Global Convergence of a Grassmannian Gradient Descent Algorithm for Subspace Estimation

Dejiao Zhang
dejiao@umich.edu
University of Michigan, Ann Arbor

Laura Balzano
girasole@umich.edu
University of Michigan, Ann Arbor

Abstract

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 tackle a particular instance of this scenario, where we seek the d-dimensional subspace spanned by a streaming data matrix. We apply the natural first order incremental gradient descent method, constraining the gradient method to the Grassmannian. In this paper, we propose an adaptive step size scheme that is greedy for the noiseless case, that maximizes the improvement of our metric of convergence at each data index t, and yields an expected improvement for the noisy case. We show that, with noise-free data, this method converges from any random initialization to the global minimum of the problem. For noisy data, we provide the expected convergence rate of the proposed algorithm per iteration.

1 Introduction

Low-rank matrix factorization is one of the foundational tools of signal processing, numerical methods, and data analysis. Suppose we wish to factorize a matrix $M = UW^T$, imposing orthogonality constraints on U or W. Solving for such matrix factorizations can be computationally burdensome, and many algorithms that attempt to speed up computation are actually solving a non-convex optimization problem, therefore coming with few guarantees.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

The Singular Value Decomposition (SVD) is the solution to a non-convex optimization problem, and there are several highly successful algorithms for solving it [Golub and Van Loan, 2012]. Unfortunately, these algorithms cannot easily be extended to problems with regularizers or missing data. Recently, several results have been published with first-of-their-kind guarantees for a variety of different gradient-type algorithms on non-convex matrix factorization problems [Jain et al., 2013, De Sa et al., 2014, Armentano et al., 2014, Chen and Wainwright, 2015, Bhojanapalli et al., 2015, Zheng and Lafferty, 2015]. These new algorithms, being gradient-based, are well-suited to extensions of the original problem that include different cost functions or regularizers. For example, with gradient methods to solve the SVD we may be able to solve Robust PCA [Candès et al., 2011, He et al., 2012, Xu et al., 2010], Sparse PCA [d'Aspremont et al., 2008], or even ℓ_1 PCA [Brooks et al., 2013] with gradient methods as well.

Our contribution is to provide a global convergence result for d-dimensional subspace estimation using an incremental gradient algorithm performed on the Grassmannian, the space of all d-dimensional subspaces of $\mathbb{R}^n$. Subspace estimation is a special case of matrix factorization with orthogonality constraints, where we seek to estimate only the subspace spanned by the columns of the left matrix factor $U \in \mathbb{R}^{n\times d}$. Our result demonstrates that this gradient algorithm *converges globally* almost surely, *i.e.,* it converges from any random initialization to the global minimizer. To the best of our knowledge, this is the first global convergence result for an incremental gradient descent method on the Grassmannian. When there is no noise, we propose a greedy step size scheme that maximizes the improvements on the defined metrics of convergence. Given this, we provide a rate of convergence in two parts: slower convergence in an initial phase starting from the random initialization, and then linear convergence for a local region around the global minimizer, where our results match those in [Balzano and Wright, 2014].

For the noisy case, we propose a step-size regimen that is simply a weighted version of the step size for noise-free data, where the weights depend on the data and noise statistics. With this step size, we provide results guaranteeing monotonic improvements on the metrics of convergence in terms of expectation.

Incremental gradient descent is our focus, motivated by streaming data applications. There are many applications of subspace estimation and tracking in medical imaging, communications, and environmental science; see more in [Edelman et al., 1998, Balzano and Wright, 2014, Balzano, 2012]. Matrix factors with orthogonality constraints, such as those given by the SVD, are also used in several data applications: they provide a unique collection of low-dimensional projections for data visualization, capture directions of maximal variance so as to give useful insights into data structure, and allow compressed storage of massive datasets with a precise notion of loss in compression.

2 Formulation and Related Work

We may formulate subspace estimation as a non-convex optimization problem as follows. Let $M \in \mathbb{R}^{n\times N}$ be a matrix that we wish to approximate with a subspace of rank d, and solve:

$$\begin{aligned} \underset{U\in\mathbb{R}^{n\times d}, W\in\mathbb{R}^{N\times d}}{\text{minimize}} \quad & \|UW^T - M\|_F^2 \\ \text{subject to} \quad & \text{span}\,(U) \in \mathcal{G}(n,d) \end{aligned} \tag{1}$$

This problem is non-convex firstly because of the product of the two optimization variables U and W and secondly because the optimization is over the Grassmannian $\mathcal{G}(n,d)$, the non-convex set of all d-dimensional subspaces in $\mathbb{R}^n$. However, several methods[1] can find the global minimizer of this problem in polynomial time under a variety of assumptions on M.

