A Riemannian Proximal Newton Method

Speaker: Wen Huang

Xiamen University

March 19, 2023

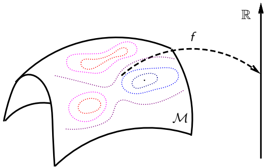
Joint work with Wutao Si, P.-A. Absil, Rujun Jiang, Simon Vary

Fuzhou University

Problem Statement

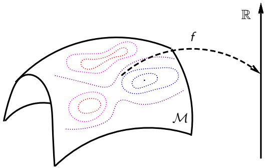
Optimization on Manifolds with Structure:

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



- $\mathcal{M}$ is a finite-dimensional Riemannian manifold;
- f is smooth and may be nonconvex; and
- $h(x)$ is continuous and convex but may be nonsmooth;

Problem Statement



Optimization on Manifolds with Structure:

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

- $\mathcal{M}$ is a finite-dimensional Riemannian manifold;
- f is smooth and may be nonconvex; and
- $h(x)$ is continuous and convex but may be nonsmooth;

Applications: sparse PCA [ZHT06], compressed model [OLCO13], sparse partial least squares regression [CSG$^+$18], sparse inverse covariance estimation [BESS19], sparse blind deconvolution [ZLK$^+$17], and clustering [HWGVD22].

Outline

- Euclidean proximal gradient method and its variants;
- Riemannian proximal gradient method and its variants;
- A Riemannian proximal Newton method;
- Numerical experiments;

Euclidean Proximal Gradient Method and its variants

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

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

Euclidean Proximal Gradient Method and its variants

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

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

- Proximal Gradient
- Accelerated versions
- Proximal inexact Newton
- Proximal quasi-Newton

Euclidean Proximal Gradient Method and its variants

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

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

Given x_0[1],

$$\begin{cases} d_k = \arg\min_p \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_{\mathrm{F}}^2 + h(x_k + p) \\ x_{k+1} = x_k + d_k. \end{cases}$$

- Proximal Gradient
- Accelerated versions
- Proximal inexact Newton
- Proximal quasi-Newton

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

Euclidean Proximal Gradient Method and its variants

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

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

- Proximal Gradient
- Accelerated versions
- Proximal inexact Newton
- Proximal quasi-Newton

Given x_0,

$$\begin{cases} d_k = \arg\min_p \langle \nabla f(x_k), p \rangle + \frac{L}{2} \|p\|_{\mathrm{F}}^2 + h(x_k + p) \\ x_{k+1} = x_k + d_k. \end{cases}$$

- $h = 0$: reduce to steepest descent method;
- Any limit point is a critical point;
- $O\left(\frac{1}{k}\right)$ sublinear convergence rate for convex f and h;
- Linear convergence rate for strongly convex f and convex h;
- Local convergence rate by KL property;

Euclidean Proximal Gradient Method and its variants

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

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

- Proximal Gradient
- Accelerated versions
- Proximal inexact Newton
- Proximal quasi-Newton

Given x_0, let $y_0 = x_0, t_0 = 1$;

$$\begin{cases} d_{y_k} = \mathrm{argmin}_p \langle \nabla f(y_k), p \rangle + \frac{L}{2} \|p\|_{\mathrm{F}}^2 + h(y_k + p) \\ x_{k+1} = y_k + d_{y_k} \\ t_{k+1} = \frac{\sqrt{4t_k^2+1}+1}{2} \\ y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}} (x_{k+1} - x_k). \end{cases}$$

Euclidean Proximal Gradient Method and its variants

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

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

- Proximal Gradient
- Accelerated versions
- Proximal inexact Newton
- Proximal quasi-Newton

Given x_0, let $y_0 = x_0, t_0 = 1$;

$$\begin{cases} d_{y_k} = \mathrm{argmin}_p \langle \nabla f(y_k), p \rangle + \frac{L}{2} \|p\|_\mathrm{F}^2 + h(y_k + p) \\ x_{k+1} = y_k + d_{y_k} \\ t_{k+1} = \frac{\sqrt{4t_k^2+1}+1}{2} \\ y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}} (x_{k+1} - x_k). \end{cases}$$

- A representative one: FISTA [BT09];
- Based on the Nesterov momentum technique;
- $O\left(\frac{1}{k^2}\right)$ sublinear convergence rate for convex f and h;

[BT09] 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.

Euclidean Proximal Gradient Method and its variants

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

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

- Proximal Gradient
- Accelerated versions
- Proximal inexact Newton
- Proximal quasi-Newton

Given x_0;

$$\begin{cases} d_k = \mathrm{argmin}_p \langle \nabla f(x_k), p \rangle + \frac{1}{2}\langle p, H_k p \rangle + h(x_k + p) \\ x_{k+1} = x_k + t_k d_k, \text{ for a step size } t_k \end{cases}$$

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

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

- Proximal Gradient
- Accelerated versions
- Proximal inexact Newton
- Proximal quasi-Newton

Given x_0;

$$\begin{cases} d_k = \mathrm{argmin}_p \langle \nabla f(x_k), p \rangle + \frac{1}{2} \langle p, H_k p \rangle + h(x_k + p) \\ x_{k+1} = x_k + t_k d_k, \text{ for a step size } t_k \end{cases}$$

- H_k is Hessian or a positive definite approximation to Hessian [LSS14, MYZZ22];
- t_k is one for sufficiently large k;
- Quadratic/Superlinear convergence rate for strongly convex f and convex h;

[LLS14] Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal newton-type methods for minimizing composite functions. SIAM Journal on Optimization, 24(3):1420-1443, 2014.

[MYZZ22] Boris S Mordukhovich, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A globally convergent proximal newton-type method in nonsmooth convex optimization. Mathematical Programming, pages 1-38, 2022.

