A Riemannian Accelerated Proximal Gradient Method

Speaker: Wen Huang

Xiamen University

October 18, 2025

Joint work with Shuailing Feng, Yuhang Jiang, Shihui Ying

Symposium on "Scientific Computing: Theory and Applications"
Sichuan University

Outline

1. Riemannian Proximal Gradient Methods Review
2. A Riemannian Accelerated Proximal Gradient Method
3. Adaptive Restart for Riemannian Accelerated Proximal Gradient Method
4. Summary

Outline

1. Riemannian Proximal Gradient Methods Review
   - Problem Statement
   - Related Work
2. A Riemannian Accelerated Proximal Gradient Method
3. Adaptive Restart for Riemannian Accelerated Proximal Gradient Method
4. Summary

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;
- h is continuous but may be nonsmooth;

Applications:

- sparse principal component analysis [ZBDZ23, GHT15, ZHT06];
- clustering [HWGVD25, LYL16, PZ18];
- image processing [OSZ17, YG17];
- compressed modes [OLCO13];
- face expression recognition [FRJA18];

Goal

Optimization on Manifolds with Structure:

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

Goal:

- Propose an accelerated proximal gradient algorithm on manifold;
- Provide a convergence analysis;

Related Work

Euclidean Setting

A proximal gradient method, initial iterate x_0:

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

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

Related Work

Euclidean Setting

A proximal gradient method, initial iterate x_0:[1]

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

FISTA in convex [BT09]:

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

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

- Based on the Nesterov momentum technique;
- Two-point iterative sequence: x_k and y_k;
- $O\left(\frac{1}{k^2}\right)$ sublinear convergence rate for convex f and h;

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

Related Work

Euclidean Setting

FISTA in strongly convex [dST+21]:

$$\text{Given } x_0, \text{ let } z_0 = x_0, A_0 = 0, q = \frac{\mu}{L} \ (\mu \geq 0);$$

$$\begin{cases} A_{k+1} = \frac{2A_k+1+\sqrt{4A_k+4qA_k^2+1}}{2(1-q)} \\ \tau_k = \frac{(A_{k+1}-A_k)(1+qA_k)}{A_{k+1}+2qA_kA_{k+1}-qA_k^2}, \quad \gamma_k = \frac{A_{k+1}-A_k}{1+qA_{k+1}} \\ y_k = x_k + \tau_k(z_k - x_k) \\ d_{y_k} = \operatorname{argmin}_p \langle \nabla f(y_k), p\rangle + \frac{L}{2}\|p\|_{\mathrm{F}}^2 + h(y_k+p) \\ x_{k+1} = y_k + d_{y_k} \\ z_{k+1} = (1-q\gamma_k)z_k + q\gamma_k y_k + \gamma_k d_k. \end{cases}$$

- Three-point iterative sequence: x_k, y_k and z_k;
- $\min\{O\left(\frac{1}{k^2}\right), O\left(1-\sqrt{q}\right)^k\}$ convergence rate for strongly convex f and convex h;

[dST+21] A. d'Aspremont, D. Scieur and A. Taylor. Acceleration methods. Foundations and Trends in Optimization, 5(1-2): 1–245, 2021.

Related Work

Euclidean Setting

FISTA in strongly convex [dST+21]:

$$\text{Given } x_0, \text{ let } z_0 = x_0, A_0 = 0, q = \frac{\mu}{L} \ (\mu \geq 0);$$

$$\begin{cases} A_{k+1} = \frac{2A_k+1+\sqrt{4A_k+4qA_k^2+1}}{2(1-q)} \\ \tau_k = \frac{(A_{k+1}-A_k)(1+qA_k)}{A_{k+1}+2qA_kA_{k+1}-qA_k^2}, \quad \gamma_k = \frac{A_{k+1}-A_k}{1+qA_{k+1}} \\ y_k = x_k + \tau_k(z_k - x_k) \\ d_{y_k} = \operatorname{argmin}_p \langle \nabla f(y_k), p\rangle + \frac{L}{2}\|p\|_{\mathrm{F}}^2 + h(y_k+p) \\ x_{k+1} = y_k + d_{y_k} \\ z_{k+1} = (1-q\gamma_k)z_k + q\gamma_k y_k + \gamma_k d_k. \end{cases}$$

- Three-point iterative sequence: x_k, y_k and z_k;
- $\min\{O\left(\frac{1}{k^2}\right), O\left(1-\sqrt{q}\right)^k\}$ convergence rate for strongly convex f and convex h;
- A unified accelerated method;

[dST+21] A. d'Aspremont, D. Scieur and A. Taylor. Acceleration methods. Foundations and Trends in Optimization, 5(1-2): 1–245, 2021.

Related Work

Riemannian Setting

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Proximal Gradient 3

How can these methods be extended to Riemannian manifolds?

Related Work

Riemannian Setting

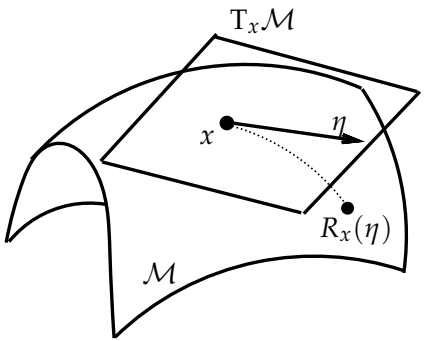
Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Proximal Gradient 3

[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.

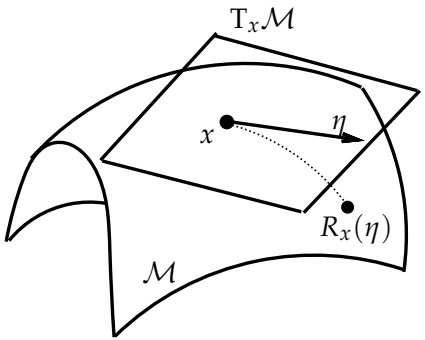
Related Work

Riemannian Setting

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Proximal Gradient 3



[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;
- For embedded submanifold;
- Solved by a semismooth Newton method;
- Any limit point is a critical point;
- 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.

Related Work

Riemannian Setting

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Proximal Gradient 3

[HW22b]: Given x_0,

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

[HW22b] W. Huang and K. Wei. Riemannian proximal gradient methods. Mathematical Programming, 194(1-2):371-413, 2022.

Related Work

Riemannian Setting

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Proximal Gradient 3

[HW22b]: Given x_0,

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

- 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;
- $O(1/k)$ rate for retraction convex f and h;
- Local convergence rate by Riemannian KL property;

[HW22b] W. Huang and K. Wei. Riemannian proximal gradient methods. Mathematical Programming, 194(1-2):371-413, 2022.

Related Work

Riemannian Setting

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Proximal Gradient 3

[BJJP25]: Given x_0,

$$\begin{cases} \text{Let } H_{x_k}(x) = h(x) + \frac{1}{2\lambda} d^2\left(x, R_{x_k}(-\lambda \operatorname{grad} f(x_k))\right); \\ x_{k+1} \text{ is a stationary point of } H_{x_k}(x); \\ \text{and } H_{x_k}(x_k) \geq H_{x_k}(x_{k+1}); \end{cases}$$

[BJJP25] R. Bergmann, H. Jasa, P. John, M. Pfeffer. The intrinsic Riemannian proximal gradient method for nonconvex optimization. arXiv:2506.09775, 2025.

Related Work

Riemannian Setting

Optimization with Structure:

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

- Proximal Gradient 1
- Proximal Gradient 2
- Proximal Gradient 3

[BJJP25]: Given x_0,

$$\begin{cases} \text{Let } H_{x_k}(x) = h(x) + \frac{1}{2\lambda} d^2\left(x, R_{x_k}(-\lambda \mathrm{grad} f(x_k))\right); \\ x_{k+1} \text{ is a stationary point of } H_{x_k}(x); \\ \text{and } H_{x_k}(x_k) \geq H_{x_k}(x_{k+1}); \end{cases}$$

- x_{k+1} can be viewed as a Riemannian proximal point of h on manifold;
- Sublinear convergence rate in nonconvex setting by Exponential map;
- Any limit point is a critical point by Exponential map;

[BJJP25] R. Bergmann, H. Jasa, P. John, M. Pfeffer. The intrinsic Riemannian proximal gradient method for nonconvex optimization. arXiv:2506.09775, 2025.

