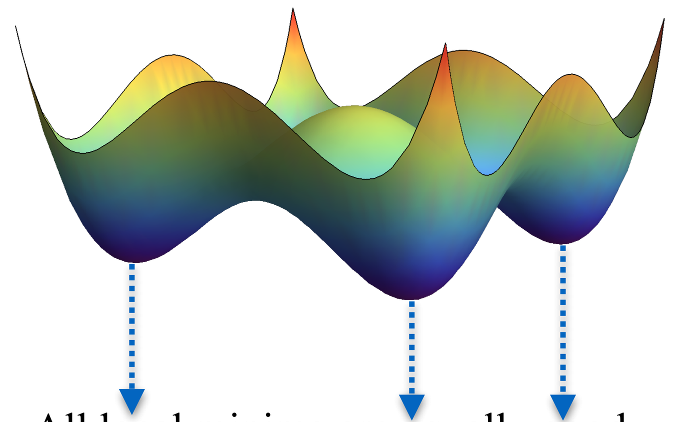
A Linearly Convergent Method for Non-Smooth Non-Convex Optimization on the Grassmannian with Applications to Robust Subspace and Dictionary Learning

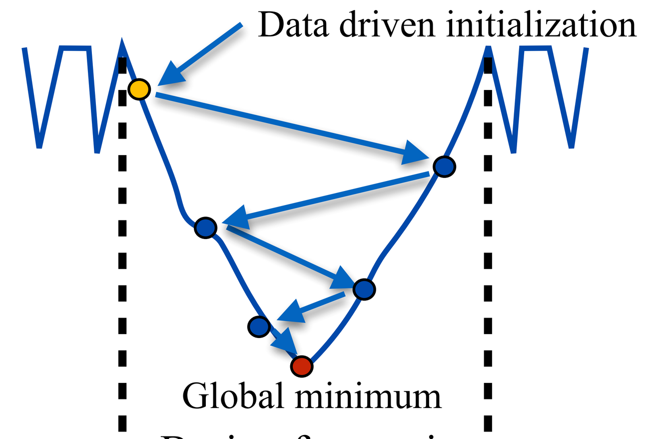
Zhihui Zhu[1], Tianyu Ding[1], Manolis Tsakiris[2], Daniel Robinson[3], René Vidal[1]
[1]Johns Hopkins University [2]ShanghaiTech University [3]Lehigh University

Introduction

- Non-convex optimization has become ubiquitous in machine learning
- New tools are needed to analyze the optimization landscape and develop efficient algorithms with guarantees of convergence to global minima. Recent advances:



Global geometric analysis (smoothness (Hessian) is often required)



Local geometric analysis (could be non-smooth)

- We focus on local analysis of non-smooth problems on the Grassmannian

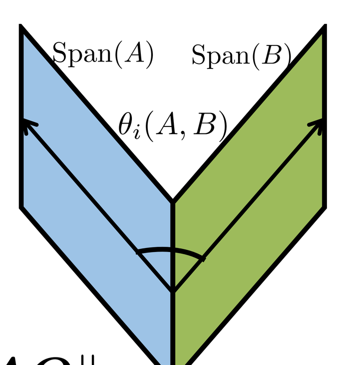
Main Contribution

- Problem: minimize $f(\boldsymbol{B})$ over $\boldsymbol{B} \in \mathbb{O}(c,D) \equiv \{\boldsymbol{B} \in \mathbb{R}^{D\times c} : \boldsymbol{B}^\top \boldsymbol{B} = \mathbf{I}_c\}$
  - $f : \mathbb{R}^{D\times c} \to \mathbb{R}$ is locally Lipschitz, possibly non-convex and non-smooth
  - f is rotation invariant, i.e., $f(\boldsymbol{B}) = f(\boldsymbol{BQ})$ for any $\boldsymbol{Q} \in \mathbb{O}(c,c)$

Contribution 1: Riemannian subgradient method (RSGM) converges locally at an *R-linear rate* if f satisfies a Riemannian regularity condition (RRC)
Contribution 2: Orthogonal Dictionary Learning (ODL) and Dual Principal Component Pursuit (DPCP) satisfy the RRC, which improves existing results
- ODL [1]: a sublinear convergence rate for RSGM
- DPCP [2]: a piecewise linear convergence rate for the sphere case, i.e., $c = 1$

