Riemannian Proximal Gradient Methods

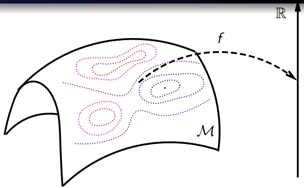
Wen Huang
Xiamen University

Symposium on the Frontiers of Mathematical Optimization Research
Guangxi University July 22, 2020

This is joint work with Ke Wei at Fudan University.

Problem Statement

Optimization on Manifolds with Structure:



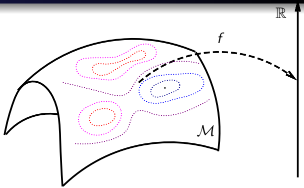
$$\min_{x \in \mathcal{M}} F(x) = f(x) + g(x),$$

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

[1] a penalized version of the ScoTLASS introduced in [JTU03].

Problem Statement

Optimization on Manifolds with Structure:



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

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

Applications: sparse PCA, sparse blind deconvolution, sparse low rank image representation, etc [JTU03, GHT15, SQ16, ZLK+17]

A sparse PCA optimization model:[1]

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

[1] a penalized version of the ScoTLASS introduced in [JTU03].

Existing Nonsmooth Optimization on Manifolds

$F : \mathcal{M} \to \mathbb{R}$ is Lipschitz continuous

- Huang (2013), Gradient sampling method without convergence analysis.
- Grohs and Hosseini (2015), Two ϵ-subgradient-based optimization methods using line search strategy and trust region strategy, respectively. Any limit point is a critical point.
- Hosseini and Uschmajew (2017), Gradient sampling method and any limit point is a critical point.
- Hosseini and Huang and Yousefpour (2018), Merge ϵ-subgradient-based and quasi-Newton ideas and show any limit point is a critical point.

Existing Nonsmooth Optimization on Manifolds

$$F : \mathcal{M} \to \mathbb{R} \text{ is convex}$$

- Zhang and Sra (2016), subgradient-based method and function value converges to the optimal $O(1/\sqrt{k})$.
- Ferreira and Oliveira (2002) proximal point method, convergence using convexity
  Bento, da Cruz Neto and Oliveira (2011), convergence using Kurdyka-Łojasiewicz (KL); and
  Bento, Ferreira and Melo (2017), function value converges to the optimal $O(1/k)$ on Hadamard manifold using convexity

Existing Nonsmooth Optimization on Manifolds

$F = f + g$, where f is L-con, and g is non-smooth

- Chen, Ma, So, and Zhang (2018), A proximal gradient method with global convergence
- Huang and Wei (2019), A Riemannian proximal gradient method with convergence rate analyses

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

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

2

A Euclidean Proximal Gradient Method

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[2]:

initial iterate: x_0,

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

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

A Euclidean Proximal Gradient Method

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[2]:

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;

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

A Euclidean Proximal Gradient Method

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[2]:

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 ∇f;

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

A Euclidean Proximal Gradient Method

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[2]:

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 ∇f;
- Proximal mapping: easy to compute;

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

A Euclidean Proximal Gradient Method

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[2]:

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 ∇f;
- Proximal mapping: easy to compute;
- Any limit point is a critical point;

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

Convergence Rates

Assumption

$\min_{x \in \mathbb{R}^{n \times m}} F(x) = f(x) + g(x)$, with convex f and g;

- $O(1/k)$ sublinear convergence rate:

$$F(x_k) - F(x_*) \leq C/k, \text{ for a constant } C;$$

Convergence Rates

Assumption

$\min_{x \in \mathbb{R}^{n \times m}} F(x) = f(x) + g(x)$, with convex f and g;

- $O(1/k)$ sublinear convergence rate:

$$F(x_k) - F(x_*) \leq C/k, \text{ for a constant } C;$$

- Optimal gradient method: $O(1/k^2)$ [Dar83, Nes83]

Convergence Rates

Assumption

$\min_{x \in \mathbb{R}^{n \times m}} F(x) = f(x) + g(x)$, with convex f and g;

