An Inexact Riemannian Proximal Gradient Method

Speaker: Wen Huang

Xiamen University

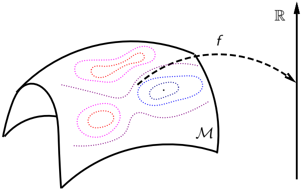
Nov. 13, 2021

Joint work with Ke Wei @Fudan University

Problem Statement

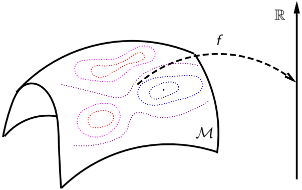
Optimization on Manifolds with Structure:

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



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

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 [ZHT06], discriminative k-means [YZW08], texture and imaging inpainting [LRZM12], co-sparse factor regression [MDC17], and low-rank sparse coding [ZGL$^+$13].

A Euclidean Proximal Gradient Method

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

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

1

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[1]:

$$\text{initial iterate:} x_0,$$
$$\begin{cases} d_k = \arg\min_{p \in \mathbb{R}^n} \langle \nabla f(x_k), p \rangle + \frac{L}{2}\|p\|_F^2 + g(x_k + p), & \text{(Proximal mapping)} \\ x_{k+1} = x_k + d_k. & \text{(Update iterates)} \end{cases}$$

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

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

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[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 ∇f;

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

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[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 ∇f;
- Proximal mapping: easy to compute;

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

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

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

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

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

A Euclidean Proximal Gradient Method

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 $\theta \in (0, 1]$:

$$\varsigma'(F(y) - F(x)) \operatorname{dist}(0, \partial F(y)) \geq 1, \quad \varsigma(t) = \frac{C}{\theta} e^{\theta}.$$

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 $C_2 > 0$;

Diffuclities 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 [CMSZ20]

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

A Riemannian Proximal Gradient Method in [CMSZ20]

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

A Riemannian Proximal Gradient Method in [CMSZ20]

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

A Riemannian Proximal Gradient Method in [CMSZ20]

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 for the Stiefel manifold by semi-smooth Newton;

A Riemannian Proximal Gradient Method in [CMSZ20]

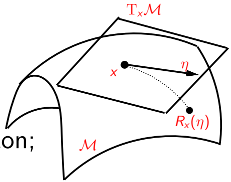
Euclidean proximal mapping

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

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 α_k;

- Only works for embedded submanifold;
- Proximal mapping is defined in tangent space;
- Convex programming;
- Solved for the Stiefel manifold by semi-smooth Newton;
- Convergence to a stationary point;



A Riemannian Proximal Gradient Method in [CMSZ20]

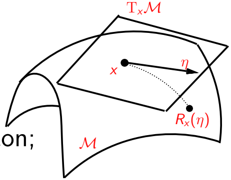
Euclidean proximal mapping

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

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 α_k;

- Only works for embedded submanifold;
- Proximal mapping is defined in tangent space;
- Convex programming;
- Solved for the Stiefel manifold by semi-smooth Newton;
- Convergence to a stationary point;
- No convergence rate results;



A Riemannian Proximal Gradient Method in [HW21]

ManPG [CMSZ20]

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

RPG [HW21]

Let $\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + g(R_{x_k}(\eta))$;

1. $\eta_k \in \mathrm{T}_{x_k} \mathcal{M}$ is a stationary point of $\ell_{x_k}(\eta)$, and $\ell_{x_k}(0) \geq \ell_k(\eta_k)$;
2. $x_{k+1} = R_{x_k}(\eta_k)$;

[1][HW21]: W. Huang, K. Wei, Riemannian Proximal Gradient Methods. Mathematical Programming, Series A, doi:10.1007/s10107-021-01632-3, 2021

A Riemannian Proximal Gradient Method in [HW21]

ManPG [CMSZ20]

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

RPG [HW21]

Let $\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + g(R_{x_k}(\eta))$;

1. $\eta_k \in \mathrm{T}_{x_k} \mathcal{M}$ is a stationary point of $\ell_{x_k}(\eta)$, and $\ell_{x_k}(0) \geq \ell_k(\eta_k)$;
2. $x_{k+1} = R_{x_k}(\eta_k)$;