Principal Angles & Distance

- $\forall \boldsymbol{A}, \boldsymbol{B} \in \mathbb{O}(c,D)$, the principal angles between Span($\boldsymbol{A}$) and Span($\boldsymbol{B}$) are defined as $\theta_i(\boldsymbol{A},\boldsymbol{B}) = \arccos(\sigma_i(\boldsymbol{A}^\top\boldsymbol{B}))$, where σ_i is the i-th singular value



- The distance between $\boldsymbol{A}, \boldsymbol{B}$ is defined as

$$\operatorname{dist}(\boldsymbol{A},\boldsymbol{B}) := \sqrt{2\sum_{i=1}^{c}\left(1-\cos(\theta_i(\boldsymbol{A},\boldsymbol{B}))\right)} = \min_{\boldsymbol{Q}\in\mathbb{O}(c,c)} \|\boldsymbol{B}-\boldsymbol{AQ}\|_F$$

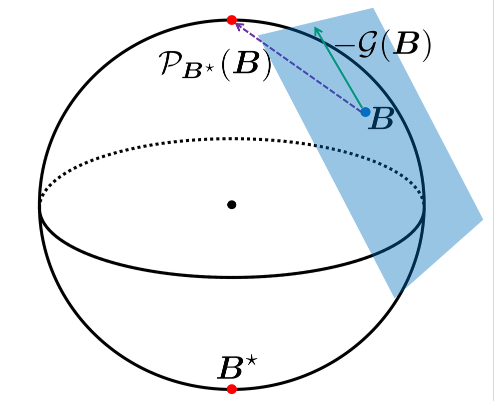
- The projection of $\boldsymbol{B}$ onto equivalence class $[\boldsymbol{A}] = \{\boldsymbol{AQ} : \boldsymbol{Q} \in \mathbb{O}(c,c)\}$ is

$$\mathcal{P}_{\boldsymbol{A}}(\boldsymbol{B}) = \boldsymbol{AQ}^\star, \quad \text{where } \boldsymbol{Q}^\star = \arg\min_{\boldsymbol{Q}\in\mathbb{O}(c,c)} \|\boldsymbol{B}-\boldsymbol{AQ}\|_F$$

Riemannian Regularity Condition (RRC)

- **Definition**: f satisfies the $(\alpha,\epsilon,\boldsymbol{B}^\star)$-RRC if for every $\boldsymbol{B} \in \mathbb{O}(c,D)$ satisfying $\operatorname{dist}(\boldsymbol{B},\boldsymbol{B}^\star) \le \epsilon$, there exists a Riemannian subgradient $\mathcal{G}(\boldsymbol{B}) \in \partial_R f(\boldsymbol{B})$ such that

$$(1) \qquad \langle \mathcal{P}_{\boldsymbol{B}^\star}(\boldsymbol{B}) - \boldsymbol{B}, -\mathcal{G}(\boldsymbol{B})\rangle \ge \alpha \operatorname{dist}(\boldsymbol{B},\boldsymbol{B}^\star)$$



- Closely related to sharpness and weak convexity for unconstrained problems [3]

Riemannian Subgradient Method (RSGM)

- Obtain a Riemannian subgradient $\mathcal{G}(\boldsymbol{B}_k)$ that satisfies (1) with $\boldsymbol{B} = \boldsymbol{B}_k$
- Compute a step size μ_k according to a certain rule
- Update the iterate $\widehat{\boldsymbol{B}}_{k+1} \leftarrow \boldsymbol{B}_k - \mu_k\mathcal{G}(\boldsymbol{B}_k)$ and $\boldsymbol{B}_{k+1} \leftarrow \text{orthonormalize}(\widehat{\boldsymbol{B}}_{k+1})$

