Global Convergence of a Grassmannian Gradient Descent Algorithm for Subspace Estimation

Dejiao Zhang & Laura Balzano
University of Michigan, Ann Arbor

Low-rank Subspace Estimation

- Finding low-rank components to fit/approximate observations is a fundamental task in data analysis. It has been observed in a variety of contexts that gradient descent methods have great success in solving low-rank factorization problems, despite the relevant problem formulation being non-convex.
- We seek the d-dimensional subspace from a streaming data matrix. We propose an adaptive step size scheme to automatically adjust learning rates for a Grassmannian gradient algorithm, which maximizes our convergence metrics at each iteration for the noise free data, and yield monotonic improvement in terms of expectation for the noisy case.
- For the noise free data, we provide a **global convergence result for the proposed algorithm from a random initialization to the true subspace.**

Problem Formulation and Algorithmic Approaches

We receive a matrix $M = [x_1, x_2, \cdots, x_N] \in \mathbb{R}^{n\times N}$ with $x_t = v_t + \xi_t$, where v_t are generated by an d-dimensional subspace $\mathcal{S} \subset \mathbb{R}^n$, and $\xi_t \in \mathbb{R}^n$ is the noise.

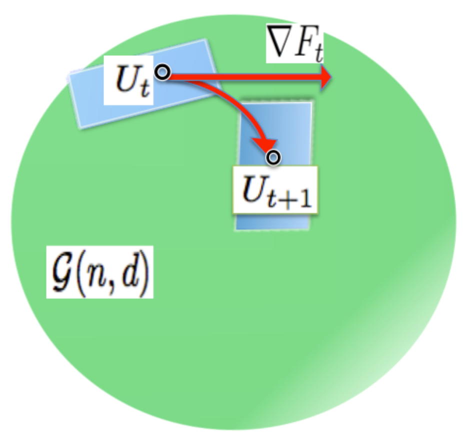
$$\begin{aligned} &\underset{U\in\mathbb{R}^{n\times d}, W\in\mathbb{R}^{N\times d}}{\text{minimize}} \quad \|UW^T - M\|_F^2 \quad \xrightarrow{\text{streaming data}} \quad \underset{U\in\mathbb{R}^{n\times d}}{\text{minimize}} \sum_{t=1}^{\infty} \min_{w_t} \|Uw_t - x_t\|_2^2 \\ &\text{s.t.} \quad \text{span}\,(U) \in \mathcal{G}(n,d) \qquad\qquad\qquad\qquad \text{s.t.} \quad \text{span}\,(U) \in \mathcal{G}(n,d) \end{aligned}$$

GROUSE ($U_t \to U_{t+1}$)

Given current estimate U_t and observation $x_t = v_t + \xi_t$ for $v_t \in R(\bar{U})$;

- Calculate weights: $w_t = \arg\min_w F_t(U_t) := \|U_t w - x_t\|_2^2$;
- Predict Projection and residual: $p_t = U_t w_t, \quad r_t = x_t - p_t$;
- Update subspace:

$$U_{t+1} = U_t + \left(\sin(\theta_t)\frac{r_t}{\|r_t\|} + (\cos(\theta_t) - 1)\frac{p_t}{\|p_t\|}\right)\frac{w_t^T}{\|w_t\|}$$



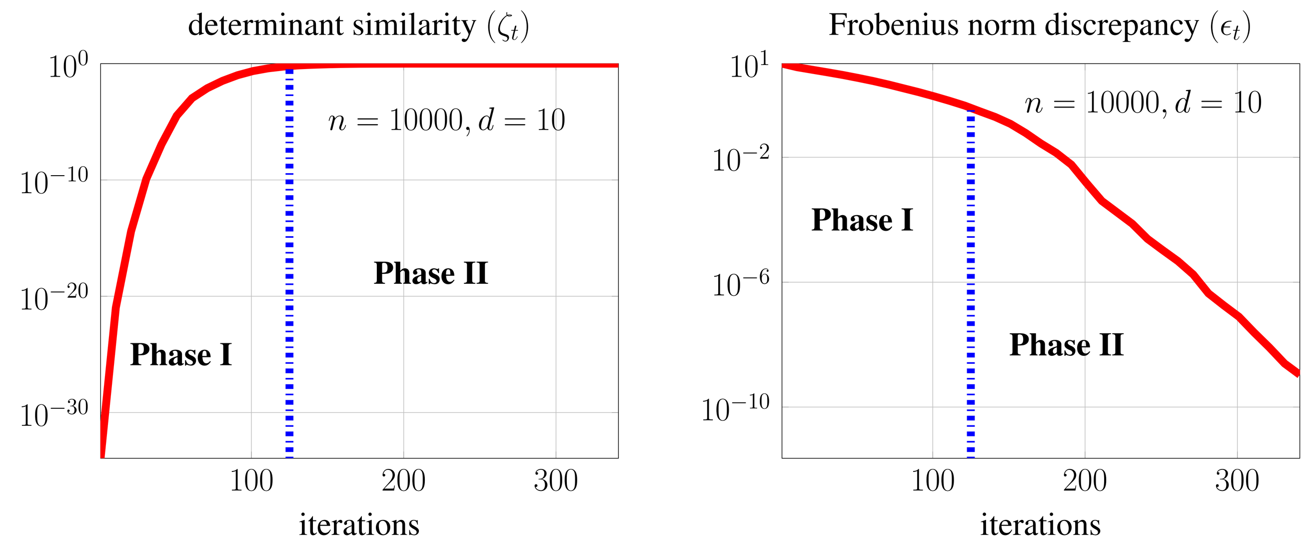
where $\theta_t = \arctan\left((1-\alpha_t)\frac{\|r_t\|}{\|p_t\|}\right)$ with $\alpha_t \in [0,1]$ depending on the statistics of the observations.