Euclidean Proximal Gradient Method and its variants

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

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

- Proximal Gradient
- Accelerated versions
- Proximal inexact Newton
- Proximal quasi-Newton

Given x_0, H_0;

$$\begin{cases} d_k = \mathrm{argmin}_p \langle \nabla f(x_k), p \rangle + \frac{1}{2} \langle p, H_k p \rangle + h(x_k + p) \\ x_{k+1} = x_k + t_k d_k, \text{ for a step size } t_k \\ \text{Update } H_k \text{ by a quasi-Newton formula} \end{cases}$$

Euclidean Proximal Gradient Method and its variants

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

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

- Proximal Gradient
- Accelerated versions
- Proximal inexact Newton
- Proximal quasi-Newton

Given x_0, H_0;

$$\begin{cases} d_k = \mathrm{argmin}_p \langle \nabla f(x_k), p \rangle + \frac{1}{2}\langle p, H_k p\rangle + h(x_k + p) \\ x_{k+1} = x_k + t_k d_k, \text{ for a step size } t_k \\ \text{Update } H_k \text{ by a quasi-Newton formula} \end{cases}$$

- Dennis-Moré condition $\Longrightarrow$ superlinear convergence rate for strongly convex f and convex h [LSS14];
- Sublinear without the accuracy assumption on H_k [ST16];

[LLS14] Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal newton-type methods for minimizing composite functions. SIAM Journal on Optimization, 24(3):1420-1443, 2014.

[ST16] K. Scheinberg and X. Tang. Practical inexact proximal quasi-Newton method with global complexity analysis. Mathematical Programming, (160):495-529, 2016.

Outline

- Euclidean proximal gradient method and its variants;
- Riemannian proximal gradient method and its variants;
- A Riemannian proximal Newton method;
- Numerical experiments;

Optimization with Structure:

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

Riemannian proximal gradient method and its variants

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Accelerated versions

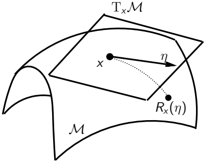
Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Accelerated versions

[CMSZ20]: Given x_0,

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

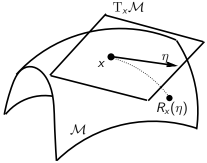


[CMSZ20] S. Chen, S. Ma, A. Man-Cho So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. SIAM Journal on Optimization, 30(1):210-239, 2020.

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Accelerated versions



[CMSZ20]: Given x_0,

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

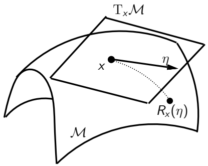
- Direction in the tangent space;
- Ambient space must be linear;
- Solved by a semismooth Newton method;

[CMSZ20] S. Chen, S. Ma, A. Man-Cho So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. SIAM Journal on Optimization, 30(1):210-239, 2020.

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Accelerated versions



[CMSZ20]: Given x_0,

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

- Direction in the tangent space;
- Ambient space must be linear;
- Solved by a semismooth Newton method;
- Any limit point is a critical point [CMSZ20, HW21b];
- No local convergence rate results;

[CMSZ20] S. Chen, S. Ma, A. Man-Cho So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. SIAM Journal on Optimization, 30(1):210-239, 2020.

[HW21b] W. Huang and K. Wei. An extension of fast iterative shrinkage-thresholding algorithm to Riemannian optimization for sparse principal component analysis. Numerical Linear Algebra with Applications, page e2409, 2021.

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Accelerated versions

[HW21a]: Given x_0,

$$\begin{cases} \text{Let } \ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + h(R_{x_k}(\eta)); \\ \eta_k \text{ is a stationary point of } \ell_{x_k} \text{ and } \ell_{x_k}(0) \geq \ell_k(\eta_k); \\ x_{k+1} = R_{x_k}(\eta_k); \end{cases}$$

[HW21a] W. Huang and K. Wei. Riemannian proximal gradient metho ds. Mathematical Programming, 2021. published online, DOI:10.1007/s10107-021-01632-3.

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Accelerated versions

[HW21a]: Given x_0,

$$\begin{cases} \text{Let } \ell_{x_k}(\eta) = \langle \mathrm{grad} f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2 + h(R_{x_k}(\eta)); \\ \eta_k \text{ is a stationary point of } \ell_{x_k} \text{ and } \ell_{x_k}(0) \geq \ell_k(\eta_k); \\ x_{k+1} = R_{x_k}(\eta_k); \end{cases}$$

- Direction in the tangent space;
- Well-defined for general manifold;
- Subproblem is difficult in general (simple for sphere);
- Any limit point is a critical point;
- $O\left(\frac{1}{k}\right)$ rate for retraction convex f and h;
- Local convergence rate by Riemannian KL property;

[HW21a] W. Huang and K. Wei. Riemannian proximal gradient metho ds. Mathematical Programming, 2021. published online, DOI:10.1007/s10107-021-01632-3.

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Accelerated versions

[HW21a]: Given x_0,

$$\begin{cases} \eta_{y_k} = \mathrm{argmin}_{\eta \in \mathrm{T}_{y_k} \mathcal{M}} \langle \mathrm{grad}\, f(y_k), \eta \rangle + \frac{L}{2} \|\eta\|_F^2 + h(y_k + \eta) \\ x_{k+1} = R_{y_k}(\eta_{y_k}) \\ t_{k+1} = \frac{\sqrt{4t_k^2+1}+1}{2} \\ y_{k+1} = R_{x_{k+1}} \left(\frac{1-t_k}{t_{k+1}} R_{x_{k+1}}^{-1}(x_k) \right) \end{cases}$$

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Accelerated versions

[HW21a]: Given x_0,

$$\begin{cases} \eta_{y_k} = \mathrm{argmin}_{\eta \in \mathrm{T}_{y_k} \mathcal{M}} \langle \mathrm{grad}\, f(y_k), \eta \rangle + \frac{L}{2} \|\eta\|_F^2 + h(y_k + \eta) \\ x_{k+1} = R_{y_k}(\eta_{y_k}) \\ t_{k+1} = \frac{\sqrt{4t_k^2+1}+1}{2} \\ y_{k+1} = R_{x_{k+1}} \left(\frac{1-t_k}{t_{k+1}} R_{x_{k+1}}^{-1}(x_k) \right) \end{cases}$$

