Convergence and complexity of Block Majorization-Minimization on Riemannian manifolds

Yuchen Li

Department of mathematics, University of Wisconsin-Madison

April 10, 2023, IFDS

Joint work with Hanbaek Lyu, Laura Balzano and Deanna Needell

Outline

Introduction

Statement of results

Examples

Numerical experiments

Problem Set-up

- **Problem Set-up**
  - (Objective function) $f : \mathcal{M}^{(1)} \times \cdots \times \mathcal{M}^{(m)} \to \mathbb{R}$ — geodesically smooth in each block
  - (Constraint Sets) $\boldsymbol{\Theta} = \Theta^{(1)} \times \cdots \times \Theta^{(m)} \subseteq \mathcal{M}^{(1)} \times \cdots \times \mathcal{M}^{(m)}$ — $\mathcal{M}^{(i)}$ complete Riemannian manifold, $\boldsymbol{\Theta}^{(i)}$ geodesically convex for rate of convergence
  - (Constrained nonconvex problem)

$$\boldsymbol{\theta}^* \in \operatorname*{arg\,min}_{\boldsymbol{\theta}=[\theta_1,\ldots,\theta_m]\in\boldsymbol{\Theta}} f(\theta_1, \ldots, \theta_m).$$

Related works

- Related works
  - (Euclidean BMM) Rate of convergence for convex problem is $\widetilde{O}(\varepsilon^{-1})$ ([HRLP15]).
  - (Riemannian MM) Rate of convergence for certain type of majorizer on specific manifolds:

    (i) Majorizer on manifolds:
    - Linear majorizer on Stiefel manifolds [BKSP21]
    - Proximal majorizer on Hadamard manifolds [BFO15]

    (ii) Majorizer on tangent spaces:
    - Tangent prox-linear on Stiefel manifolds ([CMMCSZ20])
    - Tangent prox-linear on Riemannian manifolds [HW22] (assuming retraction convexity)

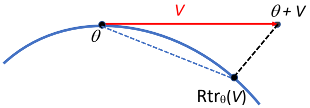


Figure: Example of a retraction.

Majorization-Minimization (MM)

- Majorization-Minimizaiton
  - Choose a majorizing surrogate $g_n(\boldsymbol{\theta})$ of f at $\boldsymbol{\theta}_{n-1}$
  - $\boldsymbol{\theta}_n \leftarrow \arg\min_{\boldsymbol{\theta}\in\Theta} g_n(\boldsymbol{\theta})$
- Ex: PGD
  - $g_n(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_{n-1}) + \langle \nabla f(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}-\boldsymbol{\theta}_{n-1}\rangle + \frac{L}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_{n-1}\|^2$ (prox-linear surr)
  - $\boldsymbol{\theta}_n = \mathrm{Proj}_{\Theta}(\boldsymbol{\theta}_{n-1} - \frac{1}{L}\nabla f(\boldsymbol{\theta}_{n-1}))$
- Ex: Linear surrogate over Stiefel Manifold
  - $g_n(\boldsymbol{\theta}) := f_n(\boldsymbol{\theta}_{n-1}) + \langle \nabla f_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}-\boldsymbol{\theta}_{n-1}\rangle$
  - $\boldsymbol{\theta}_n = \mathrm{Proj}_{\mathcal{V}^{n\times k}}(-\nabla f_n(\boldsymbol{\theta}_{n-1}))$

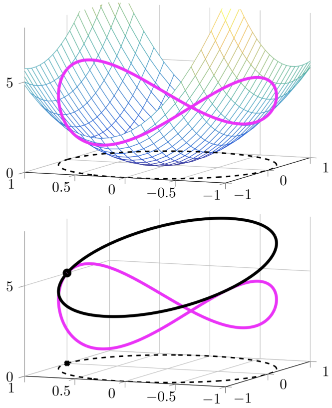


Figure: Example of linear surrogate over Stiefel manifold (Excerpted from [BKSP21])

Euclidean Block MM

- (Euclidean) Block Majorization-minimization: For $n = 1, \ldots, N$ and $i = 1, \ldots, m$

$$\begin{cases} g_n^{(i)} \leftarrow \left[\text{Majorizing surrogate of } f_n^{(i)}(\theta) := f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)\right] \\ \theta_n^{(i)} \in \arg\min_{\theta \in \Theta^{(i)} \subseteq \mathbb{R}^{I_i}} g_n^{(i)}(\theta) \end{cases}$$

  - Sequentially update each block while fixing the rest.
  - Special case: Block PGD (block coordinate descent)

Riemannian Block MM

- Riemannian Block MM: For $n = 1, \ldots, N$ and $i = 1, \ldots, m$

$$g_n^{(i)} \leftarrow \left[\text{Majorizing surrogate of } \theta \mapsto f_n^{(i)}(\theta) := f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)\right]$$

