Riemannian proximal gradient methods and variants

Speaker: Wen Huang

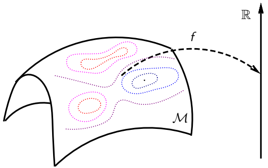
Xiamen University

July 10, 2023

Beijing University

Optimization on Manifolds with Structure:

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

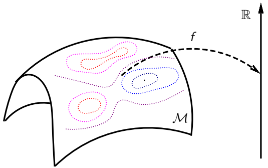


- $\mathcal{M}$ is a finite-dimensional Riemannian manifold;
- f is smooth and may be nonconvex; and
- $h(x)$ is continuous and convex but may be nonsmooth;

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].

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, 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
- Xiao, Liu, and Yuan (2021), Infeasible approach over the Stiefel manifold
- Zhou, Bao, and Ding (2022), An augmented Lagrangian method on matrix manifolds
- Huang and Wei (2021-2023), A Riemannian proximal gradient method, an inexact Riemannian proximal gradient method, and a modified FISTA on embedded manifolds
- Wang and Yang (2023), A proximal quasi-Newton method on manifolds on the Stiefel manifold
- Huang, Meng, Gallivan, and Van Dooren (2023), An inexact proximal gradient method on embedded submanifolds
- Beck and Rosset (2023), A dynamic smoothing technique

Outline

Optimization with Structure:

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

- Proximal gradient methods
- Inexact proximal gradient methods
- A proximal Newton method

[HW2021]: W. Huang and K. Wei, Riemannian proximal gradient methods, Mathematics Programming, 194, 371-413, 2022.
[HW2023]: An inexact Riemannian proximal gradient method, Computational Optimization and Applications, 85, 1-32, 2023
[HWGV2023]: A Riemannian optimization approach to clustering problems, arxiv, 2023
[SAHJV2023]: A Riemannian proximal Newton method, arxiv, 2023

Outline

Optimization with Structure:

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

- Proximal gradient methods
  - Euclidean version
  - Riemannian version in [CMSZ20]
  - Riemannian version in [HW21a]
- Inexact proximal gradient methods
- A proximal Newton method

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

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

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

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

---

initial iterate: x_0,

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

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

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

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

---

initial iterate: x_0,

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

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

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

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

---

initial iterate: x_0,

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

- $h = 0$: reduce to steepest descent method;
- L: greater than the Lipschitz constant of ∇f;

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

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

---

initial iterate: x_0,

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

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

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

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

---

initial iterate: x_0,

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

- $h = 0$: 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;

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

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

---

initial iterate: x_0,

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

- $h = 0$: 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;
- $O\left(\frac{1}{k}\right)$ sublinear convergence rate for convex f and h;

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

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

---

initial iterate: x_0,

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

- $h = 0$: 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;
- $O\left(\frac{1}{k}\right)$ sublinear convergence rate for convex f and h;
- Linear convergence rate for strongly convex f and convex h;

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

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

---

initial iterate: x_0,

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

- $h = 0$: 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;
- $O\left(\frac{1}{k}\right)$ sublinear convergence rate for convex f and h;
- Linear convergence rate for strongly convex f and convex h;
- Local convergence rate by KL property;

Optimization with Structure: $\mathcal{M}$

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

Optimization with Structure: $\mathcal{M}$

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

Euclidean proximal mapping

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

In the Riemannian setting:

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

A Riemannian proximal mapping [CMSZ20]

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

- Only works for embedded submanifold;

[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 mapping [CMSZ20]

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

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

A Riemannian proximal mapping [CMSZ20]

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

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

Proximal Gradient Method

Riemannian version in [CMSZ20]

[CMSZ20]

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

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

[XLWZ18]: X. Xiao, Y. Li, Z. Wen, and L. Zhang, A regularized semi-smooth Newton method with projection steps for composite convex programs. Journal of Scientific Computing, 76(1):364-389, 2018.