- A representative on in [HW21b], also see [HW21a];
- Observe acceleration empirically;
- No $O(\frac{1}{k^2})$ convergence rate results;

Optimization with Structure:

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

No proximal Newton or quasi-Newton methods
on Riemannian manifold

Optimization with Structure:

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

No proximal Newton or quasi-Newton methods
on Riemannian manifold

Task: Develop a Riemannian proximal Newton method
that has superlinear local convergence rate

Outline

- Euclidean proximal gradient method and its variants;
- Riemannian proximal gradient method and its variants;
- A Riemannian proximal Newton method;
- Numerical experiments;

Outline

- Euclidean proximal gradient method and its variants;
- Riemannian proximal gradient method and its variants;
- A Riemannian proximal Newton method;
- Numerical experiments;

Note that we focus on:

- $\mathcal{M}$ is an Riemannian embedded submanifold of a Euclidean space;
- $h(x) = \mu\|x\|_1$;

Euclidean version:

$$\begin{cases} d_k = \mathrm{argmin}_p \langle \nabla f(x_k), p\rangle + \frac{1}{2}\langle p, \nabla^2 f(x_k) p\rangle + h(x_k + p) \\ x_{k+1} = x_k + d_k \end{cases}$$

A native generalization by replacing the Euclidean gradient and Hessian by the Riemannian gradient and Hessian:

$$\begin{cases} \eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k}\mathcal{M}} \langle \mathrm{grad}\, f(x_k), \eta\rangle + \frac{1}{2}\langle \eta, \mathrm{Hess}\, f(x_k)\eta\rangle + h(x_k + \eta) \\ x_{k+1} = R_{x_k}(\eta_k) \end{cases}$$

Euclidean version:

$$\begin{cases} d_k = \operatorname{argmin}_p \langle \nabla f(x_k), p \rangle + \frac{1}{2} \langle p, \nabla^2 f(x_k) p \rangle + h(x_k + p) \\ x_{k+1} = x_k + d_k \end{cases}$$

A native generalization by replacing the Euclidean gradient and Hessian by the Riemannian gradient and Hessian:

$$\begin{cases} \eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \mathrm{grad}\, f(x_k), \eta \rangle + \frac{1}{2} \langle \eta, \mathrm{Hess}\, f(x_k) \eta \rangle + h(x_k + \eta) \\ x_{k+1} = R_{x_k}(\eta_k) \end{cases}$$

Does it converge superlinearly locally?

Euclidean version:

$$\begin{cases} d_k = \mathrm{argmin}_p \langle \nabla f(x_k), p \rangle + \frac{1}{2} \langle p, \nabla^2 f(x_k) p \rangle + h(x_k + p) \\ x_{k+1} = x_k + d_k \end{cases}$$

A native generalization by replacing the Euclidean gradient and Hessian by the Riemannian gradient and Hessian:

$$\begin{cases} \eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \mathrm{grad}\, f(x_k), \eta \rangle + \frac{1}{2} \langle \eta, \mathrm{Hess}\, f(x_k) \eta \rangle + h(x_k + \eta) \\ x_{k+1} = R_{x_k}(\eta_k) \end{cases}$$

Does it converge superlinearly locally?

Not necessarily!

Consider the Sparse PCA over sphere:

$$\min_{x \in \mathbb{S}^{n-1}} -x^{\mathrm{T}} A^{\mathrm{T}} A x + \mu \|x\|_1,$$

where $f(x) = -x^{\mathrm{T}} A^{\mathrm{T}} A x$, $h(x) = \mu \|x\|_1$.

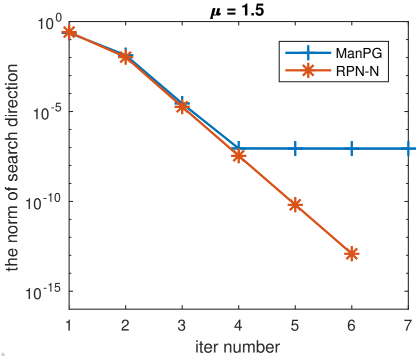


Figure: Comparisons of native generalization (RPN-N) and the proximal gradient method (ManPG) in [CMSZ20].

Euclidean version:

$$\begin{cases} d_k = \mathrm{argmin}_p \langle \nabla f(x_k), p\rangle + \frac{1}{2}\langle p, \nabla^2 f(x_k) p\rangle + h(x_k + p) \\ x_{k+1} = x_k + d_k \end{cases}$$

A native generalization by replacing the Euclidean gradient and Hessian by the Riemannian gradient and Hessian:

$$\begin{cases} \eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k}\mathcal{M}} \langle \mathrm{grad}\, f(x_k), \eta\rangle + \frac{1}{2}\langle \eta, \mathrm{Hess}\, f(x_k)\eta\rangle + h(x_k + \eta) \\ x_{k+1} = R_{x_k}(\eta_k) \end{cases}$$

- $x_k + \eta$ in h is only a first order approximation;

Euclidean version:

$$\begin{cases} d_k = \mathrm{argmin}_p \langle \nabla f(x_k), p\rangle + \frac{1}{2}\langle p, \nabla^2 f(x_k) p\rangle + h(x_k + p) \\ x_{k+1} = x_k + d_k \end{cases}$$

A native generalization by replacing the Euclidean gradient and Hessian by the Riemannian gradient and Hessian:

$$\begin{cases} \eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k}\mathcal{M}} \langle \mathrm{grad} f(x_k), \eta\rangle + \frac{1}{2}\langle \eta, \mathrm{Hess} f(x_k)\eta\rangle + h(x_k + \eta) \\ x_{k+1} = R_{x_k}(\eta_k) \end{cases}$$

$$\begin{cases} \eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k}\mathcal{M}} \langle \mathrm{grad} f(x_k), \eta\rangle + \frac{1}{2}\langle \eta, \mathrm{Hess} f(x_k)\eta\rangle + h(x_k + \eta + \frac{1}{2}\Pi(\eta, \eta)) \\ x_{k+1} = R_{x_k}(\eta_k) \end{cases}$$