- General framework for Riemannian optimization;

[1] [HW21]: W. Huang, K. Wei, Riemannian Proximal Gradient Methods. Mathematical Programming, Series A, doi:10.1007/s10107-021-01632-3, 2021

A Riemannian Proximal Gradient Method in [HW21]

ManPG [CMSZ20]

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

RPG [HW21]

Let $\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + g(R_{x_k}(\eta))$;

1. $\eta_k \in \mathrm{T}_{x_k} \mathcal{M}$ is a stationary point of $\ell_{x_k}(\eta)$, and $\ell_{x_k}(0) \geq \ell_k(\eta_k)$;
2. $x_{k+1} = R_{x_k}(\eta_k)$;

- General framework for Riemannian optimization;
- Any limit point is a critical point;

[1][HW21]: W. Huang, K. Wei, Riemannian Proximal Gradient Methods. Mathematical Programming, Series A, doi:10.1007/s10107-021-01632-3, 2021

A Riemannian Proximal Gradient Method in [HW21]

ManPG [CMSZ20]

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

RPG [HW21]

Let $\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + g(R_{x_k}(\eta))$;

1. $\eta_k \in \mathrm{T}_{x_k} \mathcal{M}$ is a stationary point of $\ell_{x_k}(\eta)$, and $\ell_{x_k}(0) \geq \ell_k(\eta_k)$;
2. $x_{k+1} = R_{x_k}(\eta_k)$;

- General framework for Riemannian optimization;
- Any limit point is a critical point;
- $O(1/k)$ sublinear convergence rate for retraction-convex f and g;

A Riemannian Proximal Gradient Method in [HW21]

ManPG [CMSZ20]

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

RPG [HW21]

Let $\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + g(R_{x_k}(\eta))$;

1. $\eta_k \in \mathrm{T}_{x_k} \mathcal{M}$ is a stationary point of $\ell_{x_k}(\eta)$, and $\ell_{x_k}(0) \geq \ell_k(\eta_k)$;
2. $x_{k+1} = R_{x_k}(\eta_k)$;

- General framework for Riemannian optimization;
- Any limit point is a critical point;
- $O(1/k)$ sublinear convergence rate for retraction-convex f and g;
- Local convergence rate by Riemannian KL property;

A Riemannian Proximal Gradient Method in [HW21]

ManPG [CMSZ20]

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

RPG [HW21]

Let $\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + g(R_{x_k}(\eta))$;

1. $\eta_k \in \mathrm{T}_{x_k} \mathcal{M}$ is a stationary point of $\ell_{x_k}(\eta)$, and $\ell_{x_k}(0) \geq \ell_k(\eta_k)$;
2. $x_{k+1} = R_{x_k}(\eta_k)$;

- General framework for Riemannian optimization;
- Any limit point is a critical point;
- $O(1/k)$ sublinear convergence rate for retraction-convex f and g;
- Local convergence rate by Riemannian KL property;
- Exploring manifold structure or using semi-smooth Newton iteratively;

Both ManPG and RPG require the Riemannian proximal mapping to be solved exactly

- Theoretically, but not practical numerically
- Can we relax this requirement and still preserve desired convergence properties?
- ManPG (no converge rate results)
- RPG (this talk)

An Inexact Riemannian Proximal Gradient Method

Outline:

- Algorithm statement
- Convergence analysis on general manifolds
- Algorithm design for the inexact Riemannian proximal mapping
- Numerical experiments

Outline:

- Algorithm statement
- Convergence analysis on general manifolds
- Algorithm design for the inexact Riemannian proximal mapping
- Numerical experiments

An Inexact Riemannian Proximal Gradient Method

Inexact RPG (IRPG)

Let $\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + g(R_{x_k}(\eta))$;

1. Find $\hat{\eta}_k \in \mathrm{T}_x \mathcal{M}$ such that

$$\|\hat{\eta}_{x_k} - \eta_{x_k}^*\| \leq q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) \text{ and } \ell_{x_k}(0) \geq \ell_{x_k}(\hat{\eta}_{x_k}),$$

   where $\varepsilon_k > 0$, and $q : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function;
2. $x_{k+1} = R_{x_k}(\eta_k)$;