  - $\theta \in \Theta^{(i)} \subseteq \mathcal{M}^{(i)}$: a Riemannian manifold

- Riemannian Block MM: For $n = 1, \dots, N$ and $i = 1, \dots, m$

$$g_n^{(i)} \leftarrow \left[\text{Majorizing surrogate of } \theta \mapsto f_n^{(i)}(\theta) := f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)\right]$$

  - $\theta \in \Theta^{(i)} \subseteq \mathcal{M}^{(i)}$: a Riemannian manifold
  - Two options for minimizing $g_n^{(i)}$:

$$\textbf{Option 1: } \theta_n^{(i)} \in \arg\min_{\theta \in \Theta^{(i)}} g_n^{(i)}(\theta); \quad \textbf{Option 2: } \begin{cases} V_n^{(i)} \in \arg\min_{V \in T_{\theta_{n-1}^{(i)}}} g_n^{(i)}(\theta_{n-1}^{(i)} + V) \\ \alpha_n^{(i)} \leftarrow \text{line search} \\ \theta_n^{(i)} = \mathrm{Rtr}_{\theta_{n-1}^{(i)}}\left(\alpha_n^{(i)} V_n^{(i)}\right) \end{cases}$$

Riemannian Block MM

- Riemannian Block MM: For $n = 1, \ldots, N$ and $i = 1, \ldots, m$

$$g_n^{(i)} \leftarrow \left[\text{Majorizing surrogate of } \theta \mapsto f_n^{(i)}(\theta) := f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)\right]$$

  - $\theta \in \Theta^{(i)} \subseteq \mathcal{M}^{(i)}$: a Riemannian manifold
  - Two options for minimizing $g_n^{(i)}$:

$$\textbf{Option 1: } \theta_n^{(i)} \in \underset{\theta \in \Theta^{(i)}}{\arg\min}\, g_n^{(i)}(\theta); \quad \textbf{Option 2: } \begin{cases} V_n^{(i)} \in \arg\min_{V \in T_{\theta_{n-1}^{(i)}}} g_n^{(i)}(\theta_{n-1}^{(i)} + V) \\ \alpha_n^{(i)} \leftarrow \text{line search} \\ \theta_n^{(i)} = \mathrm{Rtr}_{\theta_{n-1}^{(i)}}\left(\alpha_n^{(i)} V_n^{(i)}\right) \end{cases}$$

- Pros and Cons:

Riemannian Block MM

- Riemannian Block MM: For $n = 1, \dots, N$ and $i = 1, \dots, m$

$$g_n^{(i)} \leftarrow \left[\text{Majorizing surrogate of } \theta \mapsto f_n^{(i)}(\theta) := f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)\right]$$

  - $\theta \in \Theta^{(i)} \subseteq \mathcal{M}^{(i)}$: a Riemannian manifold
  - Two options for minimizing $g_n^{(i)}$:

Option 1: $\theta_n^{(i)} \in \arg\min_{\theta \in \Theta^{(i)}} g_n^{(i)}(\theta)$; **Option 2:** $\begin{cases} V_n^{(i)} \in \arg\min_{V \in T_{\theta_{n-1}^{(i)}}} g_n^{(i)}(\theta_{n-1}^{(i)} + V) \\ \alpha_n^{(i)} \leftarrow \text{line search} \\ \theta_n^{(i)} = \mathrm{Rtr}_{\theta_{n-1}^{(i)}}\left(\alpha_n^{(i)} V_n^{(i)}\right) \end{cases}$

- Pros and Cons:
  - **Option 1** works for more general surrogates and objective functions, but the convergence analysis is more complicated

Riemannian Block MM

- Riemannian Block MM: For $n = 1, \dots, N$ and $i = 1, \dots, m$

$$g_n^{(i)} \leftarrow \left[\text{Majorizing surrogate of } \theta \mapsto f_n^{(i)}(\theta) := f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)\right]$$

  - $\theta \in \Theta^{(i)} \subseteq \mathcal{M}^{(i)}$: a Riemannian manifold
  - Two options for minimizing $g_n^{(i)}$:

$$\textbf{Option 1: } \theta_n^{(i)} \in \underset{\theta \in \Theta^{(i)}}{\arg\min}\, g_n^{(i)}(\theta); \quad \textbf{Option 2: } \begin{cases} V_n^{(i)} \in \arg\min_{V \in T_{\theta_{n-1}^{(i)}}} g_n^{(i)}(\theta_{n-1}^{(i)} + V) \\ \alpha_n^{(i)} \leftarrow \text{line search} \\ \theta_n^{(i)} = \mathrm{Rtr}_{\theta_{n-1}^{(i)}}\left(\alpha_n^{(i)} V_n^{(i)}\right) \end{cases}$$

- Pros and Cons:
  - **Option 1** works for more general surrogates and objective functions, but the convergence analysis is more complicated
  - **Option 2** enjoys much simpler convergence analysis, but currently only allow prox-linear surrogates for Euclidean submanifolds, also the objective function need to be smooth in ambient space.