- $x_k + \eta$ in h is only a first order approximation;
- If an second order approximation is used, then the subproblem is difficult to solve;

A Riemannian proximal Newton method (RPN)

1. Compute
$v(x_k) = \mathrm{argmin}_{v \in \mathrm{T}_{x_k} \mathcal{M}} \; f(x_k) + \langle \nabla f(x_k), v \rangle + \frac{1}{2t} \|v\|_F^2 + h(x_k + v);$
2. Find $u(x_k) \in \mathrm{T}_{x_k} \mathcal{M}$ by solving
$$J(x_k)[u(x_k)] = -v(x_k),$$
where $J(x_k) = -\left[\mathrm{I}_n - \Lambda_{x_k} + t\Lambda_{x_k}(\nabla^2 f(x_k) - \mathcal{L}_{x_k})\right]$, Λ_{x_k} and $\mathcal{L}_{x_k}$ are defined later ;
3. $x_{k+1} = R_{x_k}(u(x_k))$;

A Riemannian proximal Newton method (RPN)

1. Compute
$$v(x_k) = \mathrm{argmin}_{v \in \mathrm{T}_{x_k} \mathcal{M}} \; f(x_k) + \langle \nabla f(x_k), v \rangle + \frac{1}{2t} \|v\|_F^2 + h(x_k + v);$$
2. Find $u(x_k) \in \mathrm{T}_{x_k} \mathcal{M}$ by solving
$$J(x_k)[u(x_k)] = -v(x_k),$$
where $J(x_k) = -\left[\mathrm{I}_n - \Lambda_{x_k} + t \Lambda_{x_k} (\nabla^2 f(x_k) - \mathcal{L}_{x_k})\right]$, Λ_{x_k} and $\mathcal{L}_{x_k}$ are defined later ;
3. $x_{k+1} = R_{x_k}(u(x_k))$;

1. Step 1: compute a Riemannian proximal gradient direction (ManPG)

A Riemannian proximal Newton method (RPN)

1. Compute
$$v(x_k) = \mathrm{argmin}_{v \in \mathrm{T}_{x_k} \mathcal{M}} \; f(x_k) + \langle \nabla f(x_k), v \rangle + \frac{1}{2t} \|v\|_F^2 + h(x_k + v);$$
2. Find $u(x_k) \in \mathrm{T}_{x_k} \mathcal{M}$ by solving
$$J(x_k)[u(x_k)] = -v(x_k),$$
where $J(x_k) = -\left[\mathrm{I}_n - \Lambda_{x_k} + t\Lambda_{x_k}(\nabla^2 f(x_k) - \mathcal{L}_{x_k})\right]$, Λ_{x_k} and $\mathcal{L}_{x_k}$ are defined later ;
3. $x_{k+1} = R_{x_k}(u(x_k))$;

1. Step 1: compute a Riemannian proximal gradient direction (ManPG)
2. Step 2: compute the Riemannian proximal Newton direction, where $J(x_k)$ is from a generalized Jacobi of $v(x_k)$;

A Riemannian proximal Newton method (RPN)

1. Compute

   $v(x_k) = \mathrm{argmin}_{v \in \mathrm{T}_{x_k} \mathcal{M}} \; f(x_k) + \langle \nabla f(x_k), v \rangle + \frac{1}{2t} \|v\|_F^2 + h(x_k + v);$

2. Find $u(x_k) \in \mathrm{T}_{x_k} \mathcal{M}$ by solving

   $$J(x_k)[u(x_k)] = -v(x_k),$$

   where $J(x_k) = -\left[\mathrm{I}_n - \Lambda_{x_k} + t\Lambda_{x_k}(\nabla^2 f(x_k) - \mathcal{L}_{x_k})\right]$, Λ_{x_k} and $\mathcal{L}_{x_k}$ are defined later ;

3. $x_{k+1} = R_{x_k}(u(x_k));$

1. Step 1: compute a Riemannian proximal gradient direction (ManPG)
2. Step 2: compute the Riemannian proximal Newton direction, where $J(x_k)$ is from a generalized Jacobi of $v(x_k)$;
3. Step 3: Update iterate by a retraction;

A Riemannian proximal Newton method (RPN)

1. Compute
$$v(x_k) = \mathrm{argmin}_{v \in \mathrm{T}_{x_k} \mathcal{M}} \; f(x_k) + \langle \nabla f(x_k), v \rangle + \frac{1}{2t} \|v\|_F^2 + h(x_k + v);$$
2. Find $u(x_k) \in \mathrm{T}_{x_k} \mathcal{M}$ by solving
$$J(x_k)[u(x_k)] = -v(x_k),$$
where $J(x_k) = -\left[\mathrm{I}_n - \Lambda_{x_k} + t\Lambda_{x_k}(\nabla^2 f(x_k) - \mathcal{L}_{x_k})\right]$, Λ_{x_k} and $\mathcal{L}_{x_k}$ are defined later ;
3. $x_{k+1} = R_{x_k}(u(x_k))$;

Next, we will show:

1. G-semismoothness of $v(x_k)$ and its generalized Jacobi;
2. Superlinear convergence rate;

Definition (G-Semismoothness [Gow04])

Let $F : \mathcal{D} \to \mathbb{R}^m$ where $\mathcal{D} \subset \mathbb{R}^n$ be an open set, $\mathcal{K} : \mathcal{D} \rightrightarrows \mathbb{R}^{m \times n}$ be a nonempty set-valued mapping. We say that F is G-semismooth at $x \in \mathcal{D}$ with respect to $\mathcal{K}$ if for any $J \in \mathcal{K}(x + d)$,

$$F(x + d) - F(x) - Jd = o(\|d\|) \text{ as } d \to 0.$$

If F is G-semismooth at any $x \in \mathcal{D}$ with respect to $\mathcal{K}$, then F is called a G-semismooth function with respect to $\mathcal{K}$.