Riemannian Version of FISTA

Euclidean version:

[BT09] convex: Given x_0, let $y_0 = x_0, t_0 = 1$;

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

- Riemannian version 1
- Riemannian version 2

Riemannian Version of FISTA

Euclidean version:

[BT09] convex: Given x_0, let $y_0 = x_0, t_0 = 1$;

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

- Riemannian version 1
- Riemannian version 2

[HW22a], AManPG: Given x_0, let $y_0 = x_0, t_0 = 1$;

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

[HW22a] 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, 29(1): e2409, 2022.

Riemannian Version of FISTA

Euclidean version:

[BT09] convex: Given x_0, let $y_0 = x_0, t_0 = 1$;

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

- Riemannian version 1

[HW22b]: Given x_0, let $y_0 = x_0, t_0 = 1$;

- Riemannian version 2

$$\begin{cases} \ell_{y_k}(\eta) = \langle \operatorname{grad} f(y_k), \eta\rangle_{y_k} + \frac{L}{2}\|\eta\|_{y_k}^2 + h(R_{y_k}(\eta)) \\ \eta_{y_k} \text{ is a stationary point of } \ell_{y_k} \text{ and } \ell_{y_k}(0) \geq \ell_{y_k}(\eta_{y_k}) \\ x_{k+1} = R_{y_k}(\eta_{y_k}) \\ t_{k+1} = \frac{1+\sqrt{4t_k^2+1}}{2} \\ y_{k+1} = R_{y_k}\left(\frac{t_{k+1}+t_k-1}{t_{k+1}}\eta_{y_k} - \frac{t_k-1}{t_{k+1}}R_{y_k}^{-1}(x_k)\right). \end{cases}$$

[HW22a] 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, 29(1): e2409, 2022.

[HW22b] W. Huang and K. Wei. Riemannian proximal gradient methods. Mathematical Programming, 194(1-2):371-413,2022.

Riemannian Version of FISTA

Euclidean version:

[BT09] convex: Given x_0, let $y_0 = x_0, t_0 = 1$;

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

- Riemannian version 1
- Riemannian version 2

$$\frac{2}{\theta L}\left(t_k^2\left(F(x_{k+1}) - F(x_*)\right) - t_{k-1}^2\left(F(x_k) - F(x_*)\right)\right) \leqslant$$
$$\| \underbrace{(t_k - 1)R_{y_k}^{-1}(x_k) + R_{y_k}^{-1}(x_*)}_{\hat{W}_k} \|^2 - \| \underbrace{(t_k - 1)R_{y_k}^{-1}(x_k) + R_{y_k}^{-1}(x_*) - t_k \eta_{y_k}}_{\tilde{W}_{k+1}} \|^2$$

- $\hat{W}_k \neq \tilde{W}_k$ in general;
- How to control the difference?

[HW22a] 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, 29(1): e2409, 2022.

[HW22b] W. Huang and K. Wei. Riemannian proximal gradient methods. Mathematical Programming, 194(1-2):371-413,2022.

Riemannian Version of FISTA

Euclidean version:

[BT09] convex: Given x_0, let $y_0 = x_0, t_0 = 1$;

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

- Riemannian version 1
- Riemannian version 2

- Observe acceleration empirically;
- No theoretical guarantee for acceleration;

[HW22a] 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, 29(1): e2409, 2022.

[HW22b] W. Huang and K. Wei. Riemannian proximal gradient methods. Mathematical Programming, 194(1-2):371-413,2022.

Related Work

Riemannian Setting

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

In smooth case: $h = 0$, Riemannian Accelerated Gradient Methods
[LSC+17] [ZS18, AS20, JS22] [AOBL21] [MR22] [KY22] [MRP23] [CK25]

- [LSC+17]: Computationally hard;
- [ZS18, AS20, JS22]: Strongly convex, close to minimizer;
- [AOBL21]: convex, early stage;
- [MR22]: constant curvature;
- [KY22]: bound constraints;
- [MRP23]: Hadamard manifold;
- [CK25]: Rate independent of curvature;
- [CB22]: Negative curvature obstruct acceleration;

Related Work

Riemannian Setting

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

In smooth case: $h = 0$, Riemannian Accelerated Gradient Methods
[LSC+17] [ZS18, AS20, JS22] [AOBL21] [MR22] [KY22] [MRP23] [CK25]

[KY22] Given x_0, let $z_0 = x_0$;

$$\begin{cases} y_k = \mathrm{Exp}_{x_k}\left(\tau_k \mathrm{Exp}_{x_k}^{-1}(z_k)\right) \\ x_{k+1} = \mathrm{Exp}_{y_k}\left(-\alpha_k \operatorname{grad} f\left(y_k\right)\right) \\ v_{y_k} = \beta_k \mathrm{Exp}_{y_k}^{-1}(z_k) - \gamma_k \operatorname{grad} f(y_k) \\ z_{k+1} = \mathrm{Exp}_{x_{k+1}}\left(\Gamma_{y_k}^{x_{k+1}}\left(v_{y_k} - \mathrm{Exp}_{y_k}^{-1}(x_{k+1})\right)\right). \end{cases}$$

[KY22] J. Kim and I. Yang. Accelerated gradient methods for geodesically convex optimization: tractable algorithms and convergence analysis. PMLR, 162: 11255–11282, 2022.

Related Work

Riemannian Setting

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

In smooth case: $h = 0$, Riemannian Accelerated Gradient Methods
[LSC+17] [ZS18, AS20, JS22] [AOBL21] [MR22] [KY22] [MRP23] [CK25]

[KY22] Given x_0, let $z_0 = x_0$;

$$\begin{cases} y_k = \mathrm{Exp}_{x_k}\left(\tau_k \mathrm{Exp}_{x_k}^{-1}(z_k)\right) \\ x_{k+1} = \mathrm{Exp}_{y_k}\left(-\alpha_k \operatorname{grad} f\left(y_k\right)\right) \\ v_{y_k} = \beta_k \mathrm{Exp}_{y_k}^{-1}(z_k) - \gamma_k \operatorname{grad} f(y_k) \\ z_{k+1} = \mathrm{Exp}_{x_{k+1}}\left(\Gamma_{y_k}^{x_{k+1}}\left(v_{y_k} - \mathrm{Exp}_{y_k}^{-1}(x_{k+1})\right)\right). \end{cases}$$

- Accelerated convergence rates for geodesically convex f and geodesically strongly convex f, respectively;

[KY22] J. Kim and I. Yang. Accelerated gradient methods for geodesically convex optimization: tractable algorithms and convergence analysis. PMLR, 162: 11255–11282, 2022.

Related Work

Riemannian Setting

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

In smooth case: $h = 0$, Riemannian Accelerated Gradient Methods [LSC+17] [ZS18, AS20, JS22] [AOBL21] [MR22] [KY22] [MRP23] [CK25]

[KY22] Given x_0, let $z_0 = x_0$;

$$\begin{cases} y_k = \mathrm{Exp}_{x_k}\left(\tau_k \mathrm{Exp}_{x_k}^{-1}(z_k)\right) \\ x_{k+1} = \mathrm{Exp}_{y_k}\left(-\alpha_k \operatorname{grad} f\left(y_k\right)\right) \\ v_{y_k} = \beta_k \mathrm{Exp}_{y_k}^{-1}(z_k) - \gamma_k \operatorname{grad} f(y_k) \\ z_{k+1} = \mathrm{Exp}_{x_{k+1}}\left(\Gamma_{y_k}^{x_{k+1}}\left(v_{y_k} - \mathrm{Exp}_{y_k}^{-1}(x_{k+1})\right)\right). \end{cases}$$

- Accelerated convergence rates for geodesically convex f and geodesically strongly convex f, respectively;
- No unified parameters for accelerated gradient methods that works for both geodesically convex and geodesically strongly convex functions;

[KY22] J. Kim and I. Yang. Accelerated gradient methods for geodesically convex optimization: tractable algorithms and convergence analysis. PMLR, 162: 11255–11282, 2022.