ManPG [CMSZ20]

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k}\mathcal{M}} \langle \nabla f(x_k), \eta \rangle + \frac{L}{2}\|\eta\|_F^2 + h(x_k + \eta)$;
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-smooth Newton algorithm [XLWZ18];
- Step size 1 is not necessary decreasing;

Proximal Gradient Method

Riemannian version in [CMSZ20]

ManPG [CMSZ20]

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k}\mathcal{M}} \langle \nabla f(x_k), \eta \rangle + \frac{L}{2}\|\eta\|_F^2 + h(x_k + \eta)$;
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-smooth Newton algorithm [XLWZ18];
- Step size 1 is not necessary decreasing;
- Convergence to a stationary point;

ManPG [CMSZ20]

1. $\eta_k = \arg\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \nabla f(x_k), \eta \rangle + \frac{L}{2} \|\eta\|_F^2 + h(x_k + \eta)$;
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-smooth Newton algorithm [XLWZ18];
- Step size 1 is not necessary decreasing;
- 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 Riemannian Proximal Gradient Method (RPG)

Let $\ell_{x_k}(\eta) = \underbrace{\langle \nabla f(x_k), \eta \rangle_{x_k} + \frac{L}{2}\|\eta\|_{x_k}^2}_{\text{Riemannian metric}} + h(\underbrace{R_{x_k}(\eta)}_{\text{replace } x_k + \eta});$

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;

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

A Riemannian Proximal Gradient Method (RPG)

$$\text{Let } \ell_{x_k}(\eta) = \underbrace{\langle \nabla f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2}_{\text{Riemannian metric}} + h(\underbrace{R_{x_k}(\eta)}_{\text{replace } x_k + \eta});$$

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;
- Step size can be fixed to be 1;

Proximal Gradient Method

Riemannian version in [HW21a]

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

A Riemannian Proximal Gradient Method (RPG)

Let $\ell_{x_k}(\eta) = \underbrace{\langle \nabla f(x_k), \eta \rangle_{x_k} + \frac{L}{2} \|\eta\|_{x_k}^2}_{\text{Riemannian metric}} + h(\underbrace{R_{x_k}(\eta)}_{\text{replace } x_k + \eta});$

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;
- Step size can be fixed to be 1;
- Convergence rate results;

Proximal Gradient Method

Riemannian version in [HW21a]

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.

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

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

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

Proximal Gradient Method

Riemannian version in [HW21a]

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 $\mathcal{M}$;

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

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., $0 \in \partial F(x_*)$.

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

$$q_x(\eta) \geq q_x(\xi) + \langle \zeta, \eta - \xi \rangle_x \;\; \forall \eta, \xi \in \mathcal{S}_x. \tag{1}$$

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

Proximal Gradient Method

Riemannian version in [HW21a]

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;

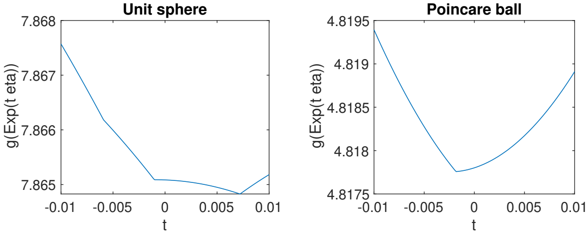
then there exists a neighborhood of x, denoted by $\mathcal{N}_x$, such that the function $h : \mathcal{M} \to \mathbb{R}$ is retraction-convex in $\mathcal{N}_x$.

Additional Assumptions:

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

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

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



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

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 for compact set $\overline{\Omega}$ with the exponential mapping;

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 $O(1/k)$:

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

Proximal Gradient Method

Riemannian version in [HW21a]

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}.

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

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

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

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

Proximal Gradient Method

Riemannian version in [HW21a]

Assumption:

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

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

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

Assumption:

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

Definition

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

- $\varsigma(0) = 0$, 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

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

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

Assumption:

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

Theoretical results:

- it holds that

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

Therefore, there exists only a unique accumulation point.

Proximal Gradient Method

Riemannian version in [HW21a]

Assumption:

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

Theoretical results:

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

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.

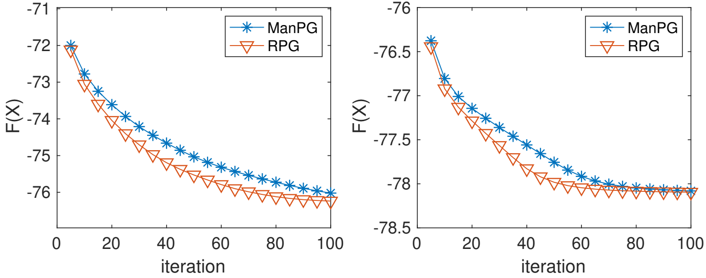


Figure: Two typical runs of ManPG, RPG, A-ManPG, and A-RPG for the Sparse PCA problem. $n = 1024$, $p = 4$, $\lambda = 2$, $m = 20$.