The standard definition of semismoothness additional requires:

- $\mathcal{K}$ is compact valued, upper semicontinuous set-valued mapping;
- F is a locally Lipschitz continuous function;
- F is directionally differentiable at x;

[Gow04] M Seetharama Gowda. Inverse and implicit function theorems for h-differentiable and semismooth functions. Optimization Methods and Software, 19(5):443-461, 2004.

$v(x)$ (dropping the subscript for simplicity)

$$v(x) = \operatorname*{argmin}_{v \in \mathrm{T}_x \mathcal{M}} f(x) + \langle \nabla f(x), v \rangle + \frac{1}{2t} \|v\|_F^2 + h(x + v);$$

$v(x)$ (dropping the subscript for simplicity)

$$v(x) = \operatorname*{argmin}_{v \in \mathrm{T}_x \mathcal{M}} \; f(x) + \langle \nabla f(x), v \rangle + \frac{1}{2t} \|v\|_F^2 + h(x+v);$$

Above problem can be rewritten as

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

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

A Riemannian proximal Newton method

G-semismoothness of $v(x)$

The Lagrangian function:

$$\mathcal{L}(v, \lambda) = \langle \xi_x, v \rangle + \frac{1}{2t} \langle v, v \rangle + h(X + v) - \langle \lambda, B_x^T v \rangle.$$

Therefore

$$\text{KKT: } \left\{ \begin{array}{c} \partial_v \mathcal{L}(v, \lambda) = 0 \\ B_x^T v = 0 \end{array} \right. \Longrightarrow \left\{ \begin{array}{c} v = \mathrm{Prox}_{th} \left(x - t(\xi_x - B_x \lambda) \right) - x \\ B_x^T v = 0 \end{array} \right.$$

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

Define

$$\mathcal{F} : \mathbb{R}^n \times \mathbb{R}^{n+d} \mapsto \mathbb{R}^{n+d} : (x; v, \lambda) \mapsto \begin{pmatrix} v + x - \mathrm{Prox}_{th}(x - t[\nabla f(x) + B_x \lambda]) \\ B_x^T v \end{pmatrix}.$$

$v(x)$ is the solution of the system $\mathcal{F}(x, v(x), \lambda(x)) = 0$;

Define

$$\mathcal{F} : \mathbb{R}^n \times \mathbb{R}^{n+d} \mapsto \mathbb{R}^{n+d} : (x; v, \lambda) \mapsto \begin{pmatrix} v + x - \mathrm{Prox}_{th}(x - t[\nabla f(x) + B_x \lambda]) \\ B_x^T v \end{pmatrix}.$$

- $\mathcal{F}$ is semismooth;
- $v(x)$ is G-semismooth by the G-semismooth Implicit Function Theorem in [Gow04, PSS03];

[Gow04] M Seetharama Gowda. Inverse and implicit function theorems for h-differentiable and semismooth functions. Optimization Methods and Software, 19(5):443-461, 2004.

[PSS03] Jong-Shi Pang, Defeng Sun, and Jie Sun. Semismo oth homeomorphisms and strong stability of semidefinite and Lorentz complementarity problems. Mathematics of Operations Research, 28(1):39-63, 2003.

Lemma (Semismooth Implicit Function Theorem)

Suppose that $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ is a semismooth function with respect to $\partial_{\mathrm{B}} F$ in an open neighborhood of (x^0, y^0) with $F(x^0, y^0) = 0$. Let $H(y) = F(x^0, y)$, if every matrix in $\partial_C H(y^0)$ is nonsingular, then there exists an open set $\mathcal{V} \subset \mathbb{R}^n$ containing x^0, a set-valued fucntion $\mathcal{K} : \mathcal{V} \to \mathbb{R}^{m \times n}$, and a G-semismooth function $f : \mathcal{V} \to \mathbb{R}^m$ with respect to $\mathcal{K}$ satisfying $f(x^0) = y^0$, for every $x \in \mathcal{V}$,

$$F(x, f(x)) = 0,$$

and the set-valued function $\mathcal{K}$ is

$$\mathcal{K} : x \mapsto \{-(A_y)^{-1} A_x : [A_x \ A_y] \in \partial_{\mathrm{B}} F\big(x, f(x)\big)\},$$

where the map $x \mapsto \mathcal{K}(x)$ is compact valued and upper semicontinuous.

Without loss of generality, we assume that the nonzero entries of x_* are in the first part, i.e., $x_* = [\bar{x}_*^T, 0^T]^T$

Assumption

Let $B_{x_*}^{\mathrm{T}} = [\bar{B}_{x_*}^{\mathrm{T}}, \hat{B}_{x_*}^{\mathrm{T}}]$, where $\bar{B}_{x_*} \in \mathbb{R}^{j \times d}$ and $\hat{B}_{x_*} \in \mathbb{R}^{(n-j) \times d}$. It is assumed that $j \geq d$ and $\bar{B}_{x_*}$ is full column rank.

Without loss of generality, we assume that the nonzero entries of x_* are in the first part, i.e., $x_* = [\bar{x}_*^T, 0^T]^T$

Assumption

Let $B_{x_*}^{\mathrm{T}} = [\bar{B}_{x_*}^{\mathrm{T}}, \hat{B}_{x_*}^{\mathrm{T}}]$, where $\bar{B}_{x_*} \in \mathbb{R}^{j \times d}$ and $\hat{B}_{x_*} \in \mathbb{R}^{(n-j) \times d}$. It is assumed that $j \geq d$ and $\bar{B}_{x_*}$ is full column rank.