- $O(1/k)$ sublinear convergence rate:

$$F(x_k) - F(x_*) \leq C/k, \text{ for a constant } C;$$

- Optimal gradient method: $O(1/k^2)$ [Dar83, Nes83]
- For example: FISTA [BT09]

initial iterate: x_0 and let $y_0 = x_0$, $t_0 = 1$,

$$\begin{cases} d_k = \arg\min_{p \in \mathbb{R}^{n \times m}} \langle \nabla f(y_k), p \rangle + \frac{L}{2} \|p\|_F^2 + g(y_k + p), \\ x_{k+1} = y_k + d_k, \\ t_{k+1} = \frac{1 + \sqrt{4t_k^2 + 1}}{2}, \\ y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}} (x_{k+1} - x_k). \end{cases}$$

Local Convergence Rates

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 $C_2 > 0$;

Difficulties in the Riemannian Setting

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?

A Riemannian Proximal Gradient Method in [CMMCSZ20]

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 [CMMCSZ20]

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;

[2] [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

A Riemannian Proximal Gradient Method in [CMMCSZ20]

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 [CMMCSZ20]

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;

[2] [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

A Riemannian Proximal Gradient Method in [CMMCSZ20]

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 [CMMCSZ20]

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;

[2] [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

A Riemannian Proximal Gradient Method in [CMMCSZ20]

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 [CMMCSZ20]

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];

A Riemannian Proximal Gradient Method in [CMMCSZ20]

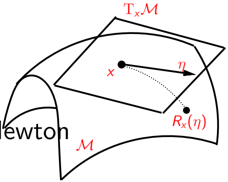
Euclidean proximal mapping

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

ManPG [CMMCSZ20]

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 α_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;



A Riemannian Proximal Gradient Method in [CMMCSZ20]

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 [CMMCSZ20]

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 α_k;

- Convergence to a stationary point;

A Riemannian Proximal Gradient Method in [CMMCSZ20]

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 [CMMCSZ20]

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

- General framework for Riemannian optimization;

New Riemannian Proximal Gradient Methods

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

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

Assumptions and Convergence Result

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.

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

Assumptions and Convergence Result

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 Ω_{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

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

Assumptions and Convergence Result

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 Ω_{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 $\mathcal{M}$;

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

Assumptions and Convergence Result

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 Ω_{x_0}.

Theoretical results:

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

Assumptions and Convergence Rate

Additional Assumptions:

- f and g are retraction-convex in $\Omega \supseteq \Omega_{x_0}$;

Definition

A function $h : \mathcal{M} \to \mathbb{R}$ is called retraction-convex 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 $R_x(\mathcal{S}_x) \subseteq \mathcal{N}$, there exists a tangent vector $\zeta \in \mathrm{T}_x\, \mathcal{M}$ such that $q_x = h \circ R_x$ satisfies

$$q_x(\eta) \geq q_x(\xi) + \langle \zeta, \eta - \xi \rangle_x \quad \forall \eta, \xi \in \mathcal{S}_x. \tag{2}$$

Note that $\zeta = \mathrm{grad}\, q_x(\xi)$ if h is differentiable; otherwise, ζ is any subgradient of q_x at ξ.

Assumptions and Convergence Rate

Additional Assumptions:

- f and g are retraction-convex in $\Omega \supseteq \Omega_{x_0}$;

Lemma

Given $x \in \mathcal{M}$ and a twice continuously differentiable function $h : \mathcal{M} \to \mathbb{R}$, if one of the following conditions holds:

- $\mathrm{Hess}\, h$ *is positive definite at x, and the retraction is second order;*
- *The manifold $\mathcal{M}$ is an embedded submanifold of $\mathbb{R}^n$ endowed with the Euclidean metric; $\mathcal{W}$ is an open subset of $\mathbb{R}^n$; $x \in \mathcal{W}$; $h : \mathcal{W} \subset \mathbb{R}^n \to \mathbb{R}$ is a μ-strongly convex function in the Euclidean setting for a sufficient large μ; the retraction is second order;*

then there exists a neighborhood of x, denoted by $\mathcal{N}_x$, such that the function $h : \mathcal{M} \to \mathbb{R}$ is retraction-convex in $\mathcal{N}_x$.