Convergence Analysis

Convergence Metrics: Let $\bar{U} \in \mathbb{R}^{n\times d}$ whose orthonormal columns span $\mathcal{S}$. Let $\Phi_{t,i}, i = 1,\ldots,d$ denote the i^{th} principal angle between subspaces Span(U_t) and Span$(\bar{U})$, then define

$$\zeta_t := \det(\bar{U}^T U_t U_t^T \bar{U}) = \prod_{i=1}^{d} \cos^2\phi_{t,i} \quad \text{and} \quad \epsilon_t := \sum_{i=1}^{d} \sin^2\phi_{t,i} = d - \|\bar{U}U_t^T\|_F^2 \,.$$

The motivation for us to analyze the convergence of GROUSE with two different metrics is that a faster convergence rate can be obtained by using different metrics in the initial and local phase.



Main Results

Assumptions. Let $v_t = \bar{U}s_t$ with $\mathbb{E}s_t = 0, \text{Cov}(s_t) = \mathbb{I}_d$. Further assume ξ_t is a Gaussian random vector i.i.d entries s.t. $\mathbb{E}\left[\|\xi_t\|^2/\|v_t\|^2 | v_t\right] \leq \sigma^2$.

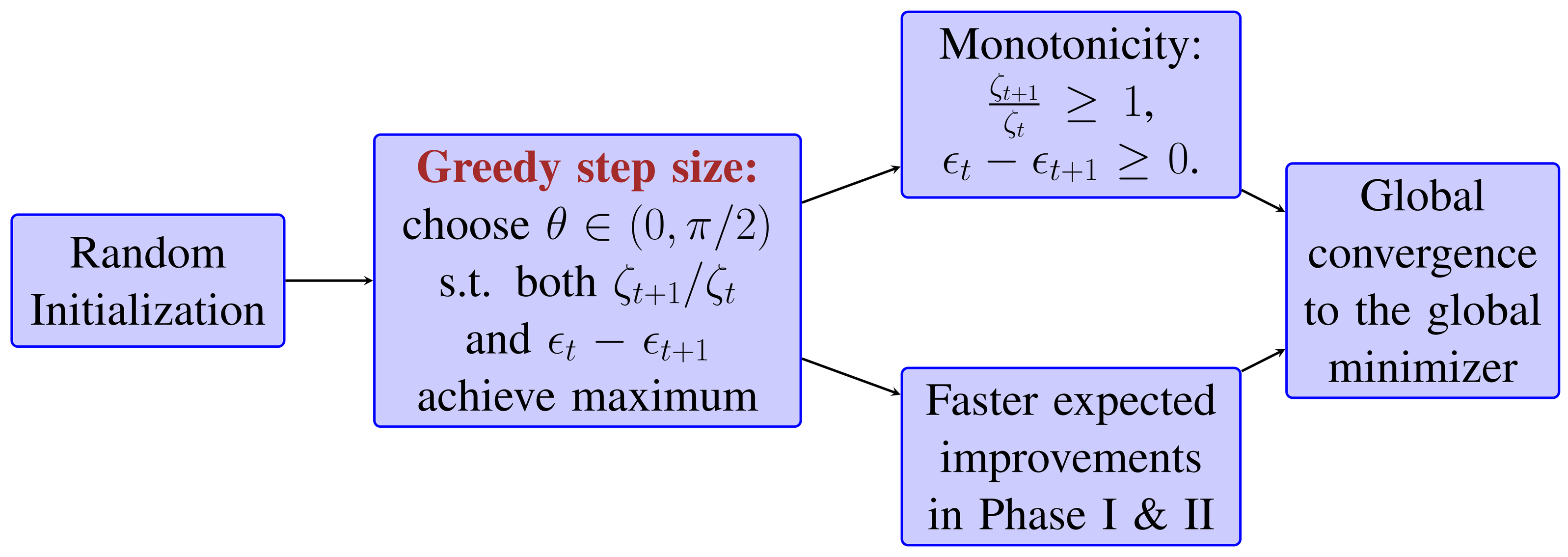
Theorem 1.(Noiseless case) Let ϵ^* be the given accuracy, for any $\rho, \rho' > 0$, after