Outline

1 Riemannian Proximal Gradient Methods Review

2 A Riemannian Accelerated Proximal Gradient Method
- The Proposed Approach
- Estimation of Convergence Rate

3 Adaptive Restart for Riemannian Accelerated Proximal Gradient Method

4 Summary

The Proposed Approach

Riemannian accelerated proximal gradient method (RAPG)

- Riemannian proximal mapping [HW22b];
- Nesterov's acceleration;
- A three-point iterative method;

The Proposed Approach

Riemannian accelerated proximal gradient method (RAPG)

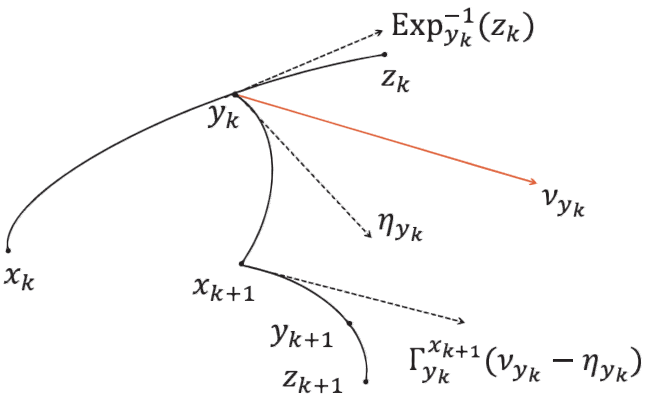
Initial iterate x_0, let $z_0 = x_0$;

1. $y_k = \mathrm{Exp}_{x_k}\left(\tau_k \mathrm{Exp}_{x_k}^{-1}(z_k)\right)$;
2. η_{y_k} is a stationary point of $\ell_{y_k}(\eta)$ on $\mathrm{T}_{y_k}\mathcal{M}$ with $\ell_{y_k}(0) \geqslant \ell_{y_k}\left(\eta_{y_k}\right)$, where $\ell_{y_k}(\eta) = \langle \mathrm{grad}\, f\left(y_k\right), \eta\rangle + \frac{\theta L}{2}\|\eta\|_{y_k}^2 + h\left(\mathrm{Exp}_{y_k}(\eta)\right)$;
3. $x_{k+1} = \mathrm{Exp}_{y_k}(\eta_{y_k})$;
4. $v_{y_k} = \beta_k \mathrm{Exp}_{y_k}^{-1}(z_k) + \gamma_k \eta_{y_k}, \quad z_{k+1} = \mathrm{Exp}_{x_{k+1}}\left(\Gamma_{y_k}^{x_{k+1}}\left(v_{y_k} - \eta_{y_k}\right)\right)$;

The Proposed Approach

Riemannian accelerated proximal gradient method (RAPG)
Initial iterate x_0, let $z_0 = x_0$;

1. $y_k = \mathrm{Exp}_{x_k}\left(\tau_k \mathrm{Exp}_{x_k}^{-1}(z_k)\right)$;
2. η_{y_k} is a stationary point of $\ell_{y_k}(\eta)$ on $\mathrm{T}_{y_k}\mathcal{M}$ with $\ell_{y_k}(0) \geqslant \ell_{y_k}\left(\eta_{y_k}\right)$, where $\ell_{y_k}(\eta) = \langle \mathrm{grad}\, f\left(y_k\right), \eta\rangle + \frac{\theta L}{2}\|\eta\|_{y_k}^2 + h\left(\mathrm{Exp}_{y_k}(\eta)\right)$;
3. $x_{k+1} = \mathrm{Exp}_{y_k}(\eta_{y_k})$;
4. $v_{y_k} = \beta_k \mathrm{Exp}_{y_k}^{-1}(z_k) + \gamma_k \eta_{y_k}, \quad z_{k+1} = \mathrm{Exp}_{x_{k+1}}\left(\Gamma_{y_k}^{x_{k+1}}\left(v_{y_k} - \eta_{y_k}\right)\right)$;

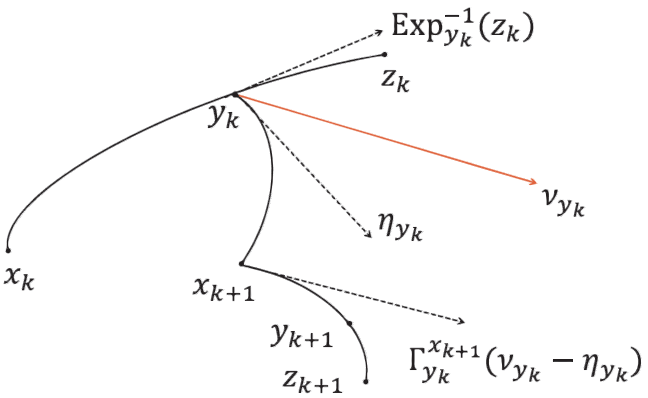


1. Step 1: compute y_k; note that x_k, y_k and z_k are on a geodesic;

The Proposed Approach

Riemannian accelerated proximal gradient method (RAPG)
Initial iterate x_0, let $z_0 = x_0$;

1. $y_k = \mathrm{Exp}_{x_k}\left(\tau_k \mathrm{Exp}_{x_k}^{-1}(z_k)\right)$;
2. η_{y_k} is a stationary point of $\ell_{y_k}(\eta)$ on $\mathrm{T}_{y_k}\mathcal{M}$ with $\ell_{y_k}(0) \geqslant \ell_{y_k}\left(\eta_{y_k}\right)$, where $\ell_{y_k}(\eta) = \langle \mathrm{grad}\, f\left(y_k\right), \eta\rangle + \frac{\theta L}{2}\|\eta\|_{y_k}^2 + h\left(\mathrm{Exp}_{y_k}(\eta)\right)$;
3. $x_{k+1} = \mathrm{Exp}_{y_k}(\eta_{y_k})$;
4. $v_{y_k} = \beta_k \mathrm{Exp}_{y_k}^{-1}(z_k) + \gamma_k \eta_{y_k}, \quad z_{k+1} = \mathrm{Exp}_{x_{k+1}}\left(\Gamma_{y_k}^{x_{k+1}}\left(v_{y_k} - \eta_{y_k}\right)\right)$;

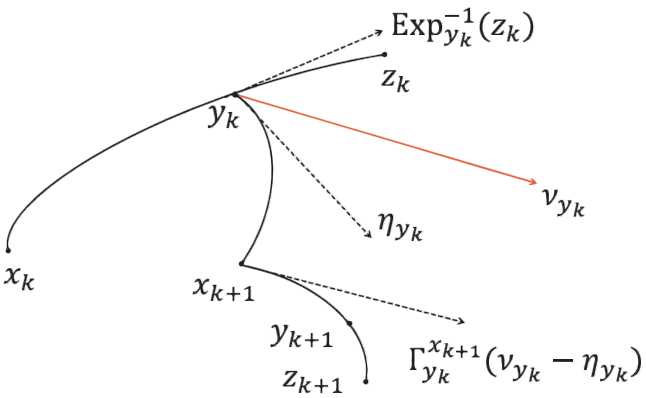


1. Step 1: compute y_k; note that x_k, y_k and z_k are on a geodesic;
2. Step 2: compute a Riemannian proximal gradient direction η_{y_k};

The Proposed Approach

Riemannian accelerated proximal gradient method (RAPG)
Initial iterate x_0, let $z_0 = x_0$;

1. $y_k = \mathrm{Exp}_{x_k}\left(\tau_k \mathrm{Exp}_{x_k}^{-1}(z_k)\right)$;
2. η_{y_k} is a stationary point of $\ell_{y_k}(\eta)$ on $\mathrm{T}_{y_k}\mathcal{M}$ with $\ell_{y_k}(0) \geqslant \ell_{y_k}\left(\eta_{y_k}\right)$, where $\ell_{y_k}(\eta) = \langle \mathrm{grad}\, f\left(y_k\right), \eta\rangle + \frac{\theta L}{2}\|\eta\|_{y_k}^2 + h\left(\mathrm{Exp}_{y_k}(\eta)\right)$;
3. $x_{k+1} = \mathrm{Exp}_{y_k}(\eta_{y_k})$;
4. $v_{y_k} = \beta_k \mathrm{Exp}_{y_k}^{-1}(z_k) + \gamma_k \eta_{y_k}, \quad z_{k+1} = \mathrm{Exp}_{x_{k+1}}\left(\Gamma_{y_k}^{x_{k+1}}\left(v_{y_k} - \eta_{y_k}\right)\right)$;