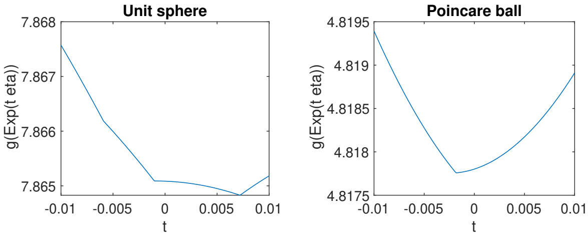
Assumptions and Convergence Rate

Additional Assumptions:

- f and g are retraction-convex in $\Omega \supseteq \Omega_{x_0}$;

Nonsmooth? Example: $g(x) = \|x\|_1$ with exponential mapping

- unit sphere: $\{x \in \mathbb{R}^n \mid x^T x = 1\}$, $n = 100$
- Poincaré ball model [GBH18]: $\{x \in \mathbb{R}^n \mid x^T x < 1\}$, $n = 100$
- $g(\mathrm{Exp}_x(t\eta_x))$ versus t



[GBH18] Ganea et al., Hyperbolic entailment cones for learning hierarchical embedding, ICML, 2018.

Assumptions and Convergence Rate

Additional Assumptions:

- f and g are retraction-convex in $\Omega \supseteq \Omega_{x_0}$;
- Retraction approximately satisfies the triangle relation in Ω: 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 $\eta_x = R_x^{-1}(y)$, $\xi_x = R_x^{-1}(z)$, $\zeta_y = R_y^{-1}(z)$.

- In the Euclidean setting: $\eta_x = R_x^{-1}(y) = y - x$, $\xi_x = R_x^{-1}(z) = z - x$, $\zeta_y = R_y^{-1}(z) = z - y$:

$$\xi_x - \eta_x = (z - x) - (y - x) = z - y = \zeta_y.$$

- Holds on the unit sphere.

Assumptions and Convergence Rate

Additional Assumptions:

- f and g are retraction-convex in $\Omega \supseteq \Omega_{x_0}$;
- Retraction approximately satisfies the triangle relation in Ω: 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 $\eta_x = R_x^{-1}(y)$, $\xi_x = R_x^{-1}(z)$, $\zeta_y = R_y^{-1}(z)$.

Table: Exponential mapping on the Stiefel manifold with the Euclidean metric $\langle \eta_x, \xi_x \rangle_x = \mathrm{trace}(\eta_x^T \xi_x)$. Left $= \left| \|\xi_x - \eta_x\|_x^2 - \|\zeta_y\|_y^2 \right|$

$(n,p) = (10,1)$		$(n,p) = (10,4)$		$(n,p) = (10,10)$	
$\|\eta_x\|$	Left	$\|\eta_x\|$	Left	$\|\eta_x\|$	Left
5.00_{-2}	7.83_{-5}	5.00_{-2}	1.83_{-5}	5.00_{-2}	2.14_{-6}
2.50_{-2}	1.80_{-5}	2.50_{-2}	4.27_{-6}	2.50_{-2}	4.72_{-7}
1.25_{-2}	4.25_{-6}	1.25_{-2}	1.01_{-6}	1.25_{-2}	1.11_{-7}
6.25_{-3}	1.03_{-6}	6.25_{-3}	2.46_{-7}	6.25_{-3}	2.68_{-8}

Additional Assumptions:

- f and g are retraction-convex in $\Omega \supseteq \Omega_{x_0}$;
- Retraction approximately satisfies the triangle relation in Ω: 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 $\eta_x = R_x^{-1}(y)$, $\xi_x = R_x^{-1}(z)$, $\zeta_y = R_y^{-1}(z)$.