Convergence Analysis of RSGM

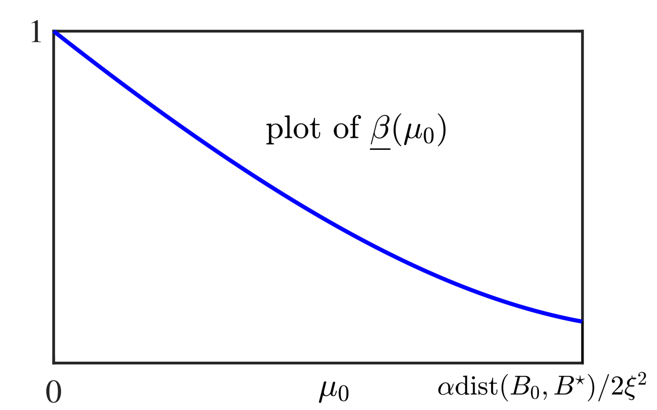
- Assumptions:
  - f satisfies the $(\alpha,\epsilon,\boldsymbol{B}^\star)$-RRC; initialization $\boldsymbol{B}_0$ satisfies $\operatorname{dist}(\boldsymbol{B}_0,\boldsymbol{B}^\star) \le \epsilon$
  - bounded Riemannian subgradient $\|\mathcal{G}(\boldsymbol{B})\|_F \le \xi$, $\forall \boldsymbol{B}$ s.t. $\operatorname{dist}(\boldsymbol{B},\boldsymbol{B}^\star) \le \epsilon$
- **Proposition (constant step size)**: Let $\mu_k \equiv \mu \le \alpha\epsilon/\xi^2$. Then

$$\operatorname{dist}(\boldsymbol{B}_k,\boldsymbol{B}^\star) \le \max\left\{\operatorname{dist}(\boldsymbol{B}_0,\boldsymbol{B}^\star) - \mu\alpha k/2,\ \mu\xi^2/\alpha\right\}$$

  - Due to non-smoothness, there exists an upper bound $\mu\xi^2/\alpha$ for all iterates
  - Tradeoff: larger μ leads to faster decrease, but larger upper bound $\mu\xi^2/\alpha$
- **Theorem**: Let $\mu_k = \mu_0\beta^k$, where
  - $\mu_0 \le \alpha\operatorname{dist}(\boldsymbol{B}_0,\boldsymbol{B}^\star)/2\xi^2$
  - $1 > \beta \ge \underline{\beta}(\mu_0) := \sqrt{1 - 2\frac{\alpha\mu_0}{\operatorname{dist}(\boldsymbol{B}_0,\boldsymbol{B}^\star)} + \frac{\mu_0^2\xi^2}{\operatorname{dist}^2(\boldsymbol{B}_0,\boldsymbol{B}^\star)}}$

  Then $\boldsymbol{B}_k$ converges to $\boldsymbol{B}^\star$ at an R-linear rate:

$$\operatorname{dist}(\boldsymbol{B}_k,\boldsymbol{B}^\star) \le \operatorname{dist}(\boldsymbol{B}_0,\boldsymbol{B}^\star)\beta^k, \ \forall\, k \ge 0.$$



- The larger (smaller) μ_0, the smaller (larger) $\underline{\beta}(\mu_0)$

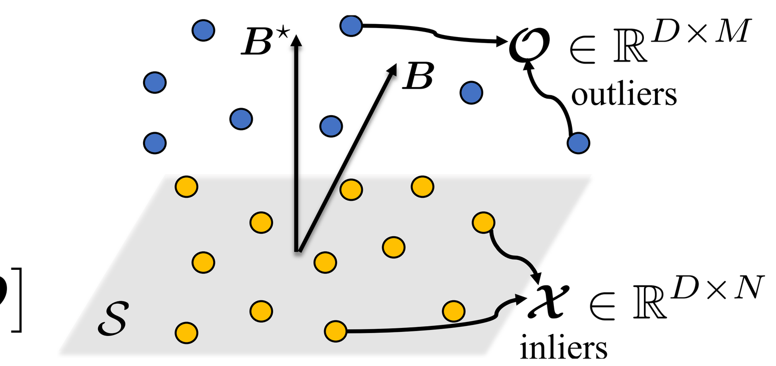
Minimization on Stiefel Manifold

- Our results can be extended to functions that are not rotation invariant by modifying the definition of distance and iterate update: $\boldsymbol{B}_{k+1} = \mathcal{P}_{\mathbb{O}(c,D)}(\widehat{\boldsymbol{B}}_{k+1})$.