In this paper, we are interested in approximating a streaming data matrix. At each step, we sample a column of M, denoted $x_t \in \mathbb{R}^n$. We consider the planted problem, where $x_t = v_t + \xi_t$ where ξ_t is noise and v_t is drawn from a continuous distribution with support on the true subspace, spanned by $\bar{U} \in \mathbb{R}^{n\times d}$ with orthonormal columns; $v_t = \bar{U}s_t$, $s_t \in \mathbb{R}^d$. When $\xi_t = 0$, we wish to find the U that minimizes

$$F(U) = \sum_{t=1}^{\infty} \min_{w_t} \|Uw_t - x_t\|_2^2 \,, \tag{2}$$

i.e., the span of the data vectors or the range of $\bar{U}$, denoted $R(\bar{U})$. When $\xi_t \neq 0$ we still discuss results in terms of the distance from $\bar{U}$. If we consider only $t = 1, \ldots, N$, Problem (2) is identical to Problem (1). The GROUSE algorithm (Grassmannian Rank-One Update Subspace Estimation) we analyze is shown as Algorithm 1, where we generate a sequence $\{U_t\}_{t=0,1,\ldots}$ of $n \times d$ matrices with orthonormal columns with the goal that $R(U_t) \to R(\bar{U})$ as $t \to \infty$. Each observed vector is used to update U_t to U_{t+1}, and we constrain the gradient descent method to the Grassmannian using a geodesic update [Edelman et al., 1998].

Because of the importance of the problem, it has been studied for decades, and there is a great deal of related work. We direct the reader to [Edelman et al., 1998, Balzano, 2012] for in-depth descriptions of algorithms and guarantees. We focus here on recent results that have *global convergence guarantees to the global minimizer* and study either gradient-type algorithms, algorithms that handle streaming data, or algorithms that maintain orthogonality constraints with manifold optimization.

First we discuss incremental methods. [De Sa et al., 2014] established the global convergence of a stochastic gradient descent method for the recovery of a positive definite matrix M in the undersampled case, where the matrix M is not measured directly but instead via linear measurements. They propose a step size scheme under which they prove global convergence results from a randomly generated initialization. Similarly, [Balsubramani et al., 2013] invokes a martingale-based argument to show the global convergence rate of the proposed incremental PCA method to the single top eigenvector in the fully sampled case. In contrast, [Arora et al., 2013] estimates the best d-dimensional subspace in the fully sampled case and provides a global convergence result by relaxing the non-convex problem to a convex one. We seek to identify the d dimensional subspace by solving the non-convex problem directly. Finally, our work is most related to [Balzano and Wright, 2014], which provides local convergence guarantees for GROUSE in both the fully sampled and undersampled case. Our work focuses on global convergence but only in the fully sampled case; we will extend the global convergence results to the undersampled case in future work.

Turning to batch methods, [Keshavan, 2012, Jain et al., 2013] provided the first theoretical

[1] For example, the power method can solve this problem if the top d singular values of M are distinct [Golub and Van Loan, 2012]. Specifically, considering $d = 1$, if the desired accuracy of the U output by the power method to the global minimizer is ϵ^*, and the first two singular values of M, $\sigma_1(M)$ and $\sigma_2(M)$ are distinct with the $\sigma_1(M) = c\sigma_2(M)$ for $c > 1$, then the power method converges in $O\left(\frac{\log(1/\epsilon^*)}{\log c}\right)$ iterations.

guarantee for an alternating minimization algorithm for low-rank matrix recovery in the undersampled case. Under typical assumptions required for the matrix recovery problems [Recht et al., 2010], they established geometric convergence to the global optimal solution. Earlier work [Keshavan et al., 2010, Ngo and Saad, 2012] considered the same undersampled problem formulation and established convergence guarantees for a steepest descent method (and a preconditioned version) on the full gradient, performed on the Grassmannian. [Chen and Wainwright, 2015, Bhojanapalli et al., 2015, Zheng and Lafferty, 2015] considered low rank semidefinite matrix estimation problems, where they reparamterized the underlying matrix as $M = UU^T$, and update U via a first order gradient descent method. However, all these results require batch processing and a decent initialization that is close enough to the optimal point, resulting in a heavy computational burden and precluding problems with streaming data. We study random initialization, and our algorithm has fast, computationally efficient updates that can be performed in an online context.