An Inexact Riemannian Proximal Gradient Method

Inexact RPG (IRPG)

Let $\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + g(R_{x_k}(\eta))$;

1. Find $\hat{\eta}_k \in \mathrm{T}_x \mathcal{M}$ such that

$$\|\hat{\eta}_{x_k} - \eta_{x_k}^*\| \leq q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) \text{ and } \ell_{x_k}(0) \geq \ell_{x_k}(\hat{\eta}_{x_k}),$$

   where $\varepsilon_k > 0$, and $q : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function;

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

Four choices of q lead to different convergence results:

1) Global $q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) = \varepsilon_k$ with $\varepsilon_k \to 0$;
2) Global $q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) = \tilde{q}(\|\hat{\eta}_{x_k}\|)$ with $\tilde{q} : \mathbb{R} \to [0, \infty)$ a continuous function satisfying $\tilde{q}(0) = 0$;
3) Unique $q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) = \varepsilon_k^2$, with $\sum_{k=0}^{\infty} \varepsilon_k < \infty$; and
4) Rate $q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) = \min(\varepsilon_k^2, \delta_q \|\hat{\eta}_{x_k}\|^2)$ with a constant $\delta_q > 0$ and $\sum_{k=0}^{\infty} \varepsilon_k < \infty$.

Inexact RPG (IRPG)

Let $\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2}\|\eta\|_{x_k}^2 + g(R_{x_k}(\eta))$;

1. Find $\hat{\eta}_k \in \mathrm{T}_x \mathcal{M}$ such that
$$\|\hat{\eta}_{x_k} - \eta^*_{x_k}\| \leq q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) \text{ and } \ell_{x_k}(0) \geq \ell_{x_k}(\hat{\eta}_{x_k}),$$
where $\varepsilon_k > 0$, and $q : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function;
2. $x_{k+1} = R_{x_k}(\eta_k)$;

Not a Riemannian generalization of any of the existing Euclidean inexact proximal gradient methods

An Inexact Riemannian Proximal Gradient Method

Inexact proximal gradient methods in the Euclidean setting:
[Com04, FP11, SRB11, VSBV13, BPR20]

[Com04]: Patrick L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators.Optimization, 53(5-6):475–504, 2004.
[FP11]: J. M. Fadili, and G. Peyre, Total variation projection with first order schemes. IEEE Transactions on Image Processing, 20(3), 657-669, 2001.
[SRB11]: M. Schmidt, N. Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. NIPS, 2001.
[VSBV13]: S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. SIAM Journal on Optimization, 23(3),1607-1633, 2013
[BPR20]: S. Bonettini, M. Prato, and S. Rebegoldi. Convergence of inexact forward–backward algorithms using the forward–backward envelope. SIAM Journal on Optimization, 30(4), 3069-3097, 2020

An Inexact Riemannian Proximal Gradient Method

Inexact proximal gradient methods in the Euclidean setting: [Com04, FP11, SRB11, VSBV13, BPR20]

- $z = \mathrm{Prox}_{\lambda g}(y) = \mathrm{argmin}_x \Phi_\lambda(x) := \lambda g(x) + \frac{1}{2}\|x - y\|^2$;

An Inexact Riemannian Proximal Gradient Method

Inexact proximal gradient methods in the Euclidean setting: [Com04, FP11, SRB11, VSBV13, BPR20]

- $z = \mathrm{Prox}_{\lambda g}(y) = \mathrm{argmin}_x \Phi_\lambda(x) := \lambda g(x) + \frac{1}{2}\|x - y\|^2$;
- z satisfies

$$(y - z)/\lambda \in \partial^E g(z) \text{ and } \mathrm{dist}(0, \partial^E \Phi_\lambda(z)) = 0.$$

An Inexact Riemannian Proximal Gradient Method

Inexact proximal gradient methods in the Euclidean setting: [Com04, FP11, SRB11, VSBV13, BPR20]

