Convergence and complexity of block majorization-minimization on Riemannian manifolds

Yuchen Li[1], Laura Balzano[2], Deanna Needell[3], Hanbaek Lyu[1]

[1]University of Wisconsin - Madison,[2]University of Michigan, [3]University of California, Los Angeles

Introduction

Consider the following multi-block constrained optimization problem on Riemannian manifolds,

$$\min_{\substack{\boldsymbol{\theta}=[\theta^{(1)},\dots,\theta^{(m)}] \\ \theta^{(i)}\in\Theta^{(i)}\subseteq\mathcal{M}^{(i)} \text{ for } i=1,\dots,m}} f(\boldsymbol{\theta}) \qquad (1)$$

i.e. the objective function f maps smoothly a point on the product manifold $\mathcal{M}^{(1)}\times\cdots\times\mathcal{M}^{(m)}$. In order to obtain a first-order optimal solution to (1), we consider various Riemannian generalizations of *Block Majorization-Minimization* (BMM) algorithm in the Euclidean space [HRLP15]. The high-level idea of BMM in the is that, in order to minimize a multi-block objective, one can minimize a majorizing surrogate of the objective in each block in a cyclic order. The same idea of applying the MM principle in a blockwise fashion may seem very natural in the Riemannian setting, but the way to do so is not necessarily unique. Below we give our RBMM algorithm which includes two different options,

RBMM:

For $i=1,2,\dots,m$:

Option 1:

Majorize on Manifold and Minimize on Manifold (MmMm)

$g_n^{(i)} \leftarrow \left[\text{Majorizing surrogate of } \theta\mapsto f_n^{(i)}(\theta) \text{ at } \theta_{n-1}^{(i)}\right]$

$\theta_n^{(i)} \in \arg\min_{\theta\in\Theta^{(i)}} g_n^{(i)}(\theta)$

Option 2:

Majorize on Tangent spaces and Minimize on Tangent Spaces (MtMt)

$\hat{g}_n^{(i)} \leftarrow \left[\text{Majorizing surrogate of } \eta\mapsto \hat{f}_n^{(i)}\circ \mathrm{Rtr}_{\theta_{n-1}^{(i)}}(\eta) \text{ at } \mathbf{0}\right]$

$\theta_n^{(i)} \leftarrow$ Minimize $\hat{g}_n^{(i)}$ on $T^*_{\theta_{n-1}^{(i)}}$ and retract onto $\Theta^{(i)}\subseteq\mathcal{M}^{(i)}$