Lastly, several convergence results for optimization on general Riemannian manifolds, including several special cases for the Grassmannian, can be found in [Absil et al., 2009]. Most of the results are very general; they include global convergence rates to local optima for steepest descent, conjugate gradient, and trust region methods, to name a few. We instead focus on solving the problem in (2) and provide global convergence rates to the global minimum.

3 Convergence analysis

We analyze Algorithm 1. At each step, the algorithm receives a vector $x_t = v_t + \xi_t \in \mathbb{R}^n$ such that $v_t = \bar{U} s_t$, $s_t \in \mathbb{R}^d$ and ξ_t is zero mean Gaussian noise. The algorithm then outputs an $n \times d$ matrix U_t with orthonormal columns at each iteration. We wish to recover $\bar{U}$, *i.e.,* the minimizer of Equation (2) when there is no noise. We would like to emphasize that in this scenario in a real application one would use the ISVD or a Gram-Schmidt procedure, but we seek convergence results for the Grassmannian gradient descent algorithm so that extensions can be made; *e.g.,* we may regularize the cost function or we may minimize some other function of the data. Reliable global convergence of the GROUSE algorithm has been observed empirically, despite the fact that the algorithm is solving a non-convex problem and operating on a non-convex manifold.

Algorithm 1 takes each vector x_t, forms the gradient of $\min_w \|Uw - x_t\|_2^2$, and takes a step in the direction of the negative gradient. The step is taken along the Grassmannian, the manifold of all d-dimensional subspaces of $\mathbb{R}^n$, and according to the step size described and justified below. In words, the algorithm works as follows: First we project our data vector onto the current subspace iterate to get the projection p_t. Then we calculate the residual r_t. The update to our subspace estimate U_t then requires only the addition of a rank-one matrix, as can be seen in Equation (4). This update is derived and explained in further detail in [Balzano et al., 2010, Edelman et al., 1998]. The rank-one update tilts U_t to no longer contain p_t but instead contain a linear combination of p_t and r_t; in other words, it moves U_t towards the observation v_t.

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 observation $x_t = v_t + \xi_t$ for $v_t \in R(\bar{U})$;
  Define $w_t := \arg\min_w \|U_t w - x_t\|_2^2$;
  Define $p_t := U_t w_t$; $r_t := x_t - U_t w_t$;
  Using step size

$$\theta_t = \arctan\left((1-\alpha_t)\frac{\|r_t\|}{\|p_t\|}\right), \tag{3}$$

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

$$U_{t+1} := U_t + \left(\frac{y_t}{\|y_t\|} - \frac{p_t}{\|p_t\|}\right)\frac{w_t^T}{\|w_t\|} \tag{4}$$

  where

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

  $t := t + 1$;
until termination

Before we present our main results on the convergence of the GROUSE algorithm, we first call out the following definitions and condition that will be used throughout our analysis.

Definition 1 (Principal Angles)**.** *We use $\phi_i\left(\bar{U}, U_t\right), i = 1, \ldots, d$ to denote the principal angles between subspaces $R(U_t)$ and $R(\bar{U})$, which are defined [[Stewart and Sun, 1990], Chapter 5] by $\cos\phi_i(\bar{U}, U_t) = \sigma_i(\bar{U}^T U_t)$.*

Definition 2 (Determinant similarity)**.** *Our first metric is $\zeta_t \in [0, 1]$, which measures the similarity between*

two subspaces and is defined as

$$\zeta_t := \det(\bar{U}^T U_t U_t^T \bar{U}) = \prod_{i=1}^{d} \cos^2 \phi_i(\bar{U}, U_t) \,. \quad (5)$$

Definition 3 (Frobenius norm discrepancy)**.** *Our second metric is $\epsilon_t \in [0, d]$, which measures the discrepancy between $R(U_t)$ and $R(\bar{U})$, and is defined as*

$$\epsilon_t := \sum_{i=1}^{d} \sin^2 \phi_i(\bar{U}, U_t) = d - \|\bar{U}^T U_t\|_F^2 \,. \quad (6)$$