$v(x)$ is a G-semismooth function of x in a neighborhood of x_*

Under the above Assumption, there exists a neighborhood $\mathcal{U}$ of x_* such that $v : \mathcal{U} \to \mathbb{R}^n : x \mapsto v(x)$ is a G-semismooth function with respect to $\mathcal{K}_v$, where

$$\mathcal{K}_v : x \mapsto \left\{-[\mathrm{I}_n,\ 0]B^{-1}A : [A\ B] \in \partial_{\mathrm{B}} \mathcal{F}\big(x, v(x), \lambda(x)\big)\right\}.$$

For $x \in \mathcal{U}$, any element of $\mathcal{K}_v(x)$ is called a generalized Jacobi of v at x.

Here, the semismooth implicit function theorem is used

The generalized Jacobi of v at x is

$$\Big\{\mathcal{J}_x \,|\mathcal{J}_x[\omega] = -\left[\mathrm{I}_n - \Lambda_x + t\Lambda_x(\nabla^2 f(x) - \mathcal{L}_x)\right]\omega - M_x B_x H_x(\mathrm{D}B_x^{\mathrm{T}}[\omega])v, \forall\omega$$
$$M_x \in \partial_C \mathrm{prox}_{th}(x)\Big\},$$

where $\Lambda_x = M_x - M_x B_x H_x B_x^T M_k$, $H_x = \left(B_x^T M_x B_x\right)^{-1}$,
$\mathcal{L}_x(\cdot) = \mathcal{W}_x(\cdot, B_x\lambda(x))$, and $\mathcal{W}_x$ denotes the Weingarten map;

- $v(x_*) = 0$;
- Set $J(x) = \mathrm{I}_n - \Lambda_x + t\Lambda_x(\nabla^2 f(x) - \mathcal{L}_x)$;
- The Riemannian proximal Newton direction: $J(x)u(x) = -v(x)$;
- Let $u(x) = (\bar{u}(x); \hat{u}(x))$, then

$$\hat{u}(x) = \hat{v} \text{ and } \bar{J}(x)\bar{u}(x) = -\bar{v}(x)$$

Assumption:

1. Let $B_{x_*}^T = [\bar{B}_{x_*}^T, \hat{B}_{x_*}^T]$, where $\bar{B}_{x_*} \in \mathbb{R}^{j \times d}$ and and $\hat{B}_{x_*} \in \mathbb{R}^{(n-j) \times d}$. It is assumed that $j \geq d$ and $\bar{B}_{x_*}$ is full column rank;

Assumption:

1. Let $B_{x_*}^T = [\bar{B}_{x_*}^T, \hat{B}_{x_*}^T]$, where $\bar{B}_{x_*} \in \mathbb{R}^{j \times d}$ and and $\hat{B}_{x_*} \in \mathbb{R}^{(n-j) \times d}$. It is assumed that $j \geq d$ and $\bar{B}_{x_*}$ is full column rank;
2. There exists a neighborhood $\mathcal{U}$ of $x_* = [\bar{x}_*^T, 0^T]^T$ on $\mathcal{M}$ such that for any $x = [\bar{x}^T, \tilde{x}^T]^T \in \mathcal{U}$, it holds that $\bar{x} + \bar{v} \neq 0$ and $\hat{x} + \hat{v} = 0$.

$$v(x) = \operatorname*{argmin}_{v \in \mathrm{T}_x \mathcal{M}} f(x) + \langle \nabla f(x), v \rangle + \frac{1}{2t} \|v\|_F^2 + h(x + v)$$

Assumption:

1. Let $B_{x_*}^T = [\bar{B}_{x_*}^T, \hat{B}_{x_*}^T]$, where $\bar{B}_{x_*} \in \mathbb{R}^{j \times d}$ and and $\hat{B}_{x_*} \in \mathbb{R}^{(n-j) \times d}$. It is assumed that $j \geq d$ and $\bar{B}_{x_*}$ is full column rank;
2. There exists a neighborhood $\mathcal{U}$ of $x_* = [\bar{x}_*^T, 0^T]^T$ on $\mathcal{M}$ such that for any $x = [\bar{x}^T, \tilde{x}^T]^T \in \mathcal{U}$, it holds that $\bar{x} + \bar{v} \neq 0$ and $\hat{x} + \hat{v} = 0$.

Theorem

Suppose that x_* be a local optimal minimizer. Under the above Assumptions, assume that $J(x_*)$ is nonsingular. Then there exists a neighborhood $\mathcal{U}$ of x_* on $\mathcal{M}$ such that for any $x_0 \in \mathcal{U}$, RPN Algorithm generates the sequence $\{x_k\}$ converging superlinearly to x_*.

Assumption:

1. Let $B_{x_*}^T = [\bar{B}_{x_*}^T, \hat{B}_{x_*}^T]$, where $\bar{B}_{x_*} \in \mathbb{R}^{j \times d}$ and and $\hat{B}_{x_*} \in \mathbb{R}^{(n-j) \times d}$. It is assumed that $j \geq d$ and $\bar{B}_{x_*}$ is full column rank;
2. There exists a neighborhood $\mathcal{U}$ of $x_* = [\bar{x}_*^T, 0^T]^T$ on $\mathcal{M}$ such that for any $x = [\bar{x}^T, \tilde{x}^T]^T \in \mathcal{U}$, it holds that $\bar{x} + \bar{v} \neq 0$ and $\hat{x} + \hat{v} = 0$.

Theorem

Suppose that x_* be a local optimal minimizer. Under the above Assumptions, assume that $J(x_*)$ is nonsingular. Then there exists a neighborhood $\mathcal{U}$ of x_* on $\mathcal{M}$ such that for any $x_0 \in \mathcal{U}$, RPN Algorithm generates the sequence $\{x_k\}$ converging superlinearly to x_*.

