Convergence Results of GROUSE

Dejiao Zhang
Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, USA
Email: dejiao@umich.edu

Laura Balzano
Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, USA
Email: girasole@umich.edu

I. Introduction

It has been observed in a variety of contexts that gradient descent methods have great success in solving low-rank matrix-factorization problems, despite the relevant problem formulation being non-convex. We consider in particular the application of natural first order incremental gradient descent method for subspace learning, constraining the gradient method to the Grassmannian $\mathcal{G}(n,d)$. This algorithm is called Grassmannian Rank-One Update Subspace Estimation (GROUSE). We have theoretical results on the guaranteed error improvement for two sampling cases: where each data vector of the streaming matrix is fully sampled or undersampled by a sampling matrix $A_t \in \mathbb{R}^{m\times n}$ with $m \ll n$. We propose an adaptive step size scheme, with which global convergence of GROUSE with fully sampled noiseless data is guaranteed, despite the nonconvex formulation and constraints. As for the undersampled data and noisy data, we prove that the proposed step size scheme yields monotonic improvement in expectation on the defined convergence metric.

We formulate the subspace estimation as a non-convex optimization problem as follows. Given are a sequence of observations $x_t = A_t(v_t + \xi_t)$ where $A_t \in \mathbb{R}^{m\times n}$ are known sampling matrices, $\xi_t \in \mathbb{R}^n$ are additive noise, and $v_t \in \mathbb{R}^n$ are drawn from an unknown d-dimensional subspace, spanned by $\bar{U} \in \mathbb{R}^{n\times d}$ with orthonormal columns. Let $U \in \mathbb{R}^{n\times d}$ be a matrix with orthonormal columns. Then we want to solve:

$$\begin{aligned} \underset{U\in\mathbb{R}^{n\times d}}{\text{minimize}} \quad & \sum_{t=1}^{T} \min_{w_t} \|A_t U w_t - x_t\|^2 \\ \text{subject to} \quad & \text{span}\,(U) \in \mathcal{G}(n,d) \end{aligned} \tag{1}$$

We study GROUSE (Fig 1) to solve the above problem, where we process one observation at a time and perform a rank-one update to generate a sequence of estimates U_t with the goal that $R(U_t) \to R(\bar{U})$, where $R(\cdot)$ denotes the column range.

II. Convergence Results

A. Full noiseless data

The following adaptive step size scheme is derived by maximizing the improvement on our convergence metric for each update of GROUSE with fully sampled noiseless data. More specifically, let $\boldsymbol{\zeta_t} = \det\left(\bar{U}^T U_t U_t^T \bar{U}\right)$ and suppose $A_t = \mathbb{I}_n, \zeta_t > 0$, then we have [1], [2]

$$\theta_t = \arg\max_{\theta} \ \zeta_{t+1}/\zeta_t = \arctan\left(\|r\|/\|p\|\right) \tag{2}$$

Given this step size scheme, global convergence of GROUSE is guaranteed for fully sampled noise-free data despite the non-convex formulation and constraints.

Theorem 1. *[1] Let $\zeta^* \in (0,1]$ be the desired accuracy of our estimated subspace. With fully sampled noiseless data, suppose the initialization is drawn uniformly from the Grassmannian $\mathcal{G}(n,d)$, then for any $\rho > 0$, after $K \geq \left(2d^2/\rho + 1\right)\tau_0 \log(n) + 2d\log\left(1/2\rho(1-\zeta^*)\right)$ (where $\tau_0 \approx 1$) updates of GROUSE, we obtain $\mathbb{P}\left(\zeta_K \geq \zeta^*\right) \geq 1 - 2\rho$.*

B. Undersampled noiseless data

For undersampled data, we consider two typical cases, missing data and compressively sampled data. We use the same step size scheme (2) as that for full data. Under mild conditions, we can prove that with probability exceeding $1 - n^{\delta d/2}(\delta \in (0,1))$, the following unified framework holds for both cases as long as we have $m \geq O(d\log n)$ samples:

$$\mathbb{E}\left[\zeta_{t+1}\middle|U_t\right] \geq \left(1 + \eta\frac{m}{n}\frac{1-\zeta_t}{d}\right)\zeta_t \tag{3}$$

where $\eta \approx 1$ is slightly different for each sampling type [1].