Riemannian Block MM

- Riemannian Block MM: For $n = 1, \dots, N$ and $i = 1, \dots, m$

$$g_n^{(i)} \leftarrow \left[\text{Majorizing surrogate of } \theta \mapsto f_n^{(i)}(\theta) := f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)\right]$$

  - $\theta \in \Theta^{(i)} \subseteq \mathcal{M}^{(i)}$: a Riemannian manifold
  - Two options for minimizing $g_n^{(i)}$:

$$\textbf{Option 1: } \theta_n^{(i)} \in \operatorname*{arg\,min}_{\theta \in \Theta^{(i)}} g_n^{(i)}(\theta); \quad \textbf{Option 2: } \begin{cases} V_n^{(i)} \in \operatorname{arg\,min}_{V \in T_{\theta_{n-1}^{(i)}}} g_n^{(i)}(\theta_{n-1}^{(i)} + V) \\ \alpha_n^{(i)} \leftarrow \text{line search} \\ \theta_n^{(i)} = \mathrm{Rtr}_{\theta_{n-1}^{(i)}}\left(\alpha_n^{(i)} V_n^{(i)}\right) \end{cases}$$

- Pros and Cons:
  - **Option 1** works for more general surrogates and objective functions, but the convergence analysis is more complicated
  - **Option 2** enjoys much simpler convergence analysis, but currently only allow prox-linear surrogates for Euclidean submanifolds, also the objective function need to be smooth in ambient space.
- Rmk: The two options coincide in the Euclidean setting with prox-linear surrogates.

Examples

- (Subspace Estimation with Grassmannian Geodesics[BRFB23])

$$X_i = U_i G_i + N_i$$

where $U_i \in \mathbb{R}^{d \times k}$ has orthonormal columns representing a point on the Grassmannian $\mathcal{G}(k, d)$; $G_i \in \mathbb{R}^{k \times \ell}$ holds weight or loading vectors; and $N_i \in \mathbb{R}^{d \times \ell}$ is an independent additive noise matrix.

  - Goal: Estimate U_i given all X_i

Examples

- (Subspace Estimation with Grassmannian Geodesics [BRFB23])

$$X_i = U_i G_i + N_i$$

where $U_i \in \mathbb{R}^{d\times k}$ has orthonormal columns representing a point on the Grassmannian $\mathcal{G}(k,d)$; $G_i \in \mathbb{R}^{k\times \ell}$ holds weight or loading vectors; and $N_i \in \mathbb{R}^{d\times \ell}$ is an independent additive noise matrix.

  - Goal: Estimate U_i given all X_i

Model U_i:

$$U_i = U(t_i) = H\cos(\Theta t_i) + Y\sin(\Theta t_i)$$

Objective function f,

$$f(U) = f(H, Y, \Theta) = \min_{\{G_i\}_{i=1}^T} \|X_i - U(t_i)G_i\|_F^2 = -\sum_{i=1}^T \|X_i^T U(t_i)\|_F^2 + c$$

  - Two blocks: $Q = [H\ Y]$ and Θ
  - $Q \in \mathcal{V}^{d\times 2k}$, a stiefel manifold

Examples

- Other examples:
  - (Optimilstic likelihood under Fisher-Rao distnce [NSAY+19])

$$\min_{\mu,\Sigma} f(\mu,\Sigma) \triangleq \left\langle M^{-1}\sum_{m=1}^{M}(x_m-\mu)(x_m-\mu)^T, \Sigma^{-1}\right\rangle + \log\det\Sigma$$

where $\Sigma \in \mathbb{S}^n_{++}$ the manifold of positive definite matrices.

  - (Robust PCA)

$$\min_{L,S} f(L,S) \triangleq \lambda\|S\|_1 + \frac{1}{2\mu}\|M-L-S\|_F^2$$

rank$(L) \leq r$, so L represents a point on low-rank manifold.

Outline

Preliminaries

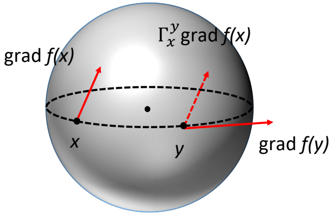
Assumption 1 (g-smooth objective and sublevel compactness) There exists a constant $L_f > 0$ such that the function $f : \boldsymbol{\Theta} = \Theta^{(1)} \times \cdots \times \Theta^{(m)} \to \mathbb{R}$ is geodesically L_f-smooth of order β in each block coordinate. Furthermore, the sublevel sets $f^{-1}((-\infty, a)) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : f(\boldsymbol{\theta}) \le a\}$ are compact for each $a \in \mathbb{R}$.