Application to DPCP

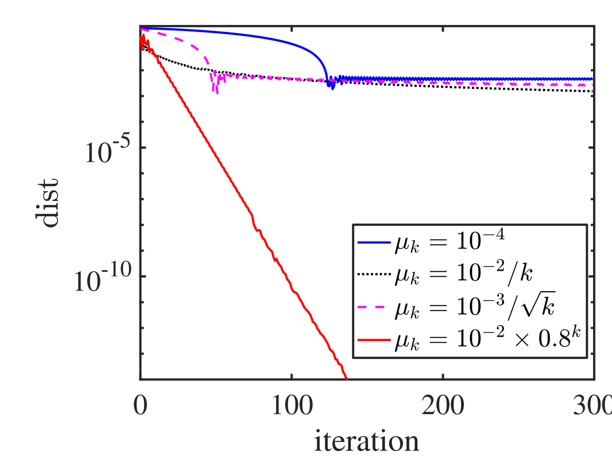
- Fit a subspace $\mathcal{S}$ of codimension $c = D - d$ to data $\boldsymbol{\mathcal{X}}$ corrupted by outliers $\boldsymbol{\mathcal{O}}$ by solving

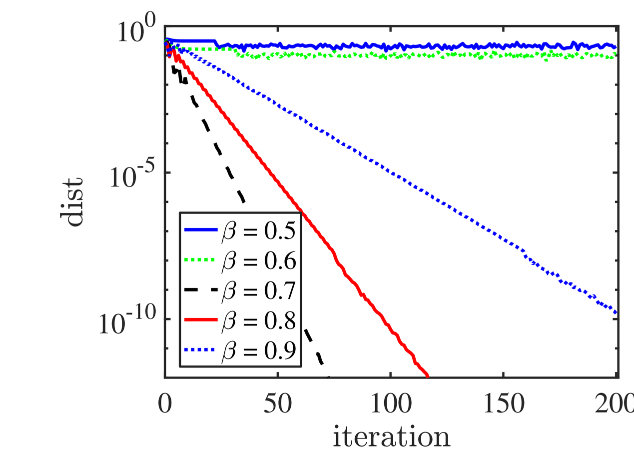
$$\min_{\boldsymbol{B}\in\mathbb{O}(c,D)} \sum_{i=1}^{N+M} \|\boldsymbol{B}^\top\widetilde{\boldsymbol{x}}_i\|_2, \quad \widetilde{\boldsymbol{\mathcal{X}}} = [\boldsymbol{\mathcal{X}}\ \boldsymbol{\mathcal{O}}]$$



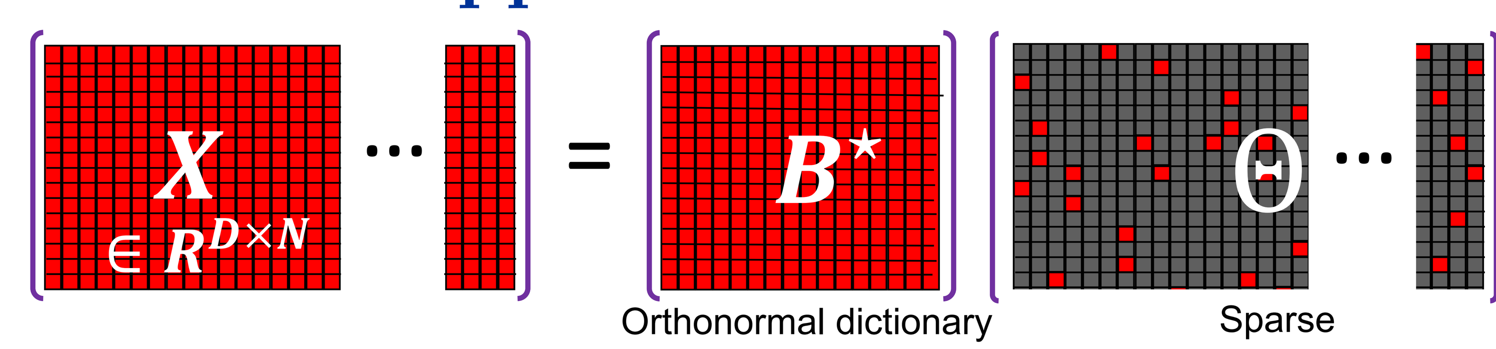
- **Theorem**: (*i*) DPCP satisfies the RRC, (*ii*) RSGM with a suitable init. converges to an orthonormal basis of $\mathcal{S}^\perp$ at an R-linear rate, and (*iii*) SVD gives a valid init. (when $M \lesssim N^2/dD$ in a random spherical model)