If the intersection of manifold and sparsity constraints forms an embedded manifold around x_*, then $\nabla^2 \bar{f}(x_*) - \bar{\mathcal{L}} \succeq 0$. If $\nabla^2 \bar{f}(x_*) - \bar{\mathcal{L}} \succ 0$, then $J(x_*)$ is nonsingular.

$$\text{Smooth case: } \min_{x \in \mathcal{M}} f(x)$$

- KKT conditions:

$$\nabla f(x) + \frac{1}{t} v + B_x \lambda = 0, \text{ and } B_x^T v = 0;$$

- Closed form solutions:

$$\lambda(x) = -B_x^{\mathrm{T}} \nabla f(x), \qquad v = -t \,\mathrm{grad}\, f(x);$$

- Action of $J(x)$: for $\omega \in \mathrm{T}_x \mathcal{M}$

$$J(x)[\omega] = -\, t P_{\mathrm{T}_x \mathcal{M}} (\nabla^2 f(x) - \mathcal{L}_x) P_{\mathrm{T}_x \mathcal{M}} \omega = -t \,\mathrm{Hess}\, f(x)[\omega]$$

- $J(x) u(x) = -v(x) \Longrightarrow \mathrm{Hess}\, f(x)[u(x)] = -\,\mathrm{grad}\, f(x)$;
- It is the Riemannian Newton method;

- Euclidean proximal gradient method and its variants;
- Riemannian proximal gradient method and its variants;
- A Riemannian proximal Newton method;
- Numerical experiments;

Numerical Experiments

Sparse PCA problem

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

where $A \in \mathbb{R}^{m \times n}$ is a data matrix and
$\mathrm{St}(r, n) = \{X \in \mathbb{R}^{n \times r} \mid X^T X = I_r\}$ is the compact Stiefel manifold.

- $R_x(\eta_x) = (x + \eta_x)(I + \eta_x^T \eta_x)^{-1/2}$;
- $t = 1/(2\|A\|_2^2)$;
- Run ManPG until $\|v\|$ reaches 10^{-4}, i.e., it reduces by a factor of 10^3. The resulting x as the input of RPN;

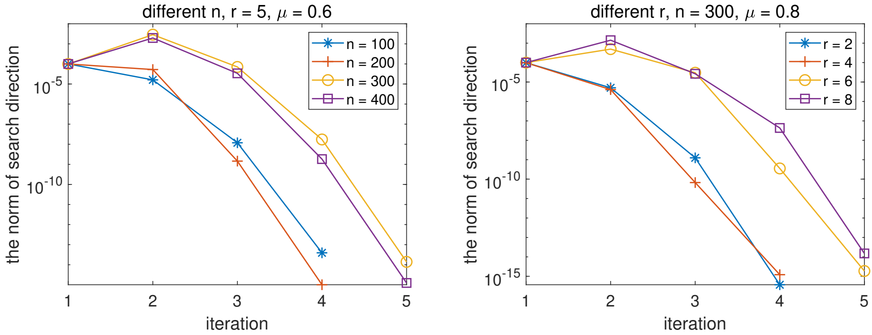


Figure: Random data. Left: different $n = \{100, 200, 300, 400\}$ with $r = 5$ and $\mu = 0.6$; Right: different $r = \{2, 4, 6, 8\}$ with $n = 300$ and $\mu = 0.8$

Numerical Experiments

A Hybrid version of ManPG and RPN

Require: $x_0 \in \mathcal{M}$, $t > 0$, $\rho \in (0, \frac{1}{2}]$, $\epsilon > 0$;

for $k = 0, 1, \ldots$ **do**
  Compute v_k by solving the Riemannian proximal gradient subproblem;
  if $\|v_k\| > \epsilon$ **then**
    Set $\alpha = 1$;
    while $F(R_{x_k}(\alpha v_k)) > F(x_k) - \frac{1}{2}\alpha\|v_k\|^2$ **do**
      $\alpha = \rho\alpha$;
    end while
    $x_{k+1} = R_{x_k}(\alpha v_k)$;
  else
    Compute u_k by solving $J(x_k)u_k = -v_k$;
    $x_{k+1} = R_{x_k}(u_k)$;
  end if
end for

Consider the simple version of sparse PCA with $r = 1$, i.e.,

$$\min_{x \in \mathbb{S}^{n-1}} -x^T A^T A x + \mu \|x\|_1,$$

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

Table: An average result of 5 random runs for random data with different setting of (n, μ). The subscript k indicates a scale of 10^k. iter-u denotes the number of using the new search direction u_k.

(n, μ)	Algo	iter	iter-v	iter-u	f	sparsity	$\|v_k\|$
(5000,1.5)	ManPG	3000	897	-	-4.59_1	0.37	7.41_{-8}
(5000,1.5)	RPN	334	-	5	-4.59_1	0.37	4.53_{-16}
(10000,1.8)	ManPG	3000	1736	-	-1.02_2	0.32	2.19_{-8}
(10000,1.8)	RPN	580	-	6	-1.02_2	0.32	5.69_{-16}
(30000,2.0)	ManPG	3000	1283	-	-3.98_2	0.22	1.19_{-8}
(30000,2.0)	RPN	347	-	5	-3.98_2	0.22	5.25_{-15}
(50000,2.2)	ManPG	3000	1069	-	-7.06_2	0.18	4.56_{-7}
(50000,2.2)	RPN	789	-	5	-7.06_2	0.18	1.41_{-14}
(80000,2.5)	ManPG	3000	834	-	-1.17_3	0.17	1.41_{-6}
(80000,2.5)	RPN	839	-	6	-1.17_3	0.17	1.94_{-15}

Stopping criteria: ManPG does not terminate until iteration attains the maximal iteration (3000), RPN terminate until $\|v_k\| \leq 10^{-12}$