Definition (Geodesic L-smoothness of order β)

The objective function $f : \mathcal{M} \to \mathbb{R}$ is geodesically L-smooth of order β ($\beta > 1$) if it satisfies

$$\left\|\operatorname{grad} f(x) - \Gamma_y^x(\operatorname{grad} f(y))\right\| \le \frac{L}{2} d^{\beta-1}(x, y)$$

for all x, $y \in \mathcal{M}$, where $\Gamma_x^y : T_x \to T_y$ is the parallel transport along a minimal geodesic joining x and y, $d(x, y)$ is the distance between x and y.



Preliminaries

Assumption 2 (g-convex constraints) Each $\Theta^{(i)}$ is geodesically convex. That is, given any two points in $\Theta^{(i)}$, there exists a distance minimizing geodesic contained in $\Theta^{(i)}$ that joins the two points.

Preliminaries

Assumption 3 (Good surrogates or good Manifold) Assume one of the three:

(i) (Option 1) Each surrogate $g_n^{(i)}$ is L_g-geodesically-smooth of order β for some constant $L_g \geq 0$ for all $n \geq 1$ and $i = 1, \ldots, m$.

Preliminaries

Assumption 3 (Good surrogates or good Manifold) Assume one of the three:

(i) (Option 1) Each surrogate $g_n^{(i)}$ is L_g-geodesically-smooth of order β for some constant $L_g \geq 0$ for all $n \geq 1$ and $i = 1, \ldots, m$.

(ii) (Option 1) The manifolds $\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(m)}$ have uniformly lower bounded injectivity radius; $g_n^{(i)}$ = proximal surrogates:

$$g_n^{(i)}(\theta) = f_n^{(i)}(\theta) + \frac{\lambda_n}{2} d^2(\theta, \theta_{n-1}^{(i)}). \quad \text{(could be 'non-}g\text{-smooth')}$$

Assumption 3 (Good surrogates or good Manifold) Assume one of the three:

(i) (Option 1) Each surrogate $g_n^{(i)}$ is L_g-geodesically-smooth of order β for some constant $L_g \geq 0$ for all $n \geq 1$ and $i = 1, \ldots, m$.

(ii) (Option 1) The manifolds $\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(m)}$ have uniformly lower bounded injectivity radius; $g_n^{(i)} =$ proximal surrogates:

$$g_n^{(i)}(\theta) = f_n^{(i)}(\theta) + \frac{\lambda_n}{2} d^2(\theta, \theta_{n-1}^{(i)}). \quad \text{(could be 'non-}g\text{-smooth')}$$

(iii) (Option 2) The manifolds $\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(m)}$ are compact; $g_n^{(i)} =$ prox-linear surrogates:

$$g_n^{(i)}(\theta) = f_n^{(i)}(\theta_{n-1}^{(i)}) + \langle \nabla f_n^{(i)}(\theta_{n-1}^{(i)}),\, \theta - \theta_{n-1}^{(i)} \rangle + \frac{\lambda_n}{2} \|\theta - \theta_{n-1}^{(i)}\|^2$$

(could be 'non-g-smooth')

Preliminaries

Assumption 3 (Good surrogates or good Manifold) Assume one of the three:

(i) (Option 1) Each surrogate $g_n^{(i)}$ is L_g-geodesically-smooth of order β for some constant $L_g \geq 0$ for all $n \geq 1$ and $i = 1, \dots, m$.

(ii) (Option 1) The manifolds $\mathcal{M}^{(1)}, \dots, \mathcal{M}^{(m)}$ have uniformly lower bounded injectivity radius; $g_n^{(i)}$ = proximal surrogates:

$$g_n^{(i)}(\theta) = f_n^{(i)}(\theta) + \frac{\lambda_n}{2} d^2(\theta, \theta_{n-1}^{(i)}). \quad \text{(could be 'non-}g\text{-smooth')}$$

(iii) (Option 2) The manifolds $\mathcal{M}^{(1)}, \dots, \mathcal{M}^{(m)}$ are compact; $g_n^{(i)}$ = prox-linear surrogates:

$$g_n^{(i)}(\theta) = f_n^{(i)}(\theta_{n-1}^{(i)}) + \langle \nabla f_n^{(i)}(\theta_{n-1}^{(i)}), \theta - \theta_{n-1}^{(i)} \rangle + \frac{\lambda_n}{2} \|\theta - \theta_{n-1}^{(i)}\|^2$$

(could be 'non-g-smooth')

- Why proximal surrogates in (ii) may not be g-smooth (i)?