Outline

Optimization with Structure:

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

- Proximal gradient methods
- Inexact proximal gradient methods
  - Inexact version of RPG [HW21a]
  - Inexact version of ManPG [CMSZ20]
- A proximal Newton method

Inexact Proximal Gradient Method

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?
- Inexact RPG
- Inexact ManPG

Inexact RPG (IRPG)

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

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

Inexact RPG (IRPG)

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

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

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

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

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

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

Inexact Proximal Gradient Method

Inexact RPG

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

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

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

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

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

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

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

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

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

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

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

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

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

- Algorithms based on strong convexity of the Euclidean proximal mapping

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

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

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

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

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

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

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

Assumption (same as the RPG):

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., $0 \in \partial F(x_*)$.

Assumption:

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

Assumption:

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

Definition

A function $h : \mathcal{M} \to \mathbb{R}$ 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 $|h \circ R(\xi_x) - h \circ R(\eta_x)| \leq L_h \|\xi_x - \eta_x\|$.

Assumption:

1. Assumptions for the global convergence
2. f is locally Lipschitz continuously differentiable
3. F satisfies the Riemannian KL property
4. 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.

Assumption:

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

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.

Assumption:

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

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 $C_2 > 0$;

IRPG

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

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;

How to find $\hat{\eta}_k$ for different q?

IRPG

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

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;

How to find $\hat{\eta}_k$ for different q?

- Only consider manifolds with a linear ambient space;

IRPG

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

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;

How to find $\hat{\eta}_k$ for different q?

- Only consider manifolds with a linear ambient space;
- Use the semi-smooth Newton method iteratively;

Inexact Proximal Gradient Method

Inexact RPG

IRPG

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

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;

How to find $\hat{\eta}_k$ for different q?

- Only consider manifolds with a linear ambient space;
- Use the semi-smooth Newton method iteratively;
- For sufficiently large $\tilde{L}$, η_k from ManPG guarantees global convergence;

ManPG [CMSZ20]

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

Above problem can be rewritten as

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

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

The Lagrangian function:

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

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

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 $h(\cdot) = \|\cdot\|_1$;
- 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 $h(\cdot) = \|\cdot\|_1$;
- 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:

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

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

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

The function q is:

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

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

- 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:

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

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

- 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;

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

Solving the Riemannian Proximal Mapping [HW21a]

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
   $$\xi_k^* \approx \arg \min_{\xi \in \mathrm{T}_{y_k} \mathcal{M}} \langle \mathcal{T}_{R_{\eta_k}}^{-\sharp} (\mathrm{grad}\, f(x) + \tilde{L} \eta_k), \xi \rangle_x + \frac{\tilde{L}}{4} \|\xi\|_F^2 + h(y_k + \xi);$$
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;

Inexact RPG for global convergence (IRPG)

1. Approximately solve
$$\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \mathrm{grad}\, f(x_k), \eta \rangle + \frac{\tilde{L}}{2} \|\eta\|_F^2 + h(x_k + \eta)$$
such that $\|\Psi_k(\Lambda)\|_F \leq \min(\phi(\hat{v}(\Lambda)), 0.5)$;
2. Let $\eta_k = \hat{v}(\Lambda)$;
3. $x_{k+1} = R_{x_k}(\eta_k)$;

- Global convergence requires a sufficient large of $\tilde{L}$;
- Step size one is used;

Inexact RPG for global convergence (IRPG)

1. Approximately solve

$$\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \mathrm{grad}\, f(x_k), \eta \rangle + \frac{\tilde{L}}{2} \|\eta\|_F^2 + h(x_k + \eta)$$

   such that $\|\Psi_k(\Lambda)\|_F \leq \min(\phi(\hat{v}(\Lambda)), 0.5)$;
2. Let $\eta_k = \hat{v}(\Lambda)$;
3. $x_{k+1} = R_{x_k}(\eta_k)$;

- Global convergence requires a sufficient large of $\tilde{L}$;
- Step size one is used;
- Is $\hat{\eta}_x$ a descent direction for any positive $\tilde{L}$?