Table: Exponential mapping on the Stiefel manifold with the canonical metric $\langle \eta_x, \xi_x \rangle_x = \mathrm{trace}(\eta_x^T (I - XX^T/2)\xi_x)$. $\mathrm{Left} = \left| \|\xi_x - \eta_x\|_x^2 - \|\zeta_y\|_y^2 \right|$

$(n,p) = (10,2)$		$(n,p) = (10,4)$		$(n,p) = (10,9)$	
$\|\eta_x\|$	Left	$\|\eta_x\|$	Left	$\|\eta_x\|$	Left
5.00_{-2}	3.55_{-5}	5.00_{-2}	1.15_{-5}	5.00_{-2}	8.39_{-6}
2.50_{-2}	8.06_{-6}	2.50_{-2}	2.58_{-6}	2.50_{-2}	1.89_{-6}
1.25_{-2}	1.90_{-6}	1.25_{-2}	6.08_{-7}	1.25_{-2}	4.45_{-7}
6.25_{-3}	4.61_{-7}	6.25_{-3}	1.47_{-7}	6.25_{-3}	1.08_{-7}

Assumptions and Convergence Rate

Additional Assumptions:

- f and g are retraction-convex in $\Omega \supseteq \Omega_{x_0}$;
- Retraction approximately satisfies the triangle relation in Ω: 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 $O(1/k)$:

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

Riemannian FISTA Method with $O(1/k^2)$?

FISTA initial iterate: x_0 and let $y_0 = x_0$, $t_0 = 1$

1. $d_k = \arg\min_{p \in \mathbb{R}^{n\times m}} \langle \nabla f(y_k), p \rangle + \frac{L}{2}\|p\|_F^2 + g(y_k + p)$
2. $x_{k+1} = y_k + d_k$
3. $t_{k+1} = \frac{1+\sqrt{4t_k^2+1}}{2}$
4. $y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}}(x_{k+1} - x_k)$

Possible Riemannian generalizations:

- Step 1: Riemannian proximal mapping
- Step 2: Retraction
- Step 4: multiple generalizations

Difficulties for $O(1/k^2)$ convergence rate, e.g.,

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

Assumptions and Local Convergence Result

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 Ω_{x_0}.

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

Assumptions and Local Convergence Result

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 β_1 such that, for all x, y in $\mathcal{M}$ with $\mathrm{dist}(x, y) < i(\mathcal{M})$, it holds that

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

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

Assumptions and Local Convergence Result

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].

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

Assumptions and Local Convergence Result

Assumption:

1. Assumptions for the global convergence
2. f is locally Lipschitz continuously differentiable
3. F satisfies the Riemannian KL property [BCNO11]

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

- $\varsigma(0) = 0$, ς 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 ∂ denotes the Riemannian generalized subdifferential. The function ς is called the desingularising function.

Assumptions and Local Convergence Result

Assumption:

1. Assumptions for the global convergence
2. f is locally Lipschitz continuously differentiable
3. F satisfies the Riemannian KL property [BCNO11]

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.

Assumptions and Local Convergence Result

Assumption:

1. Assumptions for the global convergence
2. f is locally Lipschitz continuously differentiable
3. F satisfies the Riemannian KL property [BCNO11]

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 $C_2 > 0$;

Riemannian KL property

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

Theorem

Given $x \in \mathcal{M}$, let $(\phi, \mathcal{U})$ denote a chart of $\mathcal{M}$ covering x, i.e., $x \in \mathcal{U}$. We assume that $F \circ \phi^{-1} : \mathbb{R}^d \to \mathbb{R}$ satisfies the Euclidean KL property at $\phi(x)$ with the desingularising function $\tilde{\varsigma}_x$, then F satisfies the Riemannian KL property at x with the desingularising function $\tilde{\varsigma}_x / C_x$, where C_x is a constant.

Similar result is given in [BCNO11]: F is a $\mathcal{C}$-function $\Longrightarrow$ F satisfies the Riemannian KL property.

Definition (Semialgebraic sets, mappings and 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 $\mathcal{A} \subseteq \mathbb{R}^m$ and $\mathcal{B} \subseteq \mathbb{R}^n$ be two semialgebraic sets. A mapping $: \mathcal{A} \to \mathcal{B}$ is semialgebraic if its graph is semialgebraic in $\mathbb{R}^{m+n}$. If $n = 1$, then the mapping is also called a semialgebraic function.

Continuous semialgebraic functions satisfy the Euclidean KL property with desingularising function in the form of $\varsigma(t) = \frac{C}{\theta} t^{\theta}$, where $\theta \in (0, 1]$ and $C > 0$.