Preliminaries

Proposition (Riemannian gradient of geodesic distance)

$\mathcal{M}$ = Complete Riemannian manifold, $p \in \mathcal{M}$ with $\overbrace{\mathrm{inj}(p)}^{\text{injectivity radius}} \geq r$. Let $h : \mathcal{M} \to \mathbb{R}$, $h(x) = d^2_{\mathcal{M}}(x, p)$. If $d(x, p) < r$, then $\mathrm{grad}(h) = -2\,\mathrm{Exp}_x^{-1}(p)$ as a vector in $T_x\mathcal{M}$.

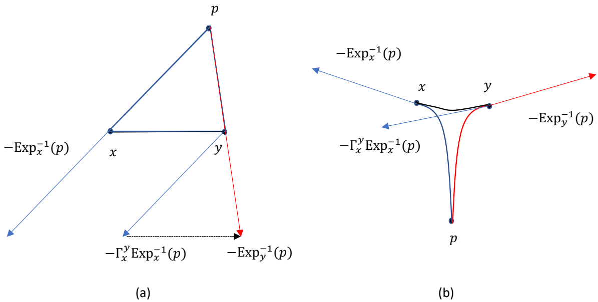


Figure: Examples on g-smoothness of $d^2(x, p)$. Panel (a) is an example in Euclidean space; Panel (b) is a counterexample in hyperbolic space.

Preliminaries

Proposition (Riemannian gradient of geodesic distance)

$\mathcal{M}$ = *Complete Riemannian manifold,* $p \in \mathcal{M}$ *with* $\overbrace{\mathrm{inj}(p)}^{\text{injectivity radius}} \geq r$*. Let* $h : \mathcal{M} \to \mathbb{R}$*,* $h(x) = d^2_{\mathcal{M}}(x, p)$*. If* $d(x, p) < r$*, then* $\mathrm{grad}(h) = -2\,\mathrm{Exp}_x^{-1}(p)$ *as a vector in* $T_x\mathcal{M}$*.*

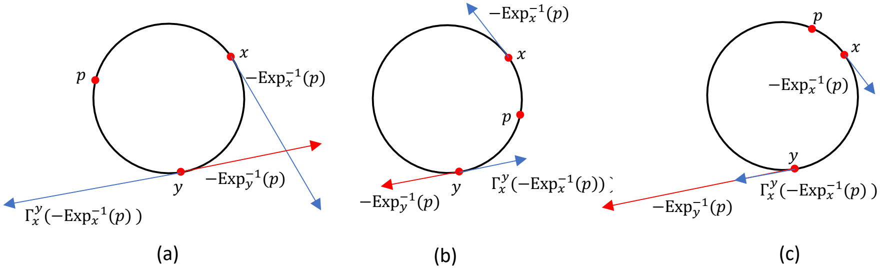


Figure: Examples on g-smoothness of $d^2(x, p)$ on S^1. Panel (a) is an counterexample; Panel (b) (c) are the cases when g-smoothness inequality becomes an equality with $L = 2$.

Preliminaries

Theorem ((LLBN '23+) Asymptotic convergence to stationary points; two blocks)

f = Objective function with $m = 2$ blocks. $(\boldsymbol{\theta}_n)_{n\geq 0}$ = Output of RBMM. Suppose Assumptions 1-3 hold. Then every limit point of $(\boldsymbol{\theta}_n)_{n\geq 0}$ is a stationary point of f over $\boldsymbol{\Theta}$.

Preliminaries

Assumption 4 (Distance-regularizing surrogates) There exists a strictly increasing function $\phi : [0, \infty) \to \mathbb{R}$ such that $\phi(0) = 0$ and

$$h_n^{(i)}(\theta) := g_n^{(i)}(\theta) - f_n^{(i)}(\theta) \geq \phi(d(\theta, \theta_{n-1}^{(i)}))$$

for all $n \geq 1$ and $i = 1, \ldots, m$.

Results

Assumption 4 (Distance-regularizing surrogates) For Option 1, there exists a strictly increasing function $\phi : [0, \infty) \to \mathbb{R}$ such that $\phi(0) = 0$ and

$$h_n^{(i)}(\theta) := g_n^{(i)}(\theta) - f_n^{(i)}(\theta) \geq \phi(d(\theta, \theta_{n-1}^{(i)}))$$

for all $n \geq 1$ and $i = 1, \ldots, m$.