Preliminaries

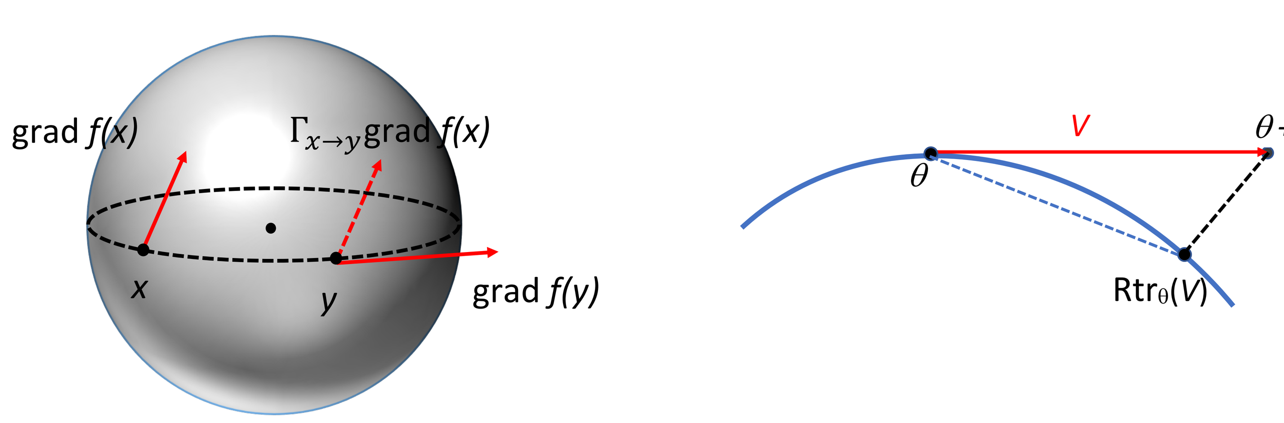

Figure 1: *Illustration of parallel transport (left) and retrations (right)*

Below we give some notations and preliminaries of Riemannian optimization,

Definition 1 (Geodesic L-smoothness) *A function $F:\prod_{i=1}^m \mathbf{M}^{(i)}\to\mathbb{R}$ is geodesically L-smooth (g-smooth in short) if for each $\mathbf{x}=(x^{(1)},\dots,x^{(m)}),\mathbf{y}=(y^{(1)},\dots,y^{(m)})\in\prod_{i=1}^m\mathbf{M}^{(i)}$ where there exists a minimizing geodesic joining $x^{(i)}$ and $y^{(i)}$ for each $i=1,\dots,m$,*

$$\left\|\mathrm{grad}_i F(\mathbf{x})-\Gamma_{y^{(i)}\to x^{(i)}}(\mathrm{grad}_i F(\mathbf{y}))\right\|\le\frac{L}{m}d(\mathbf{x},\mathbf{y})$$

where $\Gamma_{y^{(i)}\to x^{(i)}}$ is the parallel transport along a minimizing geodesic joining $x^{(i)}$ and $y^{(i)}$ in $\mathbf{M}^{(i)}$, and $d(\mathbf{x},\mathbf{y})$ is geodesic distance.

For Option 1, the majorizing surrogate $g_n^{(i)}$ is chosen so that

(1) (Majorization) $g_n^{(i)}(x)-f_n^{(i)}(x)\ge 0$ for all $x\in\mathbf{M}^{(i)}$;

(2) (Sharpness) $g_n^{(i)}(\theta_{n-1}^{(i)})=f_n^{(i)}(\theta_{n-1}^{(i)})$.

Results

Below we give some assumptions for analysis of Option 1 in RBMM,

(A1) *For Option 1, we assume the following hold. For the objective:*

(i) *There exists a constant $L_f>0$ such that the function $f:\Theta=\Theta^{(1)}\times\cdots\times\Theta^{(m)}\to\mathbb{R}$ is geodesically L_f-smooth in each block coordinate.*

For surrogates, one of the following holds:

(ii-g) *(g-smooth surrogates) Each surrogate $g_n^{(i)}$ is L_g-geodesically-smooth for some constant $L_g\ge 0$ for all $n\ge 1$ and $i=1,\dots,m$.*

(ii-p) *(Proximal surrogates) Each $g_n^{(i)}$ is a proximal surrogate: For each $n\ge 1$ and some constant $\lambda_n\ge L_f$,*

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

Moreover, $\lambda_n=O(1)$.

Furthermore, we require the following for the constraint sets:

(iii) *For each $i=1,\dots,m$, there exists a uniform lower bound $r_0>0$ for $r_{\mathrm{cvx}}(x)$ over $x\in\Theta^{(i)}$.*

(A1-1) (Distance-regularizing surrogates) *There exists a strictly increasing function $\phi:[0,\infty)\to\mathbb{R}$ such that $\phi(0)=0$ and*

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

for all $n\ge 1$ and $i=1,\dots,m$.

