\| < C_1 d^k$ for $C_1 > 0$ and $d \in (0, 1)$;
    - if $\theta \in (0, 0.5)$, then $\|x_k - x_*\| < C_2 k^{\frac{-1}{1-2\theta}}$ for $C_2 > 0$;

How to verify if a function satisfies the Riemannian KL property?

**Theorem**

*Restriction of a semialgebraic Function onto Stiefel manifold satisfies the Riemannian KL property with desingularising function in the form of* $\varsigma(t) = \frac{C}{\theta} t^{\theta}$, *where* $\theta \in (0, 1]$ *and* $C > 0$.

How to verify if a function satisfies the Riemannian KL property?

**Theorem**

*Restriction of a semialgebraic Function onto Stiefel manifold satisfies the Riemannian KL property with desingularising function in the form of* $\varsigma(t) = \frac{C}{\theta} t^{\theta}$, *where* $\theta \in (0, 1]$ *and* $C > 0$.

---

**Definition (Semialgebraic functions)**

1. A subset $\mathcal{S}$ of $\mathbb{R}^n$ is called semialgebraic if there exists a finite number of polynomial function $g_{ij}, h_{ij} : \mathbb{R}^n \to \mathbb{R}$ such that

$$\mathcal{S} = \cup_{j=1}^{p} \cap_{i=1}^{q} \{u \in \mathbb{R}^n \mid g_{ij}(u) = 0 \text{ and } h_{ij}(u) < 0\}.$$

2. Let $\mathcal{A} \subset \mathbb{R}^n$ be semialgebraic. A function : $\mathcal{A} \to \mathbb{R}$ is semialgebraic if its graph is semialgebraic in $\mathbb{R}^{n+1}$.

How to verify if a function satisfies the Riemannian KL property?

**Theorem**

*Restriction of a semialgebraic Function onto Stiefel manifold satisfies the Riemannian KL property with desingularising function in the form of $\varsigma(t) = \frac{C}{\theta} t^{\theta}$, where $\theta \in (0, 1]$ and $C > 0$.*

---

Function $-\mathrm{trace}(X^T A^T A X) + \lambda \|X\|_1$ is a semialgebraic function on $\mathbb{R}^{n \times p}$.

Two sparse PCA models:

- first model: [GHT15]

$$\min_{X \in OB(p,n)} \|X^T A^T A X - D^2\|_F^2 + \lambda \|X\|_1,$$

where $A \in \mathbb{R}^{m \times n}$ is a data matrix, $D$ is the diagonal matrix with dominant singular values of $A$,
$OB(p,n) = \{X \in \mathbb{R}^{n \times p} \mid \mathrm{diag}(X^T X) = I_p\}$, $p \leq m$;

- second model

$$\min_{X \in \mathrm{St}(p,n)} -\mathrm{trace}(X^T A^T A X) + \lambda \|X\|_1.$$

$$\text{First: } \min_{X \in OB(p,n)} \|X^T A^T A X - D^2\|_F^2 + \lambda \|X\|_1, .$$

Table: An average result of 10 random tests. $n = 128$, $m = 20$, $r = 4$. $\delta = (L\|x_{k+1} - x_k\|)^2$. The subscript $k$ indicates a scale of $10^k$.

| $\lambda$ | Algo | iter | time | $f$ | $\delta$ | spar. | navar |
|---|---|---|---|---|---|---|---|
| 3 | ManPG | 11791 | 1.40 | $8.33_1$ | $5.11_{-6}$ | 0.54 | 0.86 |
| | RPG | 11679 | 0.94 | $8.33_1$ | $5.11_{-6}$ | 0.54 | 0.86 |

- ManPG: the method in [CMSZ20];
- RPG: the new Riemannian proximal gradient;

See more numerical experiments in [HW21].

Second: $\min_{X \in \mathrm{St}(p,n)} -\mathrm{trace}(X^T A^T A X) + \lambda \|X\|_1.$

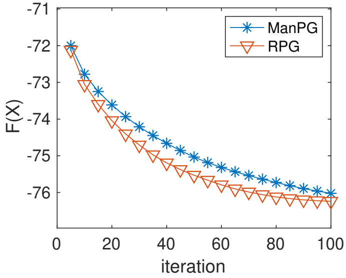



Figure: Two typical runs of ManPG, RPG, A-ManPG, and A-RPG for the Sparse PCA problem. $n = 1024$, $p = 4$, $\lambda = 2$, $m = 20$.

See more numerical experiments in [HW21].