Theorem (Asymptotic convergence to stationary points; many blocks)

Let f denote the objective function with $m \geq 2$. Let $(\boldsymbol{\theta}_n)_{n\geq 0}$ be a output of RBMM. Suppose Assumptions 1, 3, 4(for Option 1) hold. Then every limit point of $(\boldsymbol{\theta}_n)_{n\geq 0}$ is a stationary point of f over $\boldsymbol{\Theta}$.

Preliminaries

Definition (ε-approxiate stationary point): we say $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ is an *ε-approxiate stationary point* of f over $\boldsymbol{\Theta}$ if

$$-\inf_{\eta \in T^*_{\boldsymbol{\theta}_n}} \left\langle \operatorname{grad} f(\boldsymbol{\theta}^*), \frac{\eta}{\|\eta\|} \right\rangle \leq \sqrt{\varepsilon}.$$

where $T^*_\theta \mathcal{M}^{(i)} := \{\eta \in T_\theta \mathcal{M}^{(i)} : \mathrm{Exp}_\theta(\eta) \in \Theta^{(i)}\}$.

Definition (*worst-case iteration complexity*) :

$$N_\varepsilon := \sup_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}} \inf \{n \geq 1 \,|\, \boldsymbol{\theta}_n \text{ is an } \varepsilon\text{-approximate stationary point of } f \text{ over } \boldsymbol{\Theta}\},$$

where $(\boldsymbol{\theta}_n)_{n \geq 0}$ is a sequence of estimates produced by the algorithm with initial estimate $\boldsymbol{\theta}_0$.

Results

Theorem (Rate of convergence for proximal surrogates on Riemannian manifolds with lower bounded injectivity radius)

f = Objective function with $m \geq 2$ blocks. $(\boldsymbol{\theta}_n)_{n\geq 0}$ =output of RBMM. Suppose Assumptions 1-3 hold. Assume [Option 1 with prox surrogates] or [Option 2 with prox-linear surrogates].

(i) *(Worst-case rate of convergence) There exists constants M and $c > 0$ independent of $\boldsymbol{\theta}_0$ such that*

$$\min_{1\leq k\leq n}\left[-\inf_{\eta\in T^*_{\boldsymbol{\theta}_n}}\left\langle \operatorname{grad} f(\boldsymbol{\theta}_n), \frac{\eta}{\|\eta\|}\right\rangle\right] \leq \frac{M}{\sqrt{n}/\log n}$$

(ii) *(Worst-case iteration complexity) The worst-case iteration complexity N_ϵ for RBMM satisfies $N_\epsilon = O(\varepsilon^{-1}\left(\log \varepsilon^{-1}\right)^2)$*

Results

Theorem (Rate of convergence for smooth surrogates)

f =objective function with $m \geq 2$ blocks. $(\boldsymbol{\theta}_n)_{n\geq 0}$ =output of RBMM. Suppose Assumptions 1-4 hold. Assume [Option 1 with g-smooth surrogates]. Suppose Assumption 5 holds with $\phi(x) = cx^{\beta}$ for some constant $c > 0$. Let $\alpha := (\beta - 1)/\beta^2$.

(i) *(Worst-case rate of convergence) There exists constants $M, c > 0$ independent of $\boldsymbol{\theta}_0$ such that*

$$\min_{1\leq k\leq n} \left[- \inf_{\eta \in T^*_{\boldsymbol{\theta}_n}} \left\langle \operatorname{grad} f(\boldsymbol{\theta}_n), \frac{\eta}{\|\eta\|} \right\rangle \right] \leq \frac{M + c\sum_{n=1}^{\infty} \Delta_n(\boldsymbol{\theta}_0)}{n^{\alpha}/(\log n)^{1/2}}$$

(ii) *(Worst-case iteration complexity) The worst-case iteration complexity N_ϵ for RBMM satisfies $N_\epsilon = O\left(\varepsilon^{-1/2\alpha}\left(\log \varepsilon^{-1}\right)\right)$.*

(iii) *(Optimal convergence rate) Further assume that the surrogate gaps $h_n^{(i)} = g_n^{(i)} - f_n^{(i)}$ satisfy $h_n^{(i)}(\theta) \leq Cd^{\beta}(\theta, \theta_n^{(i)})$ for some constant $C > 0$. Then the results in* **(i)-(ii)** *hold with the improved exponent $\alpha = (\beta - 1)/\beta$.*

Outline

Examples

- (Euclidean BMM) When specialized on the standard Euclidean manifold, our RBMM becomes the standard Euclidean Block MM (e.g., see BSUM in [HRLP15])

Examples

- (Euclidean BMM) When specialized on the standard Euclidean manifold, our RBMM becomes the standard Euclidean Block MM (e.g., see BSUM in [HRLP15])
  - Our general result gives convergence rate $\widetilde{O}(\varepsilon^{-1})$ even for nonconvex objectives with convex constraints.