$D = 100$
$c = 10$
$N = 1500$
$\frac{M}{M+N} = 0.7$



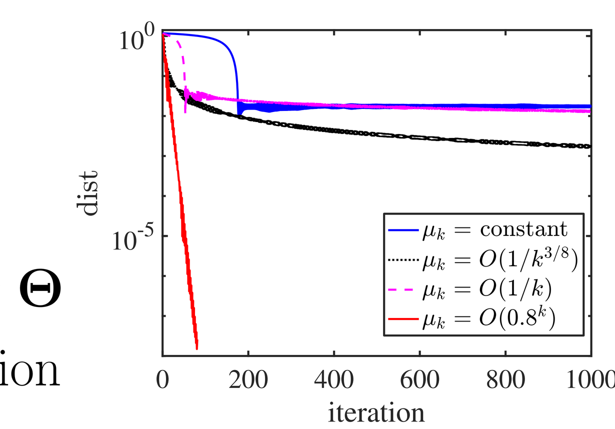


Application to ODL

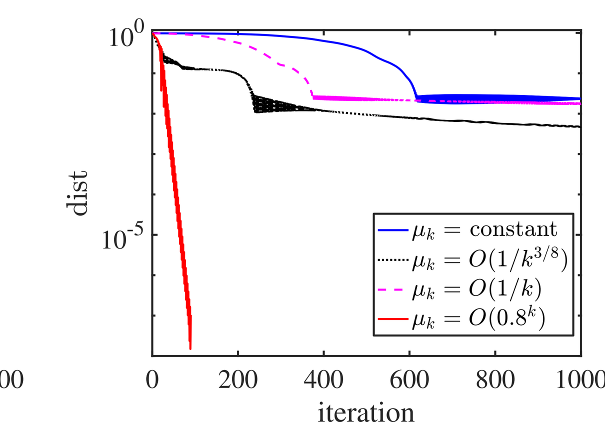


- Find one column $\boldsymbol{b}^\star$ of $\boldsymbol{B}^\star$ by solving $\min_{\boldsymbol{b}\in\mathbb{O}(1,D)} \|\boldsymbol{b}^\top\boldsymbol{X}\|_1$
- ODL satisfies the RRC [1]

$D = 70$
$N = 5857 \approx 10D^{1.5}$
sparsity level 0.3 for $\boldsymbol{\Theta}$
A random initialization



$\min_{\boldsymbol{b}\in\mathbb{O}(1,D)} \|\boldsymbol{b}^\top\boldsymbol{X}\|_1$



$\min_{\boldsymbol{B}\in\mathbb{O}(D,D)} \|\boldsymbol{B}^\top\boldsymbol{X}\|_1$

- Find all columns of $\boldsymbol{B}^\star$ by solving $\min_{\boldsymbol{B}\in\mathbb{O}(D,D)} \|\boldsymbol{B}^\top\boldsymbol{X}\|_1$ (using RSGM on the Stiefel manifold instead of Grassmannian)

[1] Bai, Jiang & Sun, Subgradient Descent Learns Orthogonal Dictionaries, In *ICLR*, 2019.
[2] Zhu et al., Dual Principal Component Pursuit: Improved Analysis and Efficient Algorithms, In *NeurIPS*, 2019.
[3] Davis et al., Subgradient Methods for Sharp Weakly Convex Functions, In *J. Optim. Theory*, 2018.
[4] Tsakiris & Vidal, Dual principal component pursuit, In *JMLR*, 2018
[5] Maunu, Zhang & Lerman, A well-tempered landscape for non-convex robust subspace recovery, In *JMLR*, 2019

This research is supported in part by NSF grant 1704458, ShanghaiTech grant 2017F0203-000-16, and the Northrop Grumman Mission Systems Research in Applications for Learning Machines (REALM) initiative.