C. Weighted Step Size Scheme for Noisy Data

A weighted step size scheme [2] $\theta_t = \arctan\left((1-\alpha)\frac{\|r\|}{\|p\|}\right)$ allows similar results for noisy data, *i.e.,* $\xi_t \neq 0$. We restrict $\alpha \in [0,1)$ with the goal that $\alpha \to 1$ as $R(U_t) \to R(\bar{U})$. The intuition behind this strategy is that the noisy part will gradually dominate $\|r\|$, we hope to include less and less information from the projection residual to our estimations as $R(U_t) \to R(\bar{U})$.

Theorem 2. *Suppose the entries of ξ_t are independent and identically distributed Gaussian random variables such that $\mathbb{E}\left[\|\xi_t\|^2/\|v_t\|^2\middle|v_t\right] \leq \sigma^2$. Then with probability at least $1 - n^{\delta_1 d/2}(\delta_1 \in (0,1))$, we obtain*

$$\mathbb{E}\left[\zeta_{t+1}\middle|U_t\right] \geq \left(1 + \eta_1\frac{m}{n}\frac{1-\zeta_t}{d}\left(1 - \frac{\sigma^2}{\frac{(1-\zeta_t)}{d} + \sigma^2}\right)\right)\zeta_t \tag{4}$$

III. Conclusion

We have shown global convergence results for GROUSE with full noiseless data, and per-iteration improvement with noise or undersampling. Leveraging techniques in stochastic process theory, it may be possible to establish convergence results for all cases in terms of the number of iterations required before GROUSE first achieves a given accuracy.

REFERENCES

[1] D. Zhang and L. Balzano, "Convergence of a grassmannian gradient descent algorithm for subspace estimation from undersampled data," *arXiv preprint arXiv:1610.00199*, 2016.

[2] ——, "Global convergence of a grassmannian gradient descent algorithm for subspace estimation," *arXiv preprint arXiv:1506.07405*, 2015.

Algorithm 1 GROUSE: Grassmannian Rank-One Update Subspace Estimation

Given U_0, an $n \times d$ matrix with orthonormal columns, with $0 < d < n$;
Set $t := 0$;
repeat
  Given sampling matrix A_t and observation x_t;
  Define $w_t := \arg\min_w \|A_t U_t w - x_t\|_2^2$;
  Define $p_t := U_t w_t$; $r_t := A_t^T \widetilde{r}_t$ with $\widetilde{r}_t := x_t - A_t p_t$;
  Using step size

$$\theta_t = \arctan\left((1-\alpha_t)\frac{\|\widetilde{r}_t\|}{\|p_t\|}\right)$$

  where $\alpha_t = C\frac{\sigma^2}{1+\sigma^2}\left(1-\frac{d}{n}\right)\frac{\|x_t\|^2}{\|\widetilde{r}_t\|^2}$ with $C > 0$ and σ^2 denotes the upper bound for the noise level (Condition 1), update with a gradient step on the Grassmannian:

$$U_{t+1} := U_t + \left(\frac{y_t}{\|y_t\|} - \frac{p_t}{\|p_t\|}\right)\frac{w_t^T}{\|w_t\|}$$

  where

$$\frac{y_t}{\|y_t\|} = \frac{p_t}{\|p_t\|}\cos(\theta_t) + \frac{r_t}{\|r_t\|}\sin(\theta_t)$$