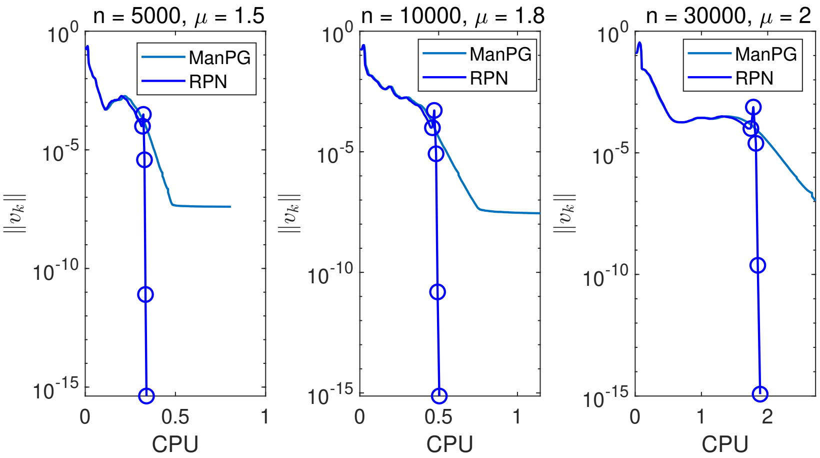


Figure: Random data: the norm of search direction v_k versus CPU for different (n, μ), where the blue circle indicates the use of the new direction u_k.

Synthetic Data [SCL+18] : we first obtain an $m \times n$ noise-free matrix, then the data matrix A is generated by adding a random noise matrix, where each entry of the noise matrix is drawn form $\mathcal{N}(0, 0.25)$, we set $m = 400$, $n = 4000$ and $\mu = 1.2$.

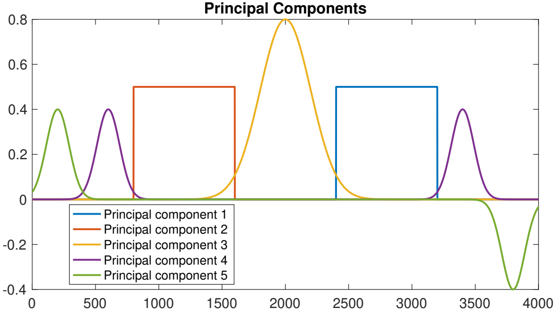


Figure: The five principal components used in the synthetic data.

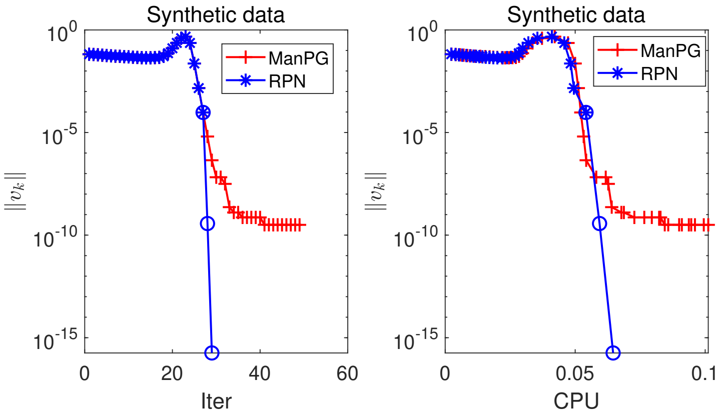


Figure: Plots of $\|v_k\|$ versus iterations and CPU times respectively, where $\|v_k\|$ is the norm of search direction, data matrix $A \in \mathbb{R}^{4000\times 400}$ is from the synthetic data, μ is set to be 1.2. Note that the blue circle indicates the use of the new direction u_k.

Summary

- Briefly review Euclidean and Riemannian proximal gradient method and its variants;
- Propose a Riemannian proximal Newton method;
- Local superlinear convergence rate is proven;
- Numerical experiments show its performance;

Future work

- Globalization;
- Other types of $h(x)$;
- General manifold;
- Riemannian proximal inexact-Newton methods;
- Riemannian proximal quasi-Newton methods;

Thank you

Thank you!

References I

Matthias Bollh ofer, Aryan Eftekhari, Simon Scheidegger, and Olaf Schenk.
Large-scale sparse inverse covariance matrix estimation.
SIAM Journal on Scientific Computing, 41(1):A380–A401, 2019.

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.

Haoran Chen, Yanfeng Sun, Junbin Gao, Yongli Hu, and Baocai Yin.
Fast optimization algorithm on riemannian manifolds and its application in low-rank learning.
Neurocomputing, 291:59 – 70, 2018.

M Seetharama Gowda.
Inverse and implicit function theorems for h-differentiable and semismooth functions.
Optimization Methods and Software, 19(5):443–461, 2004.

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

Wen Huang and Ke Wei.
An extension of fast iterative shrinkage-thresholding algorithm to Riemannian optimization for sparse principal component analysis.
Numerical Linear Algebra with Applications, page e2409, 2021.

References II

Wen Huang, Meng Wei, Kyle A. Gallivan, and Paul Van Dooren.
A Riemannian Optimization Approach to Clustering Problems, 2022.

Jason D Lee, Yuekai Sun, and Michael A Saunders.
Proximal newton-type methods for minimizing composite functions.
SIAM Journal on Optimization, 24(3):1420–1443, 2014.

Boris S Mordukhovich, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang.
A globally convergent proximal newton-type method in nonsmooth convex optimization.
Mathematical Programming, pages 1–38, 2022.

Vidvuds Ozolinš, Rongjie Lai, Russel Caflisch, and Stanley Osher.
Compressed modes for variational problems in mathematics and physics.
Proceedings of the National Academy of Sciences, 110(46):18368–18373, 2013.

Jong-Shi Pang, Defeng Sun, and Jie Sun.
Semismooth homeomorphisms and strong stability of semidefinite and lorentz complementarity problems.
Mathematics of Operations Research, 28(1):39–63, 2003.

K. Sjöstrand, L. Clemmensen, R. Larsen, G. Einarsson, and B. Ersboll.
SpaSM: A matlab toolbox for sparse statistical modeling.
Journal of Statistical Software, Articles, 84(10):1–37, 2018.

K. Scheinberg and X. Tang.
Practical inexact proximal quasi-newton method with global complexity analysis.
Mathematical Programming, (160):495–529, February 2016.

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

References III

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.