Riemannian KL property on the Stiefel manifold

Restriction of a semialgebraic Function onto Stiefel manifold satisfies the Riemannian KL property

1. For any point $x \in \mathrm{St}(p, n)$, construct a chart $(\phi, \mathcal{U})$ such that $x \in \mathcal{U}$ and ϕ is a semialgebraic mapping
2. Inverse of ϕ is semialgebraic mapping
3. The composition function $f \circ \phi^{-1}$ is a semialgebraic function
4. $f \circ \phi^{-1}$ satisfies the Euclidean KL property with desingularising function $\varsigma(t) = \frac{C}{\theta} t^{\theta}$
5. $f : \mathcal{M} \to \mathbb{R}$ satisfies the Riemannian KL property with desingularising function $\varsigma(t) = \frac{\tilde{C}}{\theta} t^{\theta}$

Riemannian KL property on the Stiefel manifold

Restriction of a semialgebraic Function onto Stiefel manifold satisfies the Riemannian KL property

1. For any point $x \in \mathrm{St}(p, n)$, construct a chart $(\phi, \mathcal{U})$ such that $x \in \mathcal{U}$ and ϕ is a semialgebraic mapping
2. Inverse of ϕ is semialgebraic mapping
3. The composition function $f \circ \phi^{-1}$ is a semialgebraic function
4. $f \circ \phi^{-1}$ satisfies the Euclidean KL property with desingularising function $\varsigma(t) = \frac{C}{\theta} t^{\theta}$
5. $f : \mathcal{M} \to \mathbb{R}$ satisfies the Riemannian KL property with desingularising function $\varsigma(t) = \frac{\tilde{C}}{\theta} t^{\theta}$

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

Verify Assumptions for sparse PCA models

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

Therefore, all assumptions for global convergence and local convergence hold for the two sparse PCA models.

- All accumulation points are critical
- Local convergence:
  - Accumulation point is unique;
  - if $\theta = 1$, then the 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 $C_2 > 0$;

Riemannian subproblem

$$\eta_x = \arg\min_{\eta \in \mathrm{T}_x \mathcal{M}} \ell_x(\eta) := \langle \nabla f(x), \eta \rangle_x + \frac{L}{2} \|\eta\|_x^2 + g(R_x(\eta))$$

In some cases, the subproblem can be solved by exploiting the structure of the manifold;

Solving the Riemannian Proximal Mapping

initial iterate: $\eta_0 \in \mathrm{T}_x \mathcal{M}$, $\sigma \in (0, 1)$, $k = 0$;

1. $y_k = R_x(\eta_k)$;
2. Compute
$$\xi_k^* = \arg\min_{\xi \in \mathrm{T}_{y_k} \mathcal{M}} \langle \mathcal{T}_{R_{\eta_k}}^{-\sharp} (\mathrm{grad}\, f(x) + \tilde{L}\eta_k), \xi \rangle_x + \frac{\tilde{L}}{4} \|\xi\|_F^2 + g(y_k + \xi);$$
3. Find $\alpha > 0$ such that $\ell_x(\eta_k + \alpha \mathcal{T}_{R_{\eta_k}}^{-1} \xi_k^*) < \ell_x(\eta_k) - \sigma\alpha \|\xi_k^*\|_x^2$;
4. $\eta_{k+1} = \eta_k + \alpha \mathcal{T}_{R_{\eta_k}}^{-1} \xi_k^*$, $k \leftarrow k + 1$ and goto Step 1;

Above algorithm is used if the ambient space is $\mathbb{R}^n$
An application of [CMMCSZ20] if $R_x^{-1}(y)$ exists.

Numerical Experiments

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

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

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.