Condition 1. *The inputs of GROUSE are $x_t = v_t + \xi_t$ where $v_t = \bar{U} s_t$ with $\mathbb{E}s_t = 0, Cov(s_t) = \mathbb{I}_d$, and ξ_t is a Gaussian random vector with entries being independently normal random variables such that $\mathbb{E}\left[\|\xi_t\|^2/\|v_t\|^2 \middle| v_t\right] \leq \sigma^2$. Further, we assume the energy of the underlying signals are finite,* i.e., *$\|v_t\|^2 < \infty$.*

3.1 Optimal Adaptive Step Size

In this section, we first derive a greedy step size scheme for each iteration t that maximizes the improvement on the defined metrics (ϵ_t, ζ_t) of convergence for the noiseless case, *i.e.,* $x_t = v_t$. Let $v_{t,\|}$ and $v_{t,\perp}$ denote the projection and residual of v_t onto $R(U_t)$. Then after each update we have the following (Appendix C):

$$\frac{\zeta_{t+1}}{\zeta_t} = \left(\frac{\|v_{t,\|}\|}{\|v_t\|}\cos\theta_t + \frac{\|v_{t,\perp}\|}{\|v_t\|}\sin\theta_t\right)^2 \quad (7a)$$

$$\epsilon_t - \epsilon_{t+1} = \frac{\left\|\bar{U}^T y_t\right\|^2}{\|y_t\|^2} - \frac{\|\bar{U}^T v_{t,\|}\|^2}{\|v_{t,\|}\|^2} \quad (7b)$$

with $\frac{y_t}{\|y_t\|} = \frac{v_{t,\|}}{\|v_{t,\|}\|}\cos\theta_t + \frac{v_{t,\perp}}{\|v_{t,\perp}\|}\sin\theta_t$. It follows that

$$\theta_t^* = \arg\max_{\theta} \frac{\zeta_{t+1}}{\zeta_t} = \arctan\left(\frac{\|v_{t,\perp}\|}{\|v_{t,\|}\|}\right)$$

This is equivalent to (3) for the noise-free case setting $\alpha_t = 0$. Using θ_t^*, we obtain monotonic improvement on the determinant increment $\frac{\zeta_{t+1}}{\zeta_t} = 1 + \frac{\|v_{t,\perp}\|^2}{\|v_t\|^2} \geq 1$. For the Frobenius norm discrepancy, we obtain $\epsilon_{t+1} - \epsilon_t = 1 - \frac{\left\|\bar{U}^T v_{t,\|}\right\|^2}{\|v_{t,\|}\|^2}$; that is, ϵ_t also achieves its maximal improvement. Therefore, when there is no noise in the observations, the proposed step size scheme described by (3) implies greedy learning rates with respect to the defined metrics (ϵ_t, ζ_t) of convergence.

For the noisy case, we propose a weighted step size schedule by restricting $\alpha_t \in (0, 1]$ with the goal that $\alpha_t \to 1$ as our estimated subspace $R(U_t)$ gradually converges to the true subspace $R(\bar{U})$. The intuition behind this strategy is that, choosing the step size in Equation (3), the update of GROUSE follows as

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

for which we have, if the noise is Gaussian distributed, $\|r_t\|^2 \sim \|v_{t,\perp}\|^2 + (1 - d/n)\|\xi_t\|^2$ (where by $a \sim b$ we mean a concentrates around b), hence the noise part will gradually dominate the projection residual as $R(U_t) \to R(\bar{U})$. It is therefore natural for us to consider incorporating less and less of the residual information into $R(U_t)$ over time. Therefore, we propose the following schedule for α:

$$\alpha_t = 1 - \frac{\|v_{t,\perp}\|^2}{\|r_t\|^2} = \frac{c\sigma^2}{1+\sigma^2}\left(1 - \frac{d}{n}\right)\frac{\|x_t\|^2}{\|r_t\|^2} \quad (8)$$

where $c > 0$. As we will show in Section 4, with this weighted learning rate scheme, we obtain improvements in expectation on both ζ_t and ϵ_t.

3.2 Convergence Without Noise

In this section, we consider the noise-free case, that is $x_t = v_t$ and $v_t \in R(\bar{U})$. The step size (Eq (3)) used in this section has $\alpha_t = 0$ for all iterations. We provide analysis of the algorithm in two separate phases. In the first phase the GROUSE algorithm will converge to a local region of the global optimal point from a random initialization within $O(d^3 log(n))$ iterations. From there, in the second phase GROUSE converges linearly to the optimal point. In each phase we use a different metric of convergence, which helps us obtain an overall faster convergence rate as compared to other work. The convergence rate with respect to only either determinant [De Sa et al., 2014] or Frobenius norm discrepancy [Jain et al., 2013] is either much slower within the local region [De Sa et al., 2014] or slower in an initial phase from random initialization [Jain et al., 2013]. This is demonstrated numerically in Figure 1.