- $z = \mathrm{Prox}_{\lambda g}(y) = \mathrm{argmin}_x \Phi_\lambda(x) := \lambda g(x) + \frac{1}{2}\|x - y\|^2$;
- z satisfies

$$(y - z)/\lambda \in \partial^E g(z) \text{ and } \mathrm{dist}(0, \partial^E \Phi_\lambda(z)) = 0.$$

- Approximation $\hat{z}$ satisfies any one of the following conditions:

$$\mathrm{dist}(0, \partial^E \Phi_\lambda(\hat{z})) \leq \frac{\varepsilon}{\lambda}, \quad \Phi_\lambda(\hat{z}) \leq \min \Phi_\lambda + \frac{\varepsilon^2}{2\lambda}, \text{ and } \frac{y - \hat{z}}{\lambda} \in \partial^E_{\frac{\varepsilon^2}{2\lambda}} g(\hat{z}),$$

An Inexact Riemannian Proximal Gradient Method

Inexact proximal gradient methods in the Euclidean setting: [Com04, FP11, SRB11, VSBV13, BPR20]

- $z = \mathrm{Prox}_{\lambda g}(y) = \mathrm{argmin}_x \Phi_\lambda(x) := \lambda g(x) + \frac{1}{2}\|x - y\|^2$;
- z satisfies

$$(y - z)/\lambda \in \partial^E g(z) \text{ and } \mathrm{dist}(0, \partial^E \Phi_\lambda(z)) = 0.$$

- Approximation $\hat{z}$ satisfies any one of the following conditions:

$$\mathrm{dist}(0, \partial^E \Phi_\lambda(\hat{z})) \leq \frac{\varepsilon}{\lambda}, \quad \Phi_\lambda(\hat{z}) \leq \min \Phi_\lambda + \frac{\varepsilon^2}{2\lambda}, \text{ and } \frac{y - \hat{z}}{\lambda} \in \partial^E_{\frac{\varepsilon^2}{2\lambda}} g(\hat{z}),$$

- Algorithms based on strong convexity of the Euclidean proximal mapping

An Inexact Riemannian Proximal Gradient Method

Inexact proximal gradient methods in the Euclidean setting:
[Com04, FP11, SRB11, VSBV13, BPR20]

- $z = \mathrm{Prox}_{\lambda g}(y) = \mathrm{argmin}_x \Phi_\lambda(x) := \lambda g(x) + \frac{1}{2}\|x - y\|^2$;
- z satisfies

$$(y - z)/\lambda \in \partial^E g(z) \text{ and } \mathrm{dist}(0, \partial^E \Phi_\lambda(z)) = 0.$$

- Approximation $\hat{z}$ satisfies any one of the following conditions:

$$\mathrm{dist}(0, \partial^E \Phi_\lambda(\hat{z})) \leq \frac{\varepsilon}{\lambda}, \quad \Phi_\lambda(\hat{z}) \leq \min \Phi_\lambda + \frac{\varepsilon^2}{2\lambda}, \text{ and } \frac{y - \hat{z}}{\lambda} \in \partial^E_{\frac{\varepsilon^2}{2\lambda}} g(\hat{z}),$$

- Algorithms based on strong convexity of the Euclidean proximal mapping
- Riemannian: may not be convex

$$\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2}\|\eta\|^2_{x_k} + g(R_{x_k}(\eta))$$

Outline:

- Algorithm statement
- Convergence analysis on general manifolds
- Algorithm design for the inexact Riemannian proximal mapping
- Numerical experiments

Assumptions and Global 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 Global 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 Global 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 Global 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:

- Suppose $\lim_{k\to\infty} q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) = 0$, then for any accumulation point x_* of $\{x_k\}$, x_* is a stationary point, i.e., $0 \in \partial F(x_*)$.