λ	Algo	iter	time	f	δ	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-Ada	1398	0.30	8.33_1	1.67_{-3}	0.54	0.86
	A-ManPG	273	0.09	8.33_1	9.19_{-4}	0.54	0.86
	A-RPG	263	0.06	8.33_1	1.12_{-3}	0.54	0.86

- ManPG: the method in [CMMCSZ20];
- RPG: the new Riemannian proximal gradient without acceleration;
- A-ManPG: Use similar technique to accelerate ManPG;
- A-RPG: the new Riemannian proximal gradient with acceleration;

Numerical Experiments

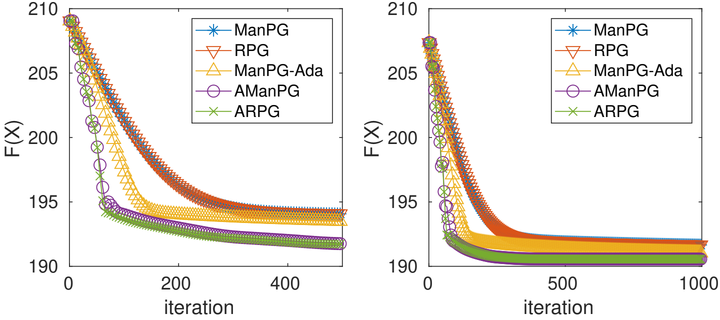


Figure: Two typical runs of ManPG, RPG, A-ManPG, and A-RPG for the Sparse PCA problem. $n = 1024$, $p = 4$, $\lambda = 2$, $m = 20$.

Numerical Experiments

Sparse PCA problem

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

where $A \in \mathbb{R}^{m \times n}$ is a data matrix.

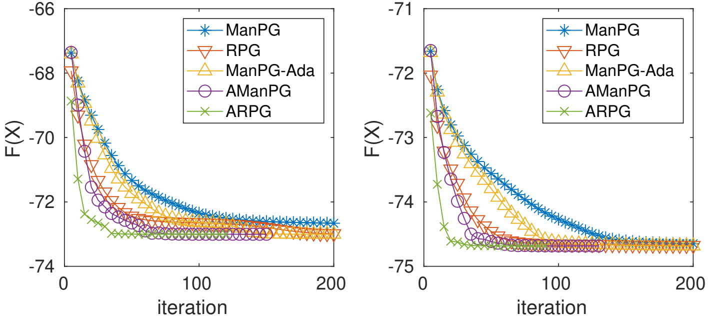


Figure: Two typical runs of ManPG, RPG, A-ManPG, and A-RPG for the Sparse PCA problem. $n = 1024$, $p = 4$, $\lambda = 2$, $m = 20$.

Summary

- Propose a Riemannian proximal gradient method;
- Global convergence to critical points
- $O(1/k)$ convergence rate using retraction-convexity
- Local convergence rate using Riemannian KL property
- Retraction of a semialgebraic functiono onto the Stiefel manifold satisfies the Riemannian KL property
- Apply the methods to sparse PCA problems on the oblique manifold and the Stiefel manifold;

- Wen Huang, Ke Wei, Riemannian Proximal Gradient Methods, arxiv:1909.06065, 2019

Thank you

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. 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.

A. Beck and M. Teboulle.
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM Journal on Imaging Sciences, 2(1):183–202, January 2009.
doi:10.1137/080716542.

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.

John Darzentas.
Problem Complexity and Method Efficiency in Optimization.
1983.

Octavian Eugen Ganea, Gary Becigneul, and Thomas Hofmann.
Hyperbolic entailment cones for learning hierarchical embeddings.
35th International Conference on Machine Learning, ICML 2018, 4:2661–2673, 2018.

References II

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.

Ian T. Jolliffe, Nickolay T. Trendafilov, and Mudassir Uddin.
A modified principal component technique based on the Lasso.
Journal of Computational and Graphical Statistics, 12(3):531–547, 2003.

Y. E. Nesterov.
A method for solving the convex programming problem with convergence rate $O(1/k^2)$.
Dokl. Akas. Nauk SSSR (In Russian), 269:543–547, 1983.

J. Shi and C. Qi.
Low-rank sparse representation for single image super-resolution via self-similarity learning.
In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1424–1428, Sep. 2016.

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.

Y. Zhang, Y. Lau, H.-W. Kuo, S. Cheung, A. Pasupathy, and J. Wright.
On the global geometry of sphere-constrained sparse blind deconvolution.
In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.