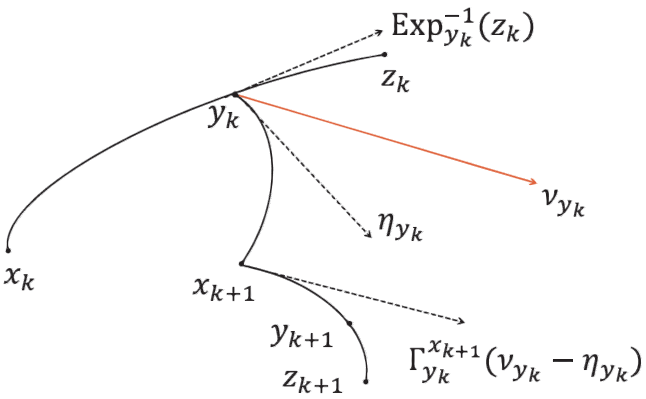


1. Step 1: compute y_k; note that x_k, y_k and z_k are on a geodesic;
2. Step 2: compute a Riemannian proximal gradient direction η_{y_k};
3. Step 3: update x_{k+1} by exponential map;

The Proposed Approach

Riemannian accelerated proximal gradient method (RAPG)
Initial iterate x_0, let $z_0 = x_0$;

1. $y_k = \mathrm{Exp}_{x_k}\left(\tau_k \mathrm{Exp}_{x_k}^{-1}(z_k)\right)$;
2. η_{y_k} is a stationary point of $\ell_{y_k}(\eta)$ on $\mathrm{T}_{y_k}\mathcal{M}$ with $\ell_{y_k}(0) \geqslant \ell_{y_k}\left(\eta_{y_k}\right)$, where $\ell_{y_k}(\eta) = \langle \mathrm{grad}\, f\left(y_k\right), \eta\rangle + \frac{\theta L}{2}\|\eta\|_{y_k}^2 + h\left(\mathrm{Exp}_{y_k}(\eta)\right)$;
3. $x_{k+1} = \mathrm{Exp}_{y_k}(\eta_{y_k})$;
4. $v_{y_k} = \beta_k \mathrm{Exp}_{y_k}^{-1}(z_k) + \gamma_k \eta_{y_k}, \quad z_{k+1} = \mathrm{Exp}_{x_{k+1}}\left(\Gamma_{y_k}^{x_{k+1}}\left(v_{y_k} - \eta_{y_k}\right)\right)$;



1. Step 1: compute y_k; note that x_k, y_k and z_k are on a geodesic;
2. Step 2: compute a Riemannian proximal gradient direction η_{y_k};
3. Step 3: update x_{k+1} by exponential map;
4. Step 4: update z_{k+1} by exponential map and parallel transport;

The Proposed Approach

Riemannian accelerated proximal gradient method (RAPG)
Initial iterate x_0, let $z_0 = x_0$;

1. $y_k = \mathrm{Exp}_{x_k}\left(\tau_k \mathrm{Exp}_{x_k}^{-1}(z_k)\right)$;
2. η_{y_k} is a stationary point of $\ell_{y_k}(\eta)$ on $\mathrm{T}_{y_k}\mathcal{M}$ with $\ell_{y_k}(0) \geqslant \ell_{y_k}\left(\eta_{y_k}\right)$, where $\ell_{y_k}(\eta) = \langle \mathrm{grad}\, f\left(y_k\right), \eta\rangle + \frac{\theta L}{2}\|\eta\|_{y_k}^2 + h\left(\mathrm{Exp}_{y_k}(\eta)\right)$;
3. $x_{k+1} = \mathrm{Exp}_{y_k}(\eta_{y_k})$;
4. $v_{y_k} = \beta_k \mathrm{Exp}_{y_k}^{-1}(z_k) + \gamma_k \eta_{y_k}, \quad z_{k+1} = \mathrm{Exp}_{x_{k+1}}\left(\Gamma_{y_k}^{x_{k+1}}\left(v_{y_k} - \eta_{y_k}\right)\right)$;

Next, we will show:

1. Assumptions on Manifolds and functions;
2. Parameter expressions for τ_k, β_k, γ_k;
3. Convergence rate of RAPG;

Assumptions on Manifolds and Functions

Assumption on Manifold:

1. Let Ω be a geodesically uniquely convex subset of $\mathcal{M}$. The diameter of Ω satisfies $\operatorname{diam}(\Omega) \leqslant D < \infty$;
2. The sectional curvature of Ω is bounded below by $\kappa_{\min}$ and above by $\kappa_{\max}$. If $\kappa_{\max} > 0$, it is additionally assumed that $D < \frac{\pi}{\sqrt{\kappa_{\max}}}$;

Assumptions on Manifolds and Functions

Assumption on Manifold:

1. Let Ω be a geodesically uniquely convex subset of $\mathcal{M}$. The diameter of Ω satisfies $\operatorname{diam}(\Omega) \leqslant D < \infty$;
2. The sectional curvature of Ω is bounded below by $\kappa_{\min}$ and above by $\kappa_{\max}$. If $\kappa_{\max} > 0$, it is additionally assumed that $D < \frac{\pi}{\sqrt{\kappa_{\max}}}$;

For the eigenvalues of the Hessian matrix of the squared distance function $\frac{1}{2}\mathrm{d}^2(\cdot, p)$ on $\Omega \subset \mathcal{M}$, where $p \in \Omega$:

- the upper bound:
$$\zeta = \begin{cases} \sqrt{-\kappa_{\min}} D \coth\left(\sqrt{-\kappa_{\min}} D\right), & \text{if } \kappa_{\min} < 0 \\ 1, & \text{if } \kappa_{\min} \geqslant 0 \end{cases}$$
- the lower bound:
$$\delta = \begin{cases} 1, & \text{if } \kappa_{\max} \leqslant 0 \\ \sqrt{\kappa_{\max}} D \cot\left(\sqrt{\kappa_{\max}} D\right), & \text{if } \kappa_{\max} > 0 \end{cases}$$

Assumptions on Manifolds and Functions

Assumption on Manifold:

1. Let Ω be a geodesically uniquely convex subset of $\mathcal{M}$. The diameter of Ω satisfies $\operatorname{diam}(\Omega) \leqslant D < \infty$;
2. The sectional curvature of Ω is bounded below by $\kappa_{\min}$ and above by $\kappa_{\max}$. If $\kappa_{\max} > 0$, it is additionally assumed that $D < \frac{\pi}{\sqrt{\kappa_{\max}}}$;

For the eigenvalues of the Hessian matrix of the squared distance function $\frac{1}{2}\mathrm{d}^2(\cdot, p)$ on $\Omega \subset \mathcal{M}$, where $p \in \Omega$:

- the upper bound:
$$\zeta = \begin{cases} \sqrt{-\kappa_{\min}} D \coth\left(\sqrt{-\kappa_{\min}} D\right), & \text{if } \kappa_{\min} < 0 \\ 1, & \text{if } \kappa_{\min} \geqslant 0 \end{cases}$$
- the lower bound:
$$\delta = \begin{cases} 1, & \text{if } \kappa_{\max} \leqslant 0 \\ \sqrt{\kappa_{\max}} D \cot\left(\sqrt{\kappa_{\max}} D\right), & \text{if } \kappa_{\max} > 0 \end{cases}$$

Choose $\xi \geqslant \zeta$.