Algorithm 1 ManPG without solving the subproblem exactly

1: Given x_0, $\nu \in (0,1)$, $\sigma \in (0, 1/(8\mu))$, $\mu > 0$;
2: **for** $k = 0, 1, \ldots$ **do**
3: Approximately solve

$$\min_{\eta \in \mathrm{T}_{x_k} \mathcal{M}} \langle \mathrm{grad}\, f(x_k), \eta \rangle + \frac{1}{2\mu} \|\eta\|_F^2 + h(x_k + \eta)$$

such that $\|\Psi_k(\Lambda)\|_F \leq \sqrt{4\mu^2 L_h^2 + \|\hat{v}_k(\Lambda)\|_F^2/2} - 2\mu L_h$;
4: Set $\eta_k = \hat{v}_k(\Lambda)$ and set $\alpha = 1$;
5: **while** $F(R_{x_k}(\alpha \eta_{x_k})) > F(x_k) - \sigma\alpha \|\eta_{x_k}\|_F^2$ **do**
6: $\alpha = \nu\alpha$;
7: **end while**
8: $x_{k+1} = R_{x_k}(\alpha \eta_{x_k})$;
9: **end for**

Assumption

The function f is Lipschitz continuously differentiable on $\mathcal{M}$ and h is Lipschitz continuous with constant L_h.

Theorem

Suppose the assumption holds. Then for any $\mu > 0$, there exists a constant $\bar{\alpha} \in (0, 1]$ such that for any $0 < \alpha < \bar{\alpha}$, the sequence $\{x_k\}$ generated by Algorithm 1 satisfies

$$F(R_{x_k}(\alpha \eta_{x_k})) - F(x_k) \leq -\frac{\alpha}{8\mu} \|\eta_{x_k}\|_F^2.$$

Moreover, the step size $\alpha > \rho \bar{\alpha}$ for all k.

Assumption

The function f is Lipschitz continuously differentiable on $\mathcal{M}$ and h is Lipschitz continuous with constant L_h.

Theorem

Suppose the assumption holds. Then any accumulation point of the sequence $\{x_k\}$ generated by Algorithm 1 is a stationary point, i.e., if x_* is an accumulation point of the above sequence, then $0 \in P_{\mathrm{T}_{x_*} \mathcal{M}} \partial F(x_*)$.

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.

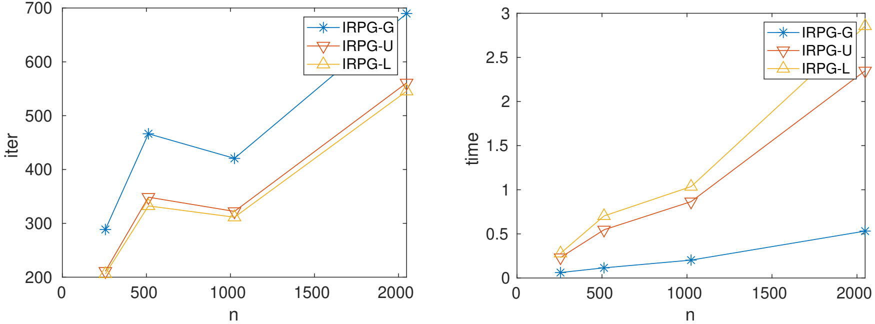


Figure: Average of 10 random runs, $p = 4$, $m = 20$, $\lambda = 2$;

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

Community detection:

$$\min_{X \in \mathcal{F}_{\mathbf{1}_n}} -\mathrm{trace}(X^T M X) + \lambda \|X\|_1,$$

where $\mathcal{F}_{\mathbf{1}_n} = \{X \in \mathbb{R}^{n \times k} : X^T X = I_k, \mathbf{1}_n \in \mathrm{span}(X)\}$

Comparing efficiency of I-AManPG and E-AManPG

	I-A	E-A	I-A	E-A	I-A	E-A	I-A	E-A
(n, q)	$(5000, 10)$		$(5000, 20)$		$(10000, 10)$		$(10000, 20)$	
iter	63	58	47	50	55	55	73	51
SSNiter	34	311	32	381	52	330	146	376
nf	140	128	105	112	123	122	161	113
ng	81	72	60	62	71	68	92	64
nR	139	127	104	111	122	121	160	112
nSG	4	13	2	5	3	10	3	13
F	-2.84_2	-2.84_2	-6.55_2	-6.56_2	-2.51_2	-2.51_2	-6.11_2	-6.14_2
$\frac{\\|\eta_{z_k}\\|}{\\|\eta_{z_0}\\|}$	6.31_{-4}	5.82_{-4}	5.32_{-4}	7.54_{-4}	5.22_{-4}	6.86_{-4}	4.02_{-4}	6.60_{-4}
time	0.84	3.04	1.51	9.81	1.54	5.19	9.48	19.21

- AManPG: add acceleration [HW21b]
- I-AManPG: Inexact version
- E-AManPG: Exact version, i.e., $\epsilon = 10^{-10}$
- An average of 10 random runs