Examples

- (Euclidean BMM) When specialized on the standard Euclidean manifold, our RBMM becomes the standard Euclidean Block MM (e.g., see BSUM in [HRLP15])
  - Our general result gives convergence rate $\widetilde{O}(\varepsilon^{-1})$ even for nonconvex objectives with convex constraints.
  - The same rate was known for convex problems [HRLP15]

Examples

- (Euclidean BMM) When specialized on the standard Euclidean manifold, our RBMM becomes the standard Euclidean Block MM (e.g., see BSUM in [HRLP15])
  - Our general result gives convergence rate $\widetilde{O}(\varepsilon^{-1})$ even for nonconvex objectives with convex constraints.
  - The same rate was known for convex problems [HRLP15]
- (Block Prox-linear and Block PGD) Consider the following block prox-linear update proposed in [XY13].

$$\theta_n^{(i)} \leftarrow \operatorname*{arg\,min}_{\theta \in \Theta^{(i)}} \left(g_n^{(i)}(\theta) := f_n^{(i)}(\theta_{n-1}^{(i)}) + \langle \nabla f_n^{(i)}(\theta_{n-1}^{(i)}),\, \theta - \theta_{n-1}^{(i)} \rangle + \frac{\lambda}{2} \|\theta - \theta_{n-1}^{(i)}\|^2 \right).$$

  - Asymptotic convergence to stationary points
  - Iteration complexity of $\widetilde{O}(\varepsilon^{-1})$

$$\begin{aligned} \theta_n^{(i)} \leftarrow \operatorname*{arg\,min}_{\theta \in \Theta^{(i)}} \left(\langle \nabla,\, \theta \rangle + \frac{\lambda}{2} \|\theta\|^2 - \lambda \langle \theta, \theta_{n-1}^{(i)} \rangle \right) &= \operatorname*{arg\,min}_{\theta \in \Theta^{(i)}} \left\| \theta - \left(\theta_{n-1}^{(i)} - \frac{1}{\lambda} \nabla \right) \right\|^2 \\ &= \mathrm{Proj}_{\Theta^{(i)}} \left(\theta_{n-1}^{(i)} - \frac{1}{\lambda} \nabla \right). \end{aligned}$$

Examples

- (Block prox-linear on Riemannian manifold)

$$\theta_n^{(i)} \leftarrow \operatorname*{arg\,min}_{\theta \in \Theta^{(i)}} \left(g_n^{(i)}(\theta) := f_n^{(i)}(\theta_{n-1}^{(i)}) + \langle \nabla f_n^{(i)}(\theta_{n-1}^{(i)}), \theta - \theta_{n-1}^{(i)} \rangle + \frac{\lambda}{2} \|\theta - \theta_{n-1}^{(i)}\|^2 \right)$$
$$= \mathrm{Proj}_{\Theta^{(i)}} \left(\theta_{n-1}^{(i)} - \frac{1}{\lambda} \nabla f_n^{(i)}(\theta_{n-1}^{(i)}) \right)$$

  - Asymptotic convergence to stationary points

Examples

- (Block prox-linear on Riemannian manifold)

$$\theta_n^{(i)} \leftarrow \operatorname*{arg\,min}_{\theta \in \Theta^{(i)}} \left(g_n^{(i)}(\theta) := f_n^{(i)}(\theta_{n-1}^{(i)}) + \langle \nabla f_n^{(i)}(\theta_{n-1}^{(i)}), \theta - \theta_{n-1}^{(i)} \rangle + \frac{\lambda}{2} \|\theta - \theta_{n-1}^{(i)}\|^2 \right)$$
$$= \mathrm{Proj}_{\Theta^{(i)}} \left(\theta_{n-1}^{(i)} - \frac{1}{\lambda} \nabla f_n^{(i)}(\theta_{n-1}^{(i)}) \right)$$

  - Asymptotic convergence to stationary points

- (Block Proximal Updates on Hadamard manifolds/Stiefel manifolds)

$$g_n^{(i)}(\theta) = f_n^{(i)}(\theta) + \frac{\lambda_n}{2} \cdot d^2 \left(\theta, \theta_{n-1}^{(i)} \right)$$

  - Asymptotic convergence to stationary points
  - Iteration complexity of $\widetilde{O}(\varepsilon^{-1})$

Hadamard manifolds includes: Euclidean spaces, Hyperbolic spaces, manifold of PD matrices