We first give asymptotic convergence result of RBMM with Option 1,

Theorem 2 (Asymptotic convergence to stationary points; many blocks) *Let f denote the objective function in (1) with $m\ge 2$. Let $(\boldsymbol{\theta}_n)_{n\ge 0}$ be an output of RBMM under (A1) (with either types of surrogages), and (A1-1) hold. Then every limit point of $(\boldsymbol{\theta}_n)_{n\ge 0}$ is a stationary point of f over $\boldsymbol{\Theta}$.*

The following theorem is the result of conplexity of Option 1 in RBMM,

Theorem 3 (Rate of convergence for smooth surrogates) *Let f denote the objective function in (1) with $m\ge 2$. Let $(\boldsymbol{\theta}_n)_{n\ge 0}$ be an output of RBMM under (A1) with g-smooth surrogates. Assume geodesic convexity of the constraint sets. Further assume that (A1-1) holds with $\phi(x)=cx^2$ for some constant $c>0$. Then the following hold:*

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

$$\min_{1\le k\le n}\left[-\sum_{i=1}^m\inf_{\eta\in T^*_{\theta_k^{(i)}},\|\eta\|\le 1}\left\langle \mathrm{grad}_i f(\boldsymbol{\theta}_k),\frac{\eta}{\min\{r_0,1\}}\right\rangle\right]\le\frac{M+c\sum_{n=1}^\infty\Delta_n}{n^{1/4}/(\log n)^{1/2}}$$

(ii) *(Worst-case iteration complexity) The worst-case iteration complexity N_ϵ for RBMM satisfies $N_\epsilon=O\left(\varepsilon^{-4}\left(\log\varepsilon^{-2}\right)\right)$.*

(iii) *(Optimal convergence rate) Further assume that the surrogate gaps $h_n^{(i)}=g_n^{(i)}-f_n^{(i)}$ satisfy $h_n^{(i)}(\theta)\le Cd^2(\theta,\theta_n^{(i)})$ for some constant $C>0$. Then*

$$\min_{1\le k\le n}\left[-\sum_{i=1}^m\inf_{\eta\in T^*_{\theta_k^{(i)}},\|\eta\|\le 1}\left\langle \mathrm{grad}_i f(\boldsymbol{\theta}_k),\frac{\eta}{\min\{r_0,1\}}\right\rangle\right]\le\frac{M+c\sum_{n=1}^\infty\Delta_n}{n^{1/2}/(\log n)^{1/2}}$$

and the worst-case iteration complexity N_ϵ for RBMM is $N_\epsilon=O\left(\varepsilon^{-2}\left(\log\varepsilon^{-2}\right)\right)$.

Examples

1. Euclidean BMM [HRLP15].
2. Block Prox-linear and Block PGD [XY13].
3. Block Euclidean prox-linear on Riemannian manifold.
4. Block Riemannian prox-linear and Block Riemannian GD.
5. Block Proximal Updates on Hadamard manifolds.
6. Block Proximal Updates on Stiefel manifolds.

Numerical experiments

CANDECOMP/PARAFAC (CP) dictionary learning: In the CANDECOMP/PARAFAC (CP) decomposition problem [KB09], given a data tensor $X\in\mathbb{R}^{I_1\times\cdots\times I_m}$ and an integer $R>0$, we would like to find the *loading matrices* $U^{(i)}\in\mathbb{R}^{I_i\times R}$ for $i=1,\cdots,m$ such that

$$X\approx\sum_{k=1}^R\bigotimes_{i=1}^m U^{(i)}[:,k],$$

where $U^{(i)}[:,k]$ denotes the k^{th} column of the $I_i\times R$ loading matrix matrix $U^{(i)}$ and $\otimes$ denotes the outer product. We could formulate the above tensor decomposition problem as the following optimization problem:

$$\underset{U^{(1)}\in\mathcal{M}^{(1)},\dots,U^{(m)}\in\mathcal{M}^{(m)}}{\operatorname{argmin}}\left(f\left(U^{(1)},\dots,U^{(m)}\right):=\left\|X-\sum_{k=1}^R\bigotimes_{i=1}^m U^{(i)}[:,k]\right\|_F^2\right),$$

where $\mathcal{M}^{(i)}\subseteq\mathbb{R}^{I_i\times R}$ is an embedded manifold, which gives additional Riemannian constraints.