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 is locally Lipschitz continuous with respect to the retraction R

Definition

A function $h : \mathcal{M} \rightarrow \mathbb{R}$ is called locally Lipschitz continuous with respect to a retraction R if for any compact subset $\mathcal{N}$ of $\mathcal{M}$, there exists a constant L_h such that for any $x \in \mathcal{N}$ and $\xi_x, \eta_x \in \mathrm{T}_x \mathcal{M}$ satisfying $R_x(\xi_x) \in \mathcal{N}$ and $R_x(\eta_x) \in \mathcal{N}$, it holds that $|h \circ R(\xi_x) - h \circ R(\eta_x)| \leq L_h \|\xi_x - \eta_x\|$.

Assumptions and Local Convergence Result

Assumption:

1. Assumptions for the global convergence
2. f is locally Lipschitz continuously differentiable
3. F is locally Lipschitz continuous with respect to the retraction R

If the manifold $\mathcal{M}$ is an embedded submanifold and function F is locally Lipschitz in the embedding space, then the function is locally Lipschitz continuous with respect to any global defined retraction R.

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

Assumptions and Local Convergence Result

Assumption:

1. Assumptions for the global convergence
2. f is locally Lipschitz continuously differentiable
3. F is locally Lipschitz continuous with respect to the retraction R
4. F satisfies the Riemannian KL property

Definition ([BdCNO11])

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 is locally Lipschitz continuous with respect to the retraction R
4. F satisfies the Riemannian KL property

Theoretical results:

- If $\|\hat{\eta}_{x_k} - \eta^*_{x_k}\| \leq \varepsilon_k^2$ for $\sum_{k=0}^{\infty} \varepsilon_k < \infty$ and $\varepsilon_k > 0$, then 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 is locally Lipschitz continuous with respect to the retraction R
4. F satisfies the Riemannian KL property

Theoretical results:

- If $\|\hat{\eta}_{x_k} - \eta^*_{x_k}\| \leq \min\left(\varepsilon_k^2, \frac{\beta}{2L_F}\|\hat{\eta}_{x_k}\|^2\right)$ for $\sum_{k=0}^{\infty} \varepsilon_k < \infty$ and $\varepsilon_k > 0$, and 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 $C_2 > 0$;

Algorithms for the Riemannian Proximal Mapping

Outline:

- Algorithm statement
- Convergence analysis on general manifolds
- Algorithm design for the inexact Riemannian proximal mapping
- Numerical experiments

Algorithms for the Riemannian Proximal Mapping

Assumptions:

- The manifold $\mathcal{M}$ has a linear ambient space
- The function g is convex and Lipschitz continuous, where the convexity and Lipschitz continuity are in the Euclidean sense.

ManPG [CMSZ20]

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

IRPG

Let $\ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + g(R_{x_k}(\eta))$;

1. Find $\hat{\eta}_k \in \mathrm{T}_x \mathcal{M}$ such that

$$\|\hat{\eta}_{x_k} - \eta_{x_k}^*\| \leq q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) \text{ and } \ell_{x_k}(0) \geq \ell_{x_k}(\hat{\eta}_{x_k}),$$

where $\varepsilon_k > 0$, and $q : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function;

ManPG can be viewed as an IRPG.

ManPG [CMSZ20]

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

Above problem can be rewritten as

$$\arg \min_{B_x^T \eta = 0} \langle \xi_x, \eta \rangle + \frac{1}{2\mu} \|\eta\|_F^2 + g(x + \eta)$$

where $B_x^T \eta = (\langle b_1, \eta \rangle, \langle b_2, \eta \rangle, \dots, \langle b_m, \eta \rangle)^T$, and $\{b_1, \dots, b_m\}$ forms an orthonormal basis of $\mathrm{N}_x \mathcal{M}$.

The Lagrangian function:

$$\mathcal{L}(\eta, \Lambda) = \langle \xi_x, \eta \rangle + \frac{1}{2\mu} \langle \eta, \eta \rangle + g(X + \eta) - \langle \Lambda, B_x^T \eta \rangle.$$

Therefore

$$\text{KKT: } \left\{ \begin{array}{c} \partial_\eta \mathcal{L}(\eta, \Lambda) = 0 \\ B_x^T \eta = 0 \end{array} \right. \Longrightarrow \left\{ \begin{array}{c} \eta = \mathrm{Prox}_{\mu g} \left(x - \mu(\xi_x - B_x \Lambda)\right) - x \\ B_x^T \eta = 0 \end{array} \right.$$

where $\mathrm{Prox}_{\mu g}(z) = \mathrm{argmin}_{v \in \mathbb{R}^{n \times p}} \frac{1}{2} \|v - z\|_F^2 + \mu g(v)$.