$$K \geq K_1 + K_2 = \left(\frac{d^3}{\rho'} + d\right)\mu_0 \log(n) + 2d\log\left(\frac{1}{\epsilon^*\rho}\right)$$

iterations of GROUSE,

$$\mathbb{P}\left(\epsilon_K \leq \epsilon^*\right) \geq 1 - \rho' - \rho \,.$$

where $\mu_0 = 1 + \frac{\log\frac{(1-\rho')}{C} + d\log(e/d)}{d\log n}$ with $C > 0$.



Theorem 4 & 5. (Noisy case) After one iteration of GROUSE we have the following:

$$\begin{aligned} \mathbb{E}\left[\zeta_{t+1}|U_t\right] &\geq \left(1 + \gamma_1\frac{1-\zeta_t}{d}\right)\zeta_t \quad \text{(Thm 4)} \\ \mathbb{E}\left[\epsilon_{t+1}|U_t\right] &\leq \left(1 - \beta_0\frac{(\cos^2\phi_{t,d} - \gamma_2)}{d}\right)\epsilon_t \quad \text{(Thm 5)} \end{aligned}$$

where $\beta_0 = \frac{1}{1+\sigma^2 d/n}$, $\gamma_1 = \beta_0\left(1 - \frac{\sigma^2}{(1-\zeta_t)/d + \sigma^2}\right)$ and $\gamma_2 = \frac{(1-d/n)\sigma^2}{\epsilon_t/d + (1-d/n)\sigma^2}$. **($\beta_0 \to 1$, $\gamma_1, \gamma_2 \to 0$ as $\sigma^2 \to 0$).**

Numerical Results

Illustration of the bounds on K_1 and K_2 for noiseless convergence in Theorem 1. We run GROUSE to convergence for a required accuracy $\epsilon^* = 1e-4$ and divide the iterations into K_1, the number to reach $\zeta_t > \frac{1}{2}$, and K_2, the remaining number to reach $\epsilon_t < \epsilon^*$.