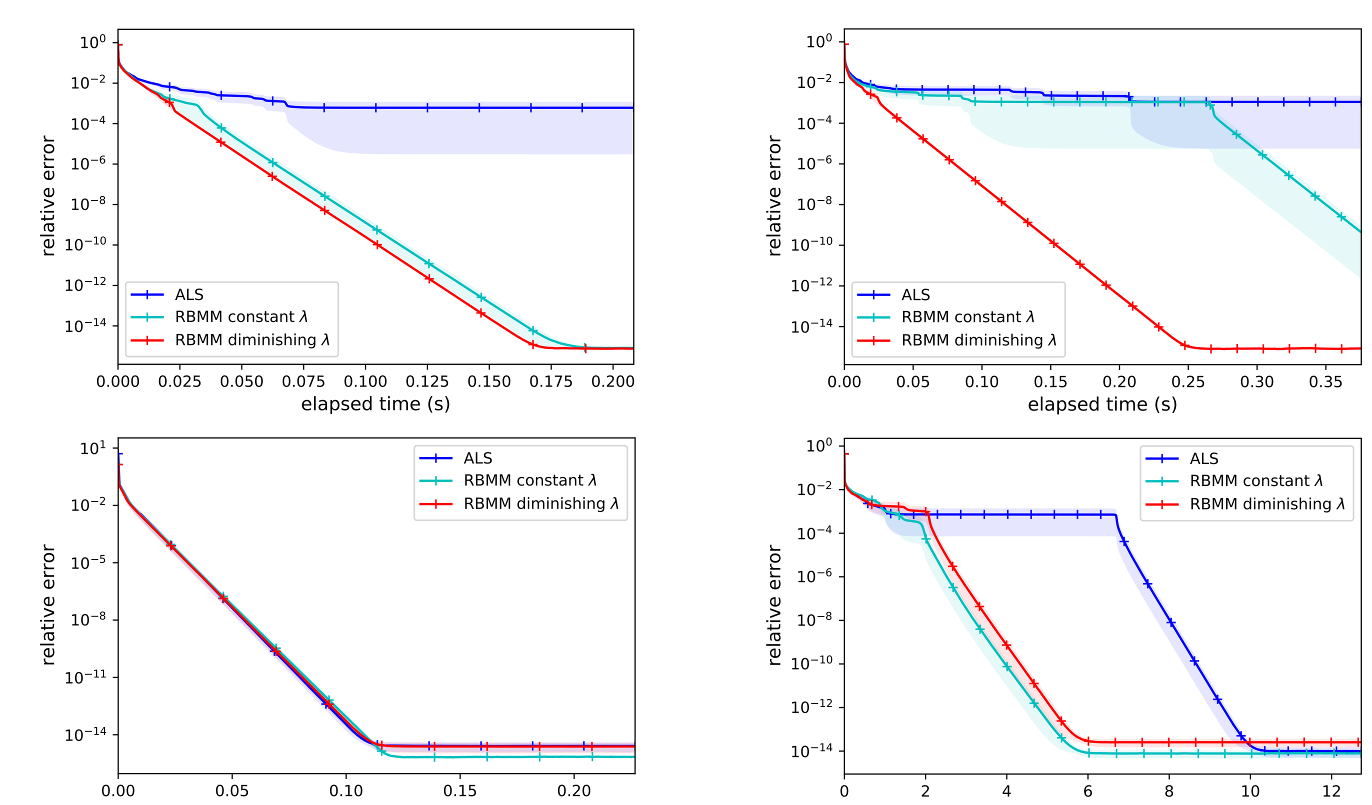

Figure 2: *Top left and top right are some typical cases of synthetic data in Euclidean case; bottom left is the typical result when first block is Stiefel manifold; bottom right is a synthetic example where first block is a point on low-rank manifold. The average relative reconstruction error with standard deviation are shown by the solid lines and shaded regions of respective colors.*

References

[HRLP15] 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.

[KB09] Tamara G Kolda and Brett W Bader, *Tensor decompositions and applications*, SIAM review **51** (2009), no. 3, 455–500.