Semi-smooth Newton method finds the Λ such that

$$\Psi(\Lambda) := B_x^T(\mathrm{Prox}_{\mu g}(x - \mu(\xi_x - B_x\Lambda)) - x) = 0$$
$$\eta_* = \mathrm{Prox}_{\mu g}(x - \mu(\xi_x - B_x\Lambda)) - x$$

- Ψ is not differentiable everywhere but semi-smooth for $g(\cdot) = \|\cdot\|_1$;
- Semi-smooth Newton:
  1. $J_\Psi(\Lambda_k)[d] = -\Psi(\Lambda_k)$, where J_Ψ is the generalized Jacobian of Ψ;
  2. $\Lambda_{k+1} = \Lambda_k + d_k$

Semi-smooth Newton method finds the Λ such that

$$\Psi(\Lambda) := B_x^T (\mathrm{Prox}_{\mu g}(x - \mu(\xi_x - B_x \Lambda)) - x) \approx 0$$

- Ψ is not differentiable everywhere but semi-smooth for $g(\cdot) = \|\cdot\|_1$;
- Semi-smooth Newton:
  1. $J_\Psi(\Lambda_k)[d] = -\Psi(\Lambda_k)$, where J_Ψ is the generalized Jacobian of Ψ;
  2. $\Lambda_{k+1} = \Lambda_k + d_k$
- Solving the equation inexactly

If $\Psi(\Lambda) = \epsilon$,

- $\eta_* = \mathrm{Prox}_{\mu g}\left(x - \mu(\xi_x - B_x \Lambda)\right) - x$ is not even in the tangent space $\mathrm{T}_x \mathcal{M}$ in this case
- Use $\hat{\eta}_x := \hat{v}(\Lambda) = P_{\mathrm{T}_x \mathcal{M}}(\mathrm{Prox}_{\mu g}\left(x - \mu(\xi_x - B_x \Lambda)\right) - x)$ instead
- How small does ϵ need to be?

If $\Psi(\Lambda) = \epsilon$,

- $\eta_* = \mathrm{Prox}_{\mu g}(x - \mu(\xi_x - B_x \Lambda)) - x$ is not even in the tangent space $\mathrm{T}_x \mathcal{M}$ in this case
- Use $\hat{\eta}_x := \hat{v}(\Lambda) = P_{\mathrm{T}_x \mathcal{M}}(\mathrm{Prox}_{\mu g}(x - \mu(\xi_x - B_x \Lambda)) - x)$ instead
- How small does ϵ need to be?

$$\|\epsilon\| \leq \min(\phi(\hat{v}(\Lambda)), 0.5),$$

with $\phi(0) = 0$ and ϕ is nondecreasing.

The function q is:

$$q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) =$$

$$\frac{2L_g \varkappa_2}{\tilde{L} - 2L_g \varkappa_2} \|\hat{\eta}_{x_k}\| + \sqrt{\frac{4L_g \varkappa_2 - 4L_g^2 \varkappa_2^2}{(\tilde{L} - 2L_g \varkappa)^2} \|\hat{\eta}_{x_k}\|^2 + \frac{4\vartheta}{\tilde{L} - 2L_g \varkappa_2} \min(\phi(\|\hat{\eta}_{x_k}\|), 0.5)}$$

- ManPG can be viewed as an inexact RPG for sufficiently large $\tilde{L}$;

The function q is:

$$q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) =$$
$$\frac{2L_g \varkappa_2}{\tilde{L} - 2L_g \varkappa_2} \|\hat{\eta}_{x_k}\| + \sqrt{\frac{4L_g \varkappa_2 - 4L_g^2 \varkappa_2^2}{(\tilde{L} - 2L_g \varkappa)^2} \|\hat{\eta}_{x_k}\|^2 + \frac{4\vartheta}{\tilde{L} - 2L_g \varkappa_2} \min(\phi(\|\hat{\eta}_{x_k}\|), 0.5)}$$