  $t := t + 1$;
until termination

Fig. 1. The GROUSE Algorithm

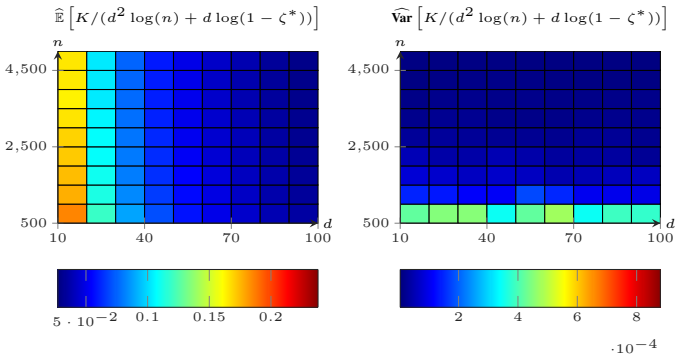

Fig. 2. Illustration of the bounds on K (Theorem 1) compared to their values in practice, averaged over 50 trials with different n and d. We run GROUSE to convergence for a required accuracy $\zeta_t = 1 - 1e\text{-}4$ and show the ratio of K to the bound described in Theorem 1, $d^2\log(n) + d\log\frac{1}{1-\zeta^*}$. We can see that, for fixed n, our theoretical results become more and more loose as we increase the dimension of the underlying subspace. However, compared to the empirical mean, the empirical variance is very small. This indicates that the relationship between our theoretical bounds and the actual iterations required by GROUSE is stable.

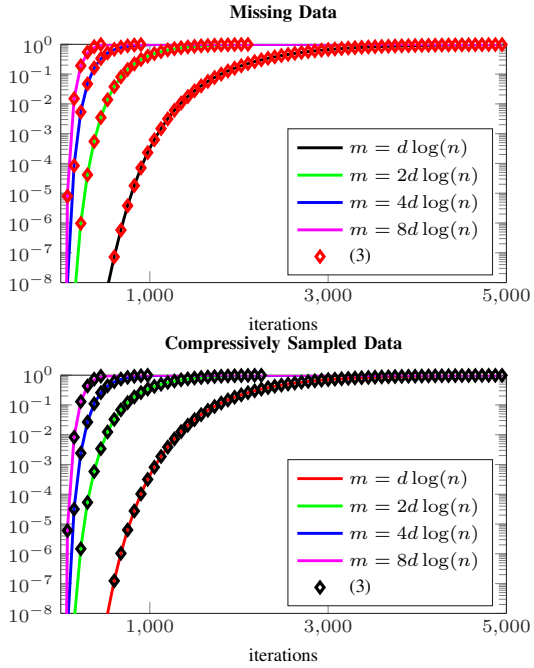

Fig. 3. We examine our theoretical result (3) for the expected improvement on ζ_t for the undersampled case. We set $n = 5000$, $d = 10$, and run GROUSE over different sampling numbers m. The plots are obtained by averaging over 50 trials. The diamonds denote the lower bound on expected convergence rates described in (3). We can see that our theoretical bounds on the expected improvement on ζ_t for both missing data and compressively sampled data are tight from any random initialization, although we have only established local convergence results for the missing data case [1] (global results for compressively sampled data).

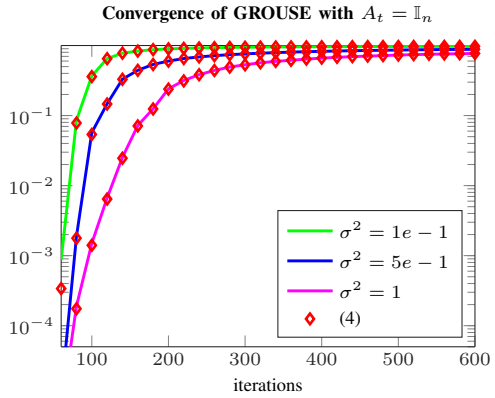

Fig. 4. Illustration of expected convergence bounds given by (4) over 50 trials. In this simulation, we set $n = 5000$, $d = 10$. We run GROUSE with multiple noise levels for the fully sampled case. The diamonds denote the lower bound on expected convergence rates described in (4). As we can see the expected improvement on ζ_t dumped by the presence of noise.