Comparing efficiency of I-AManPG and E-AManPG

	I-A	E-A	I-A	E-A	I-A	E-A	I-A	E-A
(n, q)	$(5000, 10)$		$(5000, 20)$		$(10000, 10)$		$(10000, 20)$	
iter	63	58	47	50	55	55	73	51
SSNiter	34	311	32	381	52	330	146	376
nf	140	128	105	112	123	122	161	113
ng	81	72	60	62	71	68	92	64
nR	139	127	104	111	122	121	160	112
nSG	4	13	2	5	3	10	3	13
F	-2.84_2	-2.84_2	-6.55_2	-6.56_2	-2.51_2	-2.51_2	-6.11_2	-6.14_2
$\frac{\\|\eta_{z_k}\\|}{\\|\eta_{z_0}\\|}$	6.31_{-4}	5.82_{-4}	5.32_{-4}	7.54_{-4}	5.22_{-4}	6.86_{-4}	4.02_{-4}	6.60_{-4}
time	0.84	3.04	1.51	9.81	1.54	5.19	9.48	19.21

Less computational time, same effectiveness

Outline

Optimization with Structure:

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

- Proximal gradient methods
- Inexact proximal gradient methods
- A proximal Newton method
  - Euclidean inexact proximal Newton methods
  - A naive Riemannian proximal Newton method
  - A proposed Riemannian proximal Newton method

Given x_0;

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

Inexact Proximal Newton Methods

Euclidean version

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];

[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.

Inexact Proximal Newton Methods

Euclidean version

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;

[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.

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.

Focus on embedded submanifolds

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:

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

Focus on embedded submanifolds

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?

Focus on embedded submanifolds

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

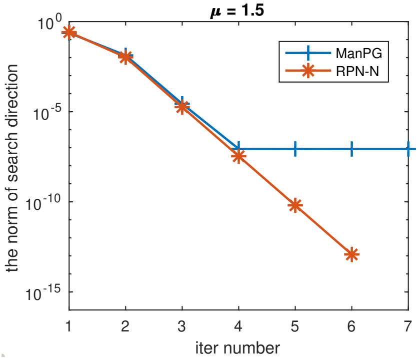


Figure: Comparisons of native generalization (RPN-N) and the proximal gradient method (ManPG) in [CMSZ20].

Euclidean version:

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

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

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

- $x_k + \eta$ 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.

Inexact Proximal Newton Methods

Riemannian version

$v(x)$ (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 $\mathrm{T}_x^{\perp} \mathcal{M}$.

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

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

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 $\mathcal{F}(x, v(x), \lambda(x)) = 0$;

Inexact Proximal Newton Methods

Riemannian version

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.

Not new but an arrangement of existing results.

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;

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;

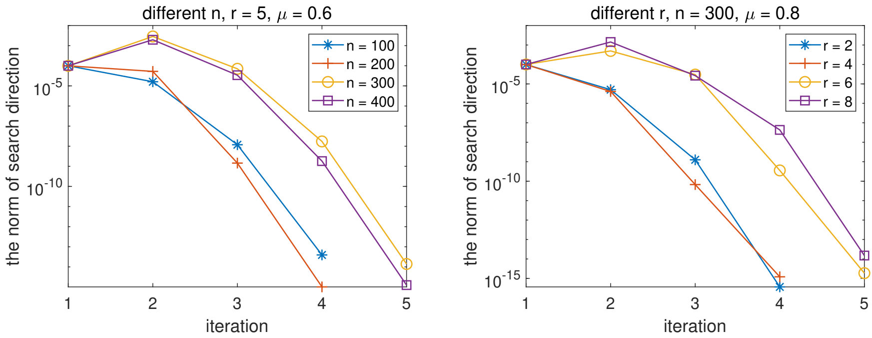


Figure: Random data. Left: different $n = \{100, 200, 300, 400\}$ with $r = 5$ and $\mu = 0.6$; Right: different $r = \{2, 4, 6, 8\}$ with $n = 300$ and $\mu = 0.8$

Summary and Future Work

Summary:

- A non-exhaustive review of nonsmooth optimization on manifolds;
- Euclidean/Riemannian proximal gradient methods;
- Inexact versions;
- Euclidean/Riemannian proximal Newton methods;

Future work:

- Accelerated version: $O(1/k^2)$ convergence rate analysis;
- Globalization for Riemannian Newton method;
- Design a Riemannian quasi-Newton method with superlinear local convergence rate;
- Generalize those methods to generic manifolds;

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

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.

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.

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.

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.

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

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.

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.

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.

References III

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.

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.

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.

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

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.