- ManPG can be viewed as an inexact RPG for sufficiently large $\tilde{L}$;
- This q may not guarantee local convergence results;

The function q is:

$$q(\varepsilon_k, \|\hat{\eta}_{x_k}\|) =$$

$$\frac{2L_g \varkappa_2}{\tilde{L} - 2L_g \varkappa_2} \|\hat{\eta}_{x_k}\| + \sqrt{\frac{4L_g \varkappa_2 - 4L_g^2 \varkappa_2^2}{(\tilde{L} - 2L_g \varkappa)^2} \|\hat{\eta}_{x_k}\|^2 + \frac{4\vartheta}{\tilde{L} - 2L_g \varkappa_2} \min(\phi(\|\hat{\eta}_{x_k}\|), 0.5)}$$

- ManPG can be viewed as an inexact RPG for sufficiently large $\tilde{L}$;
- This q may not guarantee local convergence results;
- Improving accuracy is needed;

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

Solving the Riemannian Proximal Mapping [HW21]

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. $\eta_{k+1} = \eta_k + \alpha \mathcal{T}_{R_{\eta_k}}^{-1} \xi_k^*$;
5. If $\xi_k^* = 0$, then stop;
6. $k \leftarrow k + 1$ and goto Step 1;

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

Solving the Riemannian Proximal Mapping [HW21]

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
   $$\xi_k^* \approx \arg\min_{\xi \in \mathrm{T}_{y_k} \mathcal{M}} \langle \mathcal{T}_{R_{\eta_k}}^{-\sharp}(\mathrm{grad}\, f(x) + \tilde{L}\eta_k), \xi \rangle_x + \frac{\tilde{L}}{4} \|\xi\|_F^2 + g(y_k + \xi);$$
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. $\eta_{k+1} = \eta_k + \alpha \mathcal{T}_{R_{\eta_k}}^{-1} \xi_k^*$;
5. If $\|\xi_k^*\|$ is sufficiently small, then stop;
6. $k \leftarrow k + 1$ and goto Step 1;

Solving the Riemannian Proximal Mapping [HW21]

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
$$\xi_k^* \approx \arg \min_{\xi \in \mathrm{T}_{y_k} \mathcal{M}} \langle \mathcal{T}_{R_{\eta_k}}^{-\sharp}(\mathrm{grad} f(x) + \tilde{L}\eta_k), \xi \rangle_x + \frac{\tilde{L}}{4}\|\xi\|_F^2 + g(y_k + \xi);$$
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. $\eta_{k+1} = \eta_k + \alpha \mathcal{T}_{R_{\eta_k}}^{-1} \xi_k^*$;
5. If $\|\xi_k^*\|$ is sufficiently small, then stop;
6. $k \leftarrow k + 1$ and goto Step 1;

- Same as the subproblem in ManPG;
- The same inexact technique can be used;

Solving the Riemannian Proximal Mapping [HW21]

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
$$\xi_k^* \approx \arg \min_{\xi \in \mathrm{T}_{y_k} \mathcal{M}} \langle \mathcal{T}_{R_{\eta_k}}^{-\sharp}(\mathrm{grad}\, f(x) + \tilde{L}\eta_k), \xi \rangle_x + \frac{\tilde{L}}{4} \|\xi\|_F^2 + g(y_k + \xi);$$
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. $\eta_{k+1} = \eta_k + \alpha \mathcal{T}_{R_{\eta_k}}^{-1} \xi_k^*$;
5. If $\|\xi_k^*\| < \psi(\varepsilon_k, \varrho, \|\eta_k\|)$ is sufficiently small, then stop;
6. $k \leftarrow k + 1$ and goto Step 1;

Suppose an error bound property holds for $\ell_x(\eta)$. Then

- $\psi = \varepsilon_k^2 \implies \|\hat{\eta}_{x_k} - \eta_{x_k}^*\| \leq C \varepsilon_k^2$;
- $\psi = \min(\varepsilon_k^2, \varrho \|\hat{\eta}_{x_k}\|^2) \implies \|\hat{\eta}_{x_k} - \eta_{x_k}^*\| \leq C \min(\varepsilon_k^2, \varrho \|\hat{\eta}_{x_k}\|^2)$;

Retraction-convexity of g implies the error bound property.