Assumptions on Manifolds and Functions

Assumption on Manifold:

1. Let Ω be a geodesically uniquely convex subset of $\mathcal{M}$. The diameter of Ω satisfies $\operatorname{diam}(\Omega) \leqslant D < \infty$;
2. The sectional curvature of Ω is bounded below by $\kappa_{\min}$ and above by $\kappa_{\max}$. If $\kappa_{\max} > 0$, it is additionally assumed that $D < \frac{\pi}{\sqrt{\kappa_{\max}}}$;

Assumption on functions:

1. The function f is geodesically L-smooth and geodesically μ-strongly convex ($\mu \geqslant 0$) in Ω;
2. The function h is ρ-retraction-convex with respect to the exponential map in Ω;

Assumptions on Manifold and Functions

ρ-retraction-convex:

$\tilde{h}_x(\eta) = h(R_x(\eta)) + \frac{\rho}{2}\|\eta\|^2$ is convex in tangent space.

- $\rho > 0$, h is said to be ρ-weakly retraction-convex with respect to R;
- $\rho = 0$, h is said to be retraction-convex with respect to R;
- $\rho < 0$, h is said to be ρ-strongly retraction-convex with respect to R.

Assumptions on Manifold and Functions

ρ-retraction-convex:

$\tilde{h}_x(\eta) = h(R_x(\eta)) + \frac{\rho}{2}\|\eta\|^2$ is convex in tangent space.

- $\rho > 0$, h is said to be ρ-weakly retraction-convex with respect to R;
- $\rho = 0$, h is said to be retraction-convex with respect to R;
- $\rho < 0$, h is said to be ρ-strongly retraction-convex with respect to R.

Weakly Retraction-Convex: A Necessary Assumption

e.g. $\|x\|_1$ is locally weakly retraction-convex on the embedded submanifold of $\mathbb{R}^n$.

Parameter Expressions for β_k, γ_k, τ_k

Under assumptions on manifold and functions:

$$A_{k+1} = \frac{\xi + 2\xi A_k + \sqrt{\xi^2 + 4\xi^2 A_k + 4\frac{\mu-\rho}{\theta L-\rho}\xi A_k^2}}{2\left(\xi - \frac{\mu-\rho}{\theta L-\rho}\right)},$$

$$\beta_k = \frac{\xi(\theta L-\rho) + (\mu-\rho)A_k}{\xi(\theta L-\rho) + (\mu-\rho)A_{k+1}}, \gamma_k = \frac{(\theta L-\rho)(A_{k+1}-A_k)}{\xi(\theta L-\rho) + (\mu-\rho)A_{k+1}}, \tau_k = \frac{\beta_k A_{k+1}}{\gamma_k A_k + \beta_k A_{k+1}};$$

Parameter Expressions for β_k, γ_k, τ_k

Under assumptions on manifold and functions:

$$A_{k+1} = \frac{\xi + 2\xi A_k + \sqrt{\xi^2 + 4\xi^2 A_k + 4\frac{\mu-\rho}{\theta L-\rho}\xi A_k^2}}{2\left(\xi - \frac{\mu-\rho}{\theta L-\rho}\right)},$$

$$\beta_k = \frac{\xi(\theta L-\rho) + (\mu-\rho)A_k}{\xi(\theta L-\rho) + (\mu-\rho)A_{k+1}}, \gamma_k = \frac{(\theta L-\rho)(A_{k+1}-A_k)}{\xi(\theta L-\rho) + (\mu-\rho)A_{k+1}}, \tau_k = \frac{\beta_k A_{k+1}}{\gamma_k A_k + \beta_k A_{k+1}};$$

Reduce to Euclidean space:

- if $\xi = 1$, $\rho = 0$, RAPG is FISTA in strongly convex [dST+21];
- otherwise, it is new as far as we known;

On manifold:

- Our parameter settings apply to both convex and strongly convex cases on manifold, leading to a unified accelerated algorithm.

Convergence Rate of RAPG

Under assumptions on manifold and functions:

- Sublinear convergence for $\mu \geqslant \rho$: $O\left(\frac{1}{k^2}\right)$;
- Linear convergence for $\mu > \rho$:

$$\min\left\{\left(1-\sqrt{\frac{\mu-\rho}{(\theta L-\rho)\xi}}\right)^k C_1,\ \frac{2}{\left(k+2\sqrt{A_0}\right)^2}C_2\right\}.$$

Assumption on functions:

1. The function f is geodesically L-smooth and geodesically μ-strongly convex ($\mu \geqslant 0$) in Ω;
2. The function h is ρ-retraction-convex with respect to the exponential map in Ω;

$$F(x) = f(x) + h(x)$$

Convergence Rate of RAPG

Sketch of the analysis

The core of our analysis is the construction of a potential function (or Lyapunov function) Φ_k that combines:

1. the function value gap;
2. the distance from the iterate to the optimal point; and
3. distortion error from curvature;

$$\Phi_k = A_k(F(x_k) - F(x_*)) + \frac{\xi(\theta L - \rho) + (\mu - \rho)A_k}{2}\left(\left\|\mathrm{Exp}_{x_k}^{-1}(z_k) - \mathrm{Exp}_{x_k}^{-1}(x_*)\right\|^2 + (\xi - 1)\left\|\mathrm{Exp}_{x_k}^{-1}(z_k)\right\|^2\right)$$

Convergence Rate of RAPG

Sketch of the analysis

The core of our analysis is the construction of a potential function (or Lyapunov function) Φ_k that combines:

1. the function value gap;
2. the distance from the iterate to the optimal point; and
3. distortion error from curvature;

$$\Phi_k = A_k(F(x_k) - F(x_*)) + \frac{\xi(\theta L - \rho) + (\mu - \rho)A_k}{2}\left(\left\|\mathrm{Exp}_{x_k}^{-1}(z_k) - \mathrm{Exp}_{x_k}^{-1}(x_*)\right\|^2 + (\xi - 1)\left\|\mathrm{Exp}_{x_k}^{-1}(z_k)\right\|^2\right)$$

A convergence rate of $O(1/A_k)$ is achieved if $\Phi_{k+1} \leqslant \Phi_k$ is satisfied.

The limit of RAPG:

- RAPG is theoretically supported only under the convexity of both f and h on manifolds;
- What happens in the nonconvex case?

We develop an improved version of the method.

Outline

1 Riemannian Proximal Gradient Methods Review

2 A Riemannian Accelerated Proximal Gradient Method

3 Adaptive Restart for Riemannian Accelerated Proximal Gradient Method
- Algorithm
- Convergence Analysis

4 Summary

Adaptive Restart for RAPG

Adaptive Restart for Riemannian Accelerated Proximal Gradient Method (AR-RAPG)

1: Set $z_0 = x_0$, $\tilde{x}_0 = x_0$, $\theta \geqslant 1$, $L = L_{\text{init}}$, $i = 0$, and $j = N_0$;
2: for $k = 0, 1, 2, \cdots$ do
3: if $k == j$ then
4: $[\tilde{x}_{i+1}, x_k, z_k, A_k, N_{i+1}, L] = \text{Safeguard}(\tilde{x}_i, x_k, z_k, A_k, N_i, L)$;
5: Set $j = j + N_{i+1}$ and $i = i + 1$;
6: end if
7: $(A_{k+1}, \beta_k, \gamma_k, \tau_k)$ are derived from the same formulas as in RAPG;
8: Compute y_k, x_{k+1}, z_{k+1} as in RAPG;
9: end for

Adaptive Restart for RAPG

Adaptive Restart for Riemannian Accelerated Proximal Gradient Method (AR-RAPG)

1: Set $z_0 = x_0$, $\tilde{x}_0 = x_0$, $\theta \geqslant 1$, $L = L_{\text{init}}$, $i = 0$, and $j = N_0$;
2: for $k = 0, 1, 2, \cdots$ do
3: if $k == j$ then
4: $[\tilde{x}_{i+1}, x_k, z_k, A_k, N_{i+1}, L] = \text{Safeguard}(\tilde{x}_i, x_k, z_k, A_k, N_i, L)$;
5: Set $j = j + N_{i+1}$ and $i = i + 1$;
6: end if
7: $(A_{k+1}, \beta_k, \gamma_k, \tau_k)$ are derived from the same formulas as in RAPG;
8: Compute y_k, x_{k+1}, z_{k+1} as in RAPG;
9: end for

- Safeguard strategy from [HW22b];
- The functions f and h are not required to be convex on manifold;
- If the convexity of the functions is not known, we simply set $\mu = 0$ and $\rho = 0$;

[HW22b] W. Huang and K. Wei. Riemannian proximal gradient methods. Mathematical Programming, 194(1-2):371-413,2022.