Outline

Optimistic likelihood

- Optimistic likelihood problem:

$$g_n^{(1)}(\mu) = \left\langle M^{-1} \sum_{m=1}^{M} (x_m - \mu)(x_m - \mu)^T, \Sigma_{n-1}^{-1} \right\rangle + \log\det \Sigma_{n-1} + \frac{\lambda_n}{2} \|\mu - \mu_{n-1}\|^2$$

$$g_n^{(2)}(\Sigma) = \langle S_n, \Sigma^{-1} \rangle + \log\det \Sigma + \frac{\lambda}{4} \left\| \log\left(\Sigma_{n-1}^{-\frac{1}{2}} \Sigma \Sigma_{n-1}^{-\frac{1}{2}} \right) \right\|_F^2$$

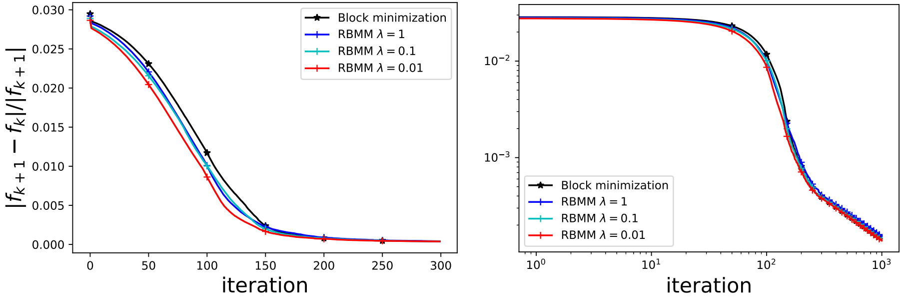


Figure: Comparison of block minimization and RBMM applied to optimistic likelihood problem under Fisher-Rao distance. RBMM is implemented with $\lambda = 0.01, 0.1, 1$ respectively.

Geodesic subspace tracking problem

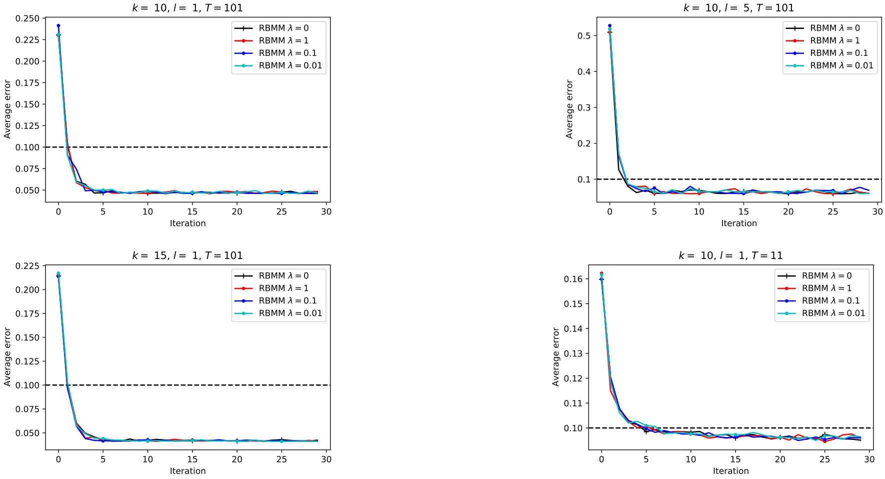


Figure: Convergence of RBMM in geodesic error under different settings. Average geodesic error is computed over 50 independent trials. The dimension is $d = 30$ and the additive Gaussian noise has standard deviation $\sigma = 0.1$. The value of other parameters are shown in the title for each panel.

Thanks!

Frame Title

- G.C. Bento, O.P. Ferreira, and P.R. Oliveira, *Proximal point method for a special class of nonconvex functions on hadamard manifolds*, Optimization **64** (2015), no. 2, 289–319.
- Arnaud Breloy, Sandeep Kumar, Ying Sun, and Daniel P Palomar, *Majorization-minimization on the stiefel manifold with application to robust sparse pca*, IEEE Transactions on Signal Processing **69** (2021), 1507–1520.
- Cameron J Blocker, Haroon Raja, Jeffrey A Fessler, and Laura Balzano, *Dynamic subspace estimation with grassmannian geodesics*, arXiv preprint arXiv:2303.14851 (2023).
- 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** (2020), no. 1, 210–239.
- Mingyi Hong, Meisam Razaviyayn, Zhi-Quan Luo, and Jong-Shi Pang, *A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing*, IEEE Signal Processing Magazine **33** (2015), no. 1, 57–77.
- Wen Huang and Ke Wei, *Riemannian proximal gradient methods*, Mathematical Programming **194** (2022), no. 1-2, 371–413.
- Viet Nguyen, Soroosh Shafieezadeh-Abadeh, Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann, *Calculating optimistic likelihoods using (geodesically) convex optimization*, NeurIPS'19: Proceedings of the 33rd International Conference on Neural Information