# Ongoing related research

Main references: Huang and Wei, Riemannian proximal gradient methods, *Mathematical Programming, Series A*, doi:10.1007/s10107-021-01632-3, 2021

---

- Riemannian proximal gradient methods without solving the subproblem exactly;
- Develop related geometry tools for other manifolds
- Riemannian KL property for more objective and manifolds
- Applications
    - Sparse PCA
    - Clustering
    - Community detection
    - Image impainting with Low rank sparse constraints

# Summary

- Riemannian optimization problem statement
- Literature review
- My related work
- A Riemannian proximal gradient method
- Ongoing related research

# References I


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

Nicolas Boumal, P-A Absil, and Coralia Cartis.
Global rates of convergence for nonconvex optimization on manifolds.
*IMA Journal of Numerical Analysis*, 39(1):1–33, 02 2018.

G. C. Bento, J. X. de Cruz Neto, and P. R. Oliveira.
Convergence of inexact descent methods for nonconvex optimization on Riemannian manifold.
*arXiv preprint arXiv:1103.4828*, 2011.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle.
Proximal alternating linearized minimization for nonconvex and nonsmooth problems.
*Mathematical Programming*, 146(1-2):459–494, 2014.

Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang.
Proximal gradient method for nonsmooth optimization over the Stiefel manifold.
*SIAM Journal on Optimization*, 30(1):210–239, 2020.

Matthieu Genicot, Wen Huang, and Nickolay T. Trendafilov.
Weakly correlated sparse components with nearly orthonormal loadings.
In *Geometric Science of Information*, pages 484–490, 2015.

W. Huang, P.-A. Absil, and K. A. Gallivan.
A Riemannian symmetric rank-one trust-region method.
*Mathematical Programming*, 150(2):179–216, February 2015.

W. Huang, P.-A. Absil, and K. A. Gallivan.
Intrinsic representation of tangent vectors and vector transport on matrix manifolds.
*Numerische Mathematik*, 136(2):523–543, 2016.

# References II


Wen Huang, P.-A. Absil, and K. A. Gallivan.
A Riemannian BFGS method without differentiated retraction for nonconvex optimization problems.
*SIAM Journal on Optimization*, 28(1):470–495, 2018.

W. Huang, P.-A. Absil, K. A. Gallivan, and P. Hand.
ROPTLIB: an object-oriented C++ library for optimization on Riemannian manifolds.
*ACM Transactions on Mathematical Software*, 4(44):43:1–43:21, 2018.

W. Huang and K. A. Gallivan.
A limited-memory Riemannian symmetric rank-one trust-region method with an efficient algorithm for its subproblem.
In *Proceedings of the 24th Internaltional Symposium on Mathematical Theory of Networks and Systems*, 2021.
https://www.math.fsu.edu/~whuang2/papers/ALMRTRSR1.htm, accepted.

W. Huang, K. A. Gallivan, and P.-A. Absil.
A Broyden Class of Quasi-Newton Methods for Riemannian Optimization.
*SIAM Journal on Optimization*, 25(3):1660–1685, 2015.

W. Huang, K. A. Gallivan, Anuj Srivastava, and P.-A. Absil.
Riemannian optimization for registration of curves in elastic shape analysis.
*Journal of Mathematical Imaging and Vision*, 54(3):320–343, 2015.
DOI:10.1007/s10851-015-0606-8.

Wen Huang, K. A. Gallivan, and Xiangxiong Zhang.
Solving phaselift by low rank Riemannian optimization methods.
In *Proceedings of the International Conference on Computational Science (ICCS2016), accepted*, 2016.

Wen Huang and Paul Hand.
Blind deconvolution by a steepest descent algorithm on a quotient manifold.
*SIAM Journal on Imaging Sciences*, 11(4):2757–2785, 2018.

S. Hosseini, W. Huang, and R. Yousefpour.
Line search algorithms for locally Lipschitz functions on Riemannian manifolds.
*SIAM Journal on Optimization*, 28(1):596–619, 2018.

W. Huang and K. Wei.
An extension of FISTA to Riemannian optimization for sparse PCA.
arXiv:1909.05485, 2019.

W. Huang and K. Wei.
Riemannian proximal gradient methods.
*Mathematical Programming*, 2021.

Xiantao Xiao, Yongfeng Li, Zaiwen Wen, and Liwei Zhang.
A regularized semi-smooth newton method with projection steps for composite convex programs.
*Journal of Scientific Computing*, 76(1):364–389, Jul 2018.

# Thank you

Thank you!