Adaptive Restart for RAPG

Safeguard

Input: $(\tilde{x}_i, x_k, z_k, A_k, N_i, L)$;
Output: $[\tilde{x}_{i+1}, x_k, z_k, A_k, N_{i+1}, L]$;

$\eta_{\tilde{x}_i}$ is a stationary point of $\ell_{\tilde{x}_i}(\eta)$ on $\mathrm{T}_{x_i}\mathcal{M}$ with $\ell_{\tilde{x}_i}(0) \geqslant \ell_{\tilde{x}_i}\left(\eta_{\tilde{x}_i}\right)$;
Set $\alpha_i = 1,\ i_{ls} = 0$;
while $F(\mathrm{Exp}_{\tilde{x}_i}(\alpha_i \eta_{\tilde{x}_i})) > F(\tilde{x}_i) - \sigma\alpha_i \|\eta_{\tilde{x}_i}\|^2$ and $i_{ls} < N_{ls}$ do
    $\alpha_i = \rho\alpha_i,\ i_{ls} = i_{ls} + 1$;
end while
if $i_{ls} == N_{ls}$ then
    $L = \tau L$ and go to Step 1; The estimation of L is too small
end if
if $F(\mathrm{Exp}_{\tilde{x}_i}(\alpha_i \eta_{\tilde{x}_i}) < F(x_k)$ then
    Safeguard takes effect
    if $N_i \neq N_{\max}$ then
        $L = \tau L$;
    end if
    $x_k = \mathrm{Exp}_{\tilde{x}_i}(\alpha_i \eta_{\tilde{x}_i}),\ z_k = x_k,\ A_k = A_0$; {Restart}
    $N_{i+1} = \max\{N_i - 1, N_{\min}\}$;
else
    x_k, z_k, and A_k keep unchanged; No restart
    $N_{i+1} = \min\{N_i + 1, N_{\max}\}$;
end if
$\tilde{x}_{i+1} = x_k$.

- Adaptively update the smoothness parameter L;
- Guarantee a decrease in the function value after a finite number of iterations;

Adaptive Restart for RAPG

Theorem (Convergence)

Under assumptions of Manifolds, if

1. Ω is compact;
2. all iterates remian in Ω;
3. f is smooth, h is locally Lipschitz continuous,

then any accumulation point $\tilde{x}_*$ of the sequence $\{\tilde{x}_i\}$ generated by AR-RAPG is a stationary point.

Numerical Experiments

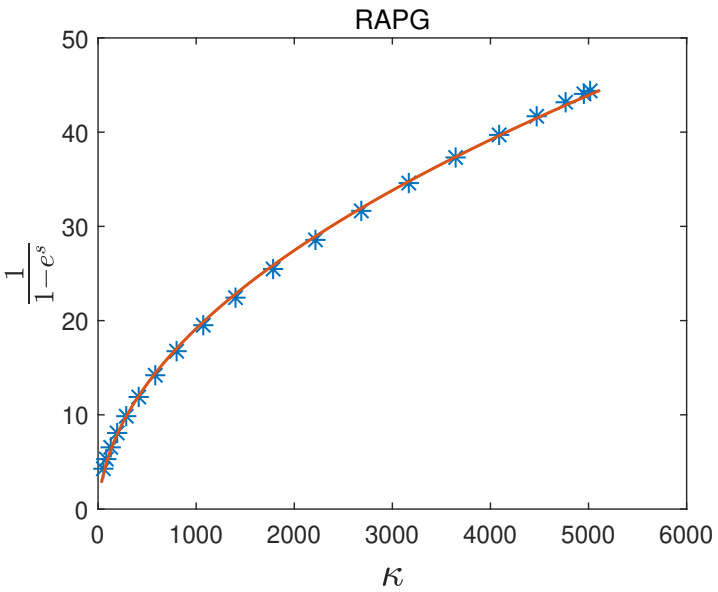
Convergence rate verification of RAPG and RPG

$$\min_{x \in \mathbb{S}^{n-1}} F(x) = \underbrace{-x^T A^T A x}_{f_1(x)} + \underbrace{\lambda \|x\|_1}_{h(x)},$$

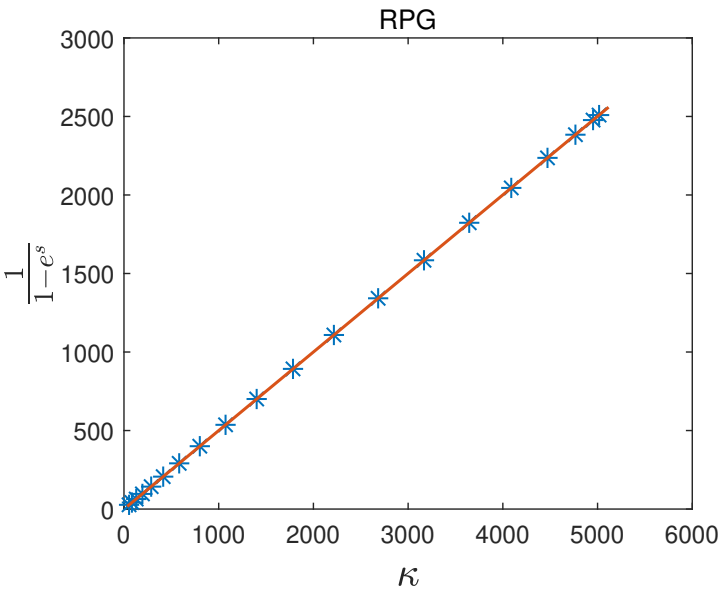
- $A = USV^T + e$;
- $S \in \mathbb{R}^{m \times n}$: first m columns are $\mathrm{diag}(m + c, m, m-1, \cdots, 2)$ with c varying from 0.01 to 1, and the remaining columns are zero;
- e is a small noise;

Numerical Experiments

Convergence rate verification of RAPG and RPG



(a) RAPG



(b) RPG

Figure: Empirical relationship between κ and $\frac{1}{1-e^s}$ for RAPG and RPG. $m = 20, n = 1000, \lambda = 10^{-4}$.

Numerical Experiments

Effectiveness of the safeguard in AR-RAPG

$$\min_{X \in \mathrm{OB}(p,n)} F(X) = \underbrace{\|X^T A^T A X - D^2\|_F^2}_{f_2(X)} + \underbrace{\lambda \|X\|_1}_{h(X)}$$

- Oblique manifold:
  $\mathrm{OB}(p,n) = \{X \in \mathbb{R}^{n \times p} \mid x_i^T x_i = 1,\ i = 1, \cdots, p\}$;
- Entries of A: standard normal distribution $\mathcal{N}(0,1)$;
- Each column of A: zero mean and unit 2-norm;