Outline:

- Algorithm statement
- Convergence analysis on general manifolds
- Algorithm design for the inexact Riemannian proximal mapping
- Numerical experiments

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.

Numerical Experiments

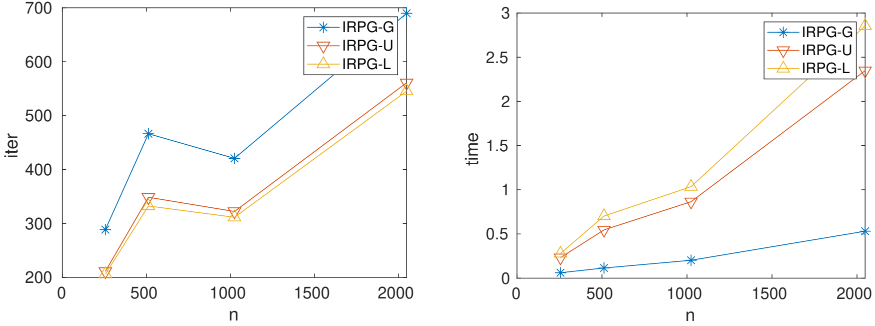


Figure: Average of 10 random runs, $p = 4$, $m = 20$, $\lambda = 2$;

- IRPG-G: an inexact version of ManPG
- IRPG-U: $\psi = \varepsilon_k^2$
- IRPG-L: $\psi = \min(\varepsilon_k^2, \varrho \|\hat{\eta}_{x_k}\|^2)$

Summary

- Review the two existing Riemannian proximal gradient methods
- Propose an inexact Riemannian proximal gradient methods
- Convergence analysis for general manifolds
- Semi-smooth Newton method for inexact Riemannian proximal mapping to guarantee global convergence
- Further improving accuracy by an iterative algorithm, accuracy is guaranteed based on error bound property.

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.

S. Bonettini, M. Prato, and S. Rebegoldi.
Convergence of inexact forward–backward algorithms using the forward–backward envelope.
SIAM Journal on Optimization, 30(4):3069–3097, 2020.

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.

Patrick L. Combettes.
Solving monotone inclusions via compositions of nonexpansive averaged operators.
Optimization, 53(5-6):475–504, 2004.

Jalal M. Fadili and Gabriel Peyré.
Total variation projection with first order schemes.
IEEE Transactions on Image Processing, 20(3):657–669, 2011.

References II

W. Huang and K. Wei.
Riemannian proximal gradient methods.
Mathematical Programming, 2021.
published online, DOI:10.1007/s10107-021-01632-3.

Xiao Liang, Xiang Ren, Zhengdong Zhang, and Yi Ma.
Repairing sparse low-rank texture.
In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 482–495, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

A. Mishra, Dipak K Dey, and K. Chen.
Sequential Co-Sparse Factor Regression.
Journal of Computational and Graphical Statistics, 26(4):814–825, 2017.

Mark Schmidt, Nicolas Roux, and Francis Bach.
Convergence rates of inexact proximal-gradient methods for convex optimization.
In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Silvia Villa, Saverio Salzo, Luca Baldassarre, and Alessandro Verri.
Accelerated and inexact forward-backward algorithms.
SIAM Journal on Optimization, 23(3):1607–1633, 2013.

Jieping Ye, Zheng Zhao, and Mingrui Wu.
Discriminative k-means for clustering.
In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja.
Low-rank sparse coding for image classification.
In *2013 IEEE International Conference on Computer Vision*, pages 281–288, 2013.

References III

Hui Zou, Trevor Hastie, and Robert Tibshirani.
Sparse principal component analysis.
Journal of Computational and Graphical Statistics, 15(2):265–286, 2006.

Thank you

Thank you!