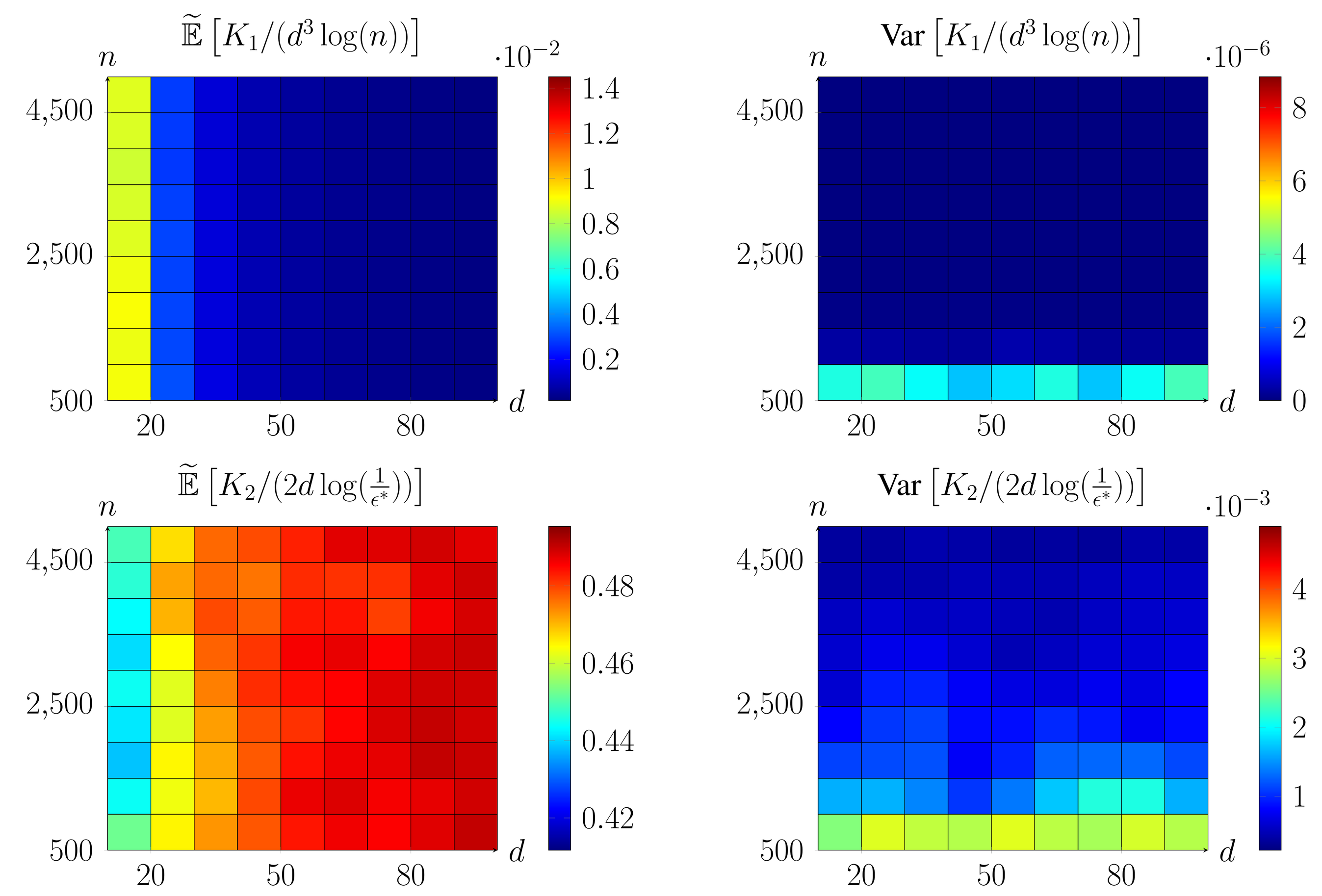
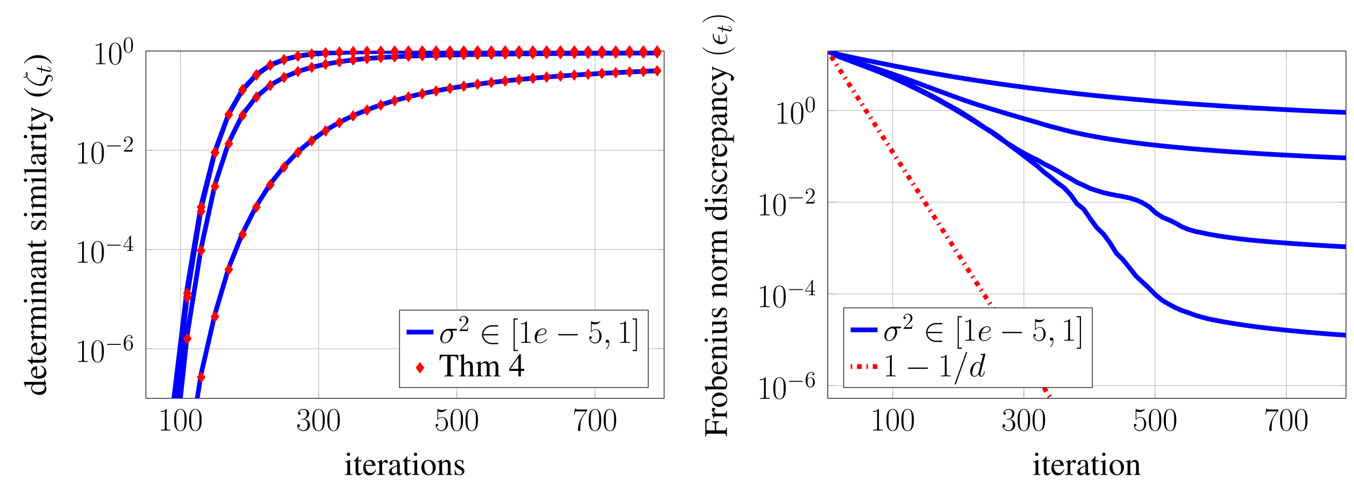


Illustration of expected convergence bounds given by Theorem 4&5 over 100 trials. In this simulation, $n = 2000$, $d = 20$ and $\sigma^2 \in \{1e-5, 1e-3, 1e-1, 1\}$.



Forthcoming Research

- We will complete the global convergence result for noisy data and extend it to more general noise model.
- We leave the global convergence results for undersampled data, including compressively sampled data and missing data as future work.

References

[1] Laura Balzano and Stephen J Wright. Local convergence of an algorithm for subspace identification from partial data. *Foundations of Computational Mathematics*, 15(5):1279–1314, 2015.

[2] Dejiao Zhang and Laura Balzano. Global convergence of a grassmannian gradient descent algorithm for subspace estimation. *arXiv preprint arXiv:1506.07405*, 2015.