Numerical Experiments

Effectiveness of the safeguard in AR-RAPG

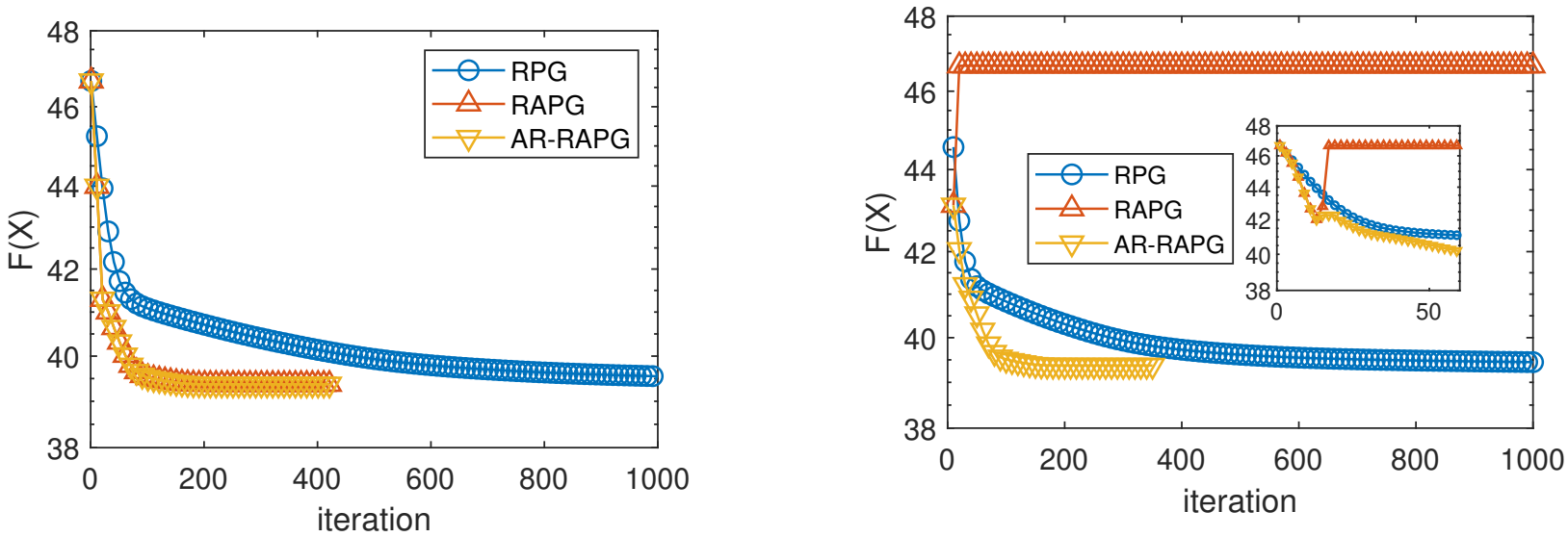


Figure: Comparison of RPG, RAPG, and AR-RAPG for the SPCA problem on oblique manifold. $\lambda = 1$, $m = 20$, $n = 200$, $p = 4$. Left: $L = 2\|D^2\|_F^2$; Right: $L = 1.2\|D^2\|_F^2$.

Numerical Experiments

Sparse PCA problem:

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

- $\mathrm{OB}(p,n) = \{X \in \mathbb{R}^{n \times p} \mid x_i^T x_i = 1, i = 1, \ldots, p\}$ denotes the oblique manifold;
- x_i being the i-th column of X;
- $A \in \mathbb{R}^{m \times n}$ is the data matrix and $p \leq m$;
- D is a diagonal matrix with the dominant singular values of A on the diagonal;

Compared with:

- ManPG, ManPG-Ada: in [CMSZ20];
- RPG: in [HW22b];

Numerical Experiments

Left: the norm of search direction;
Right: function value.

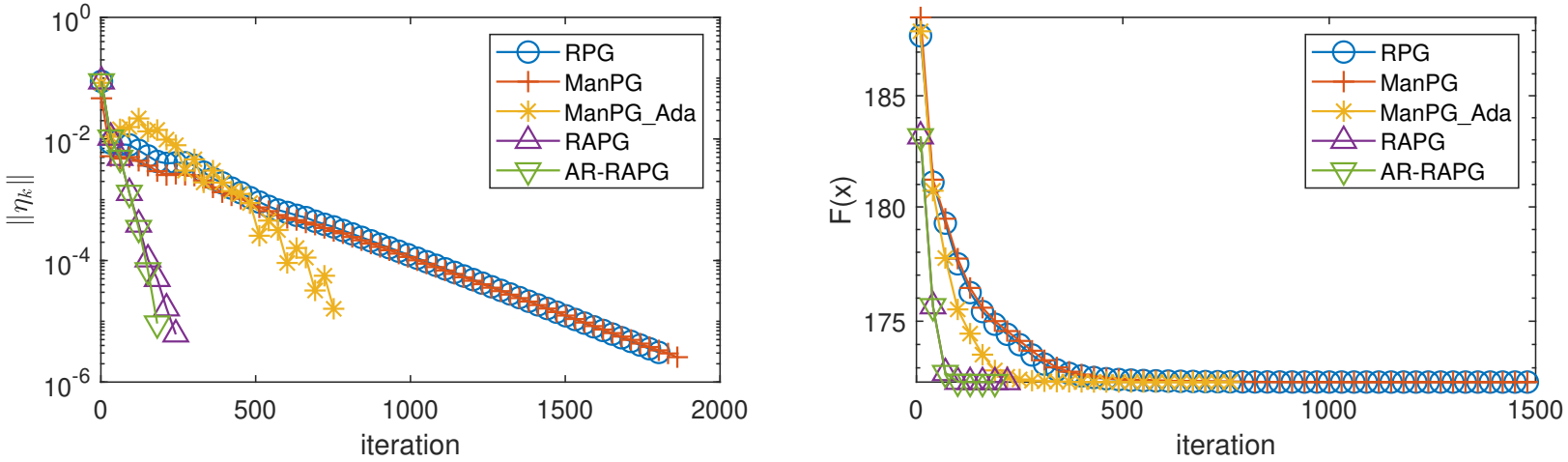


Figure: SPCA problem on oblique manifold. $n = 200,\ m = 20,\ p = 4$.

Numerical Experiments

For $m = 20,\ p = 4,\ n = \{32, 64, 128, 256\}$.
Left: number of iterations;
Right: CPU time.

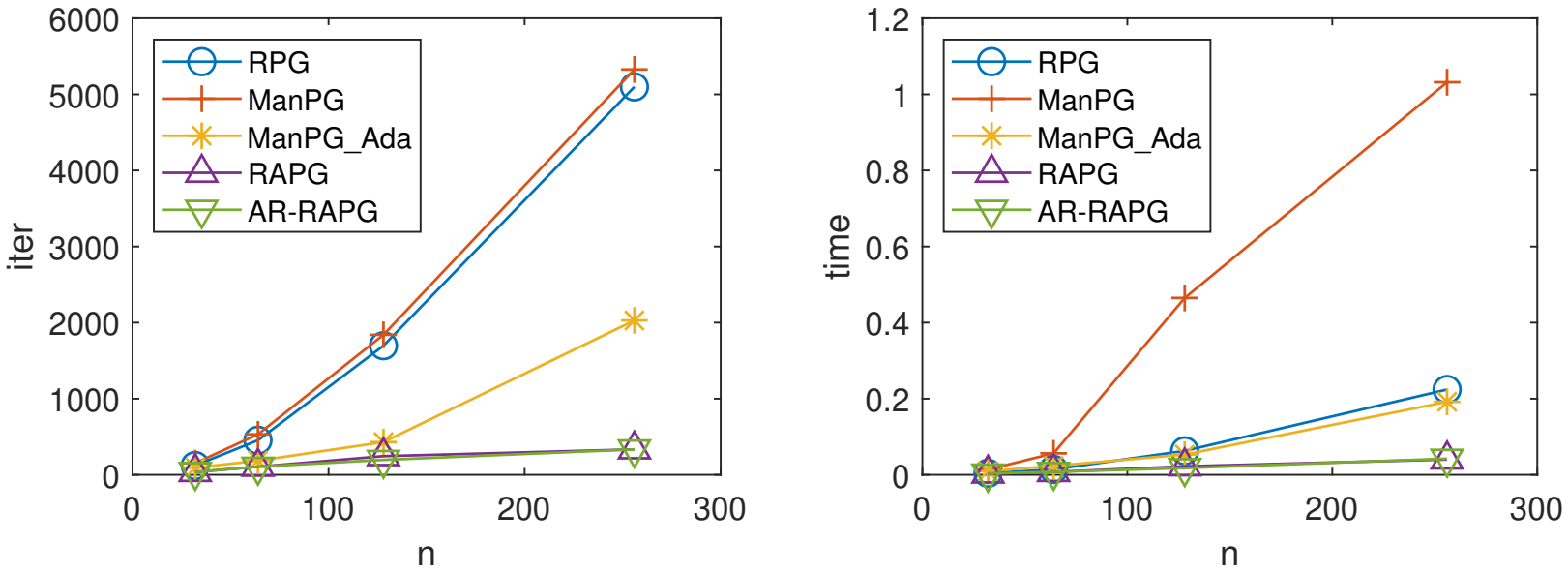


Figure: SPCA problem on oblique manifold.

Numerical Experiments

For $m = 20,\ n = 128,\ p = \{1, 2, 3, 4\}$.
Left: number of iterations;
Right: CPU time.

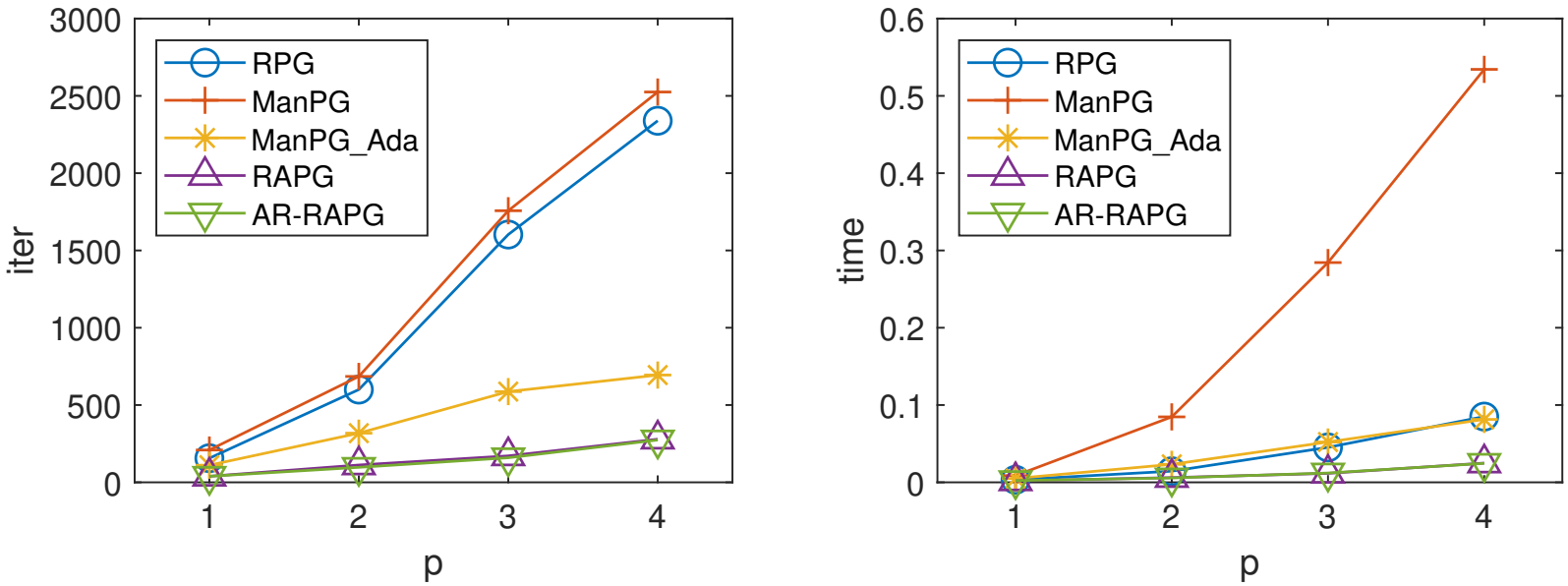


Figure: SPCA problem on oblique manifold.

Outline

1. Riemannian Proximal Gradient Methods Review
2. A Riemannian Accelerated Proximal Gradient Method
3. Adaptive Restart for Riemannian Accelerated Proximal Gradient Method
4. Summary

Summary

- Propose a Riemannian accelerated proximal gradient method;
- Accelerated convergence rate is proven;
- Propose an adaptive restart Riemannian accelerated proximal gradient method;
- Global convergence to critical points is proven;
- Numerical experiments show its performance;

ArXiv preprint: S. Feng, Y. Jiang, W. Huang*, S. Ying, A Riemannian Accelerated Proximal Gradient Method, arXiv:2509.21897, 2025.

Thank you!

References I

Foivos Alimisis, Antonio Orvieto, Gary Becigneul, and Aurelien Lucchi.

Momentum improves optimization on Riemannian manifolds.
In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1351–1359, 2021.

Kwangjun Ahn and Suvrit Sra.

From Nesterov's estimate sequence to Riemannian acceleration.
In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 84–118, 2020.

Ronny Bergmann, Hajg Jasa, Paula John, and Max Pfeffer.

The intrinsic riemannian proximal gradient method for nonconvex optimization.
arXiv preprint arXiv:2506.09775, 2025.

Amir Beck and Marc Teboulle.

A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM journal on imaging sciences, 2(1):183–202, 2009.

Christopher Criscitiello and Nicolas Boumal.

Negative curvature obstructs acceleration for strongly geodesically convex optimization, even with exact first-order oracles.
In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 496–542, 2022.

Christopher Criscitiello and Jungbin Kim.

Horospherically convex optimization on Hadamard manifolds part I: Analysis and algorithms.
arXiv preprint arXiv:2505.16970, 2025.

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.

References II

Alexandre d'Aspremont, Damien Scieur, Adrien Taylor, et al.

Acceleration methods.
Foundations and Trends® in Optimization, 5(1-2):1–245, 2021.

Yunfang Fu, Qiuqi Ruan, Yi Jin, and Gaoyun An.

Sparse orthogonal tucker decomposition for 2d+3d facial expression recognition.
In *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pages 516–521, 2018.

Matthieu Genicot, Wen Huang, and Nickolay T Trendafilov.

Weakly correlated sparse components with nearly orthonormal loadings.
In *Geometric Science of Information*, pages 484–490, Cham, 2015. Springer International Publishing.

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, 29(1):e2409, 2022.

Wen Huang and Ke Wei.

Riemannian proximal gradient methods.
Mathematical Programming, 194(1-2):371–413, 2022.

Wen Huang, Meng Wei, Kyle A Gallivan, and Paul Van Dooren.

A Riemannian optimization approach to clustering problems.
Journal of Scientific Computing, 103(1):8, 2025.

Jikai Jin and Suvrit Sra.

Understanding Riemannian acceleration via a proximal extragradient framework.
In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2924–2962, 2022.

References III

Jungbin Kim and Insoon Yang.

Accelerated gradient methods for geodesically convex optimization: tractable algorithms and convergence analysis.
In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11255–11282, 2022.

Yuan Yuan Liu, Fan Hua Shang, James Cheng, Hong Cheng, and Licheng Jiao.

Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds.
In *Advances in Neural Information Processing Systems*, pages 4868–4877, 2017.

Canyi Lu, Shuicheng Yan, and Zhouchen Lin.

Convex sparse spectral clustering: Single-view to multi-view.
IEEE Transactions on Image Processing, 25(6):2833–2843, 2016.

David Martínez-Rubio.

Global Riemannian acceleration in hyperbolic and spherical spaces.
In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 768–826, 2022.

David Martínez-Rubio and Sebastian Pokutta.

Accelerated Riemannian optimization: Handling constraints with a prox to bound geometric penalties.
In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 359–393, 2023.

Vidvuds Ozoliņš, 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.

S. Osher, Zuoqiang Shi, and Wei Zhu.

Low dimensional manifold model for image processing.
SIAM Journal on Imaging Sciences, 10(4):1669–1690, 2017.

References IV

Seyoung Park and Hongyu Zhao.

Spectral clustering based on learning similarity matrix.
Bioinformatics, 34(12):2069–2076, 2018.

Ganzhao Yuan and Bernard Ghanem.

ℓ_0TV: A sparse optimization method for impulse noise image restoration.
IEEE transactions on pattern analysis and machine intelligence, 41(2):352–364, 2017.

Yuhao Zhou, Chenglong Bao, Chao Ding, and Jun Zhu.

A semismooth Newton based augmented lagrangian method for nonsmooth optimization on matrix manifolds.
Mathematical Programming, 201(1):1–61, 2023.

Hui Zou, Trevor Hastie, and Robert Tibshirani.

Sparse principal component analysis.
Journal of computational and graphical statistics, 15(2):265–286, 2006.

Hongyi Zhang and Suvrit Sra.

An estimate sequence for geodesically convex optimization.
In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1703–1723, 2018.