Theorem 1 (Global Convergence of GROUSE)**.** *Suppose Condition 1 and that no noise is contained in the observations,* i.e., *$x_t = v_t$. Let $\epsilon^* > 0$ be the desired accuracy of our estimated subspace using the metric in Definition 3. Initialize the starting point U_0 of GROUSE as the orthonormalization of an $n \times d$ matrix with entries being standard normal variables. Then for any $\rho, \rho' > 0$, after*

$$\begin{aligned} K &\geq K_1 + K_2 \\ &= \left(\frac{d^3}{\rho'} + d\right)\mu_0 \log(n) + 2d\log\left(\frac{1}{\epsilon^*\rho}\right) \end{aligned} \quad (9)$$

iterations of GROUSE (Algorithm 1),

$$\mathbb{P}\left(\epsilon_K \leq \epsilon^*\right) \geq 1 - \rho' - \rho \,. \quad (10)$$

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

The proof of Theorem 1 is a direct combination of our analysis in two phases of the algorithm, stated in Theorem 2 and Theorem 3 below.

Theorem 2 (Initial convergence of the determinant similarity ζ_t to $\frac{1}{2}$)**.** *Under the conditions of Theorem 1, for any* $\rho' \in (0,1)$*, after*

$$K_1 \geq \left(\frac{d^3}{\rho'} + d\right)\mu_0 \log(n)$$

iterations of GROUSE (Algorithm 1),

$$\mathbb{P}\left(\zeta_{K_1} \geq \frac{1}{2}\right) \geq 1 - \rho'$$

where μ_0 *is the same as that in Theorem 1.*

Analyzing the determinant similarity turns out to be the key to proving convergence in this initial phase of GROUSE. The determinant similarity increases quickly toward 1 in the first phase. This also gives insight into how the GROUSE algorithm manages to seek the global minimum of a non-convex problem formulation: GROUSE is not attracted to stationary points that are not the global minimum. For our problem, all other stationary points U_{stat} have $\det(\bar{U}^T U_{stat} U_{stat}^T \bar{U}) = 0$, because they have at least one direction orthogonal to $\bar{U}$ [Balzano, 2012]. If the initial point U_0 has determinant similarity with $\bar{U}$ strictly greater than zero, and GROUSE increases the determinant similarity monotonically (as we mentioned in Section 3.1 and prove in Section 4), then we are guaranteed to stay away from other stationary points. Since we initialize GROUSE using U_0 uniformly from the Grassmannian, as the orthonormal basis of a random matrix $V \in R^{n\times d}$ with entries being independent standard Gaussian random variables, we guarantee $\zeta_0 > 0$ with probability one.

Theorem 3 (Local convergence of the Frobenius norm discrepancy ϵ_t to 0)**.** *Suppose at iteration* k *we have* $\zeta_k \geq 1/2$*. Then for any* $\rho \in (0,1)$ *and given accuracy* ϵ^**, after*

$$K_2 \geq 2d\log\left(\frac{1}{\epsilon^*\rho}\right)$$

additional iterations of GROUSE Algorithm 1, we have

$$\mathbb{P}(\epsilon_{k+K_2} \leq \epsilon^*) \geq 1 - \rho\,.$$

In the first phase, we require $O\left(d^3\log(n)/\rho'\right)$ iterations to reach the local region of the global minimum, where $1-\rho'$ is the probability with which we'll reach the local region. In simulations (Section 5, Figure 2) with isotropic Gaussian data vectors from the subspace, we actually see that $O(d^3\log(n))$ iterations are many more than enough to reach the local region, without fail. Our analysis, though, only requires zero-mean uncorrelated identically distributed random data vectors. Bounds on higher moments may admit a tighter analysis, which we leave for future work.

The second phase only requires $O(d\log(1/\epsilon^*\rho))$ iterations to converge to ϵ^* accuracy in the Frobenius norm discrepancy metric given in Definition 3. This result is true to what we see in practice, as you can see in Figure 2. The analysis behind this result provides a tighter version of [[Balzano and Wright, 2014], Theorem 3.2] that both grows the local region of convergence and (slightly) improves the rate to be less dependent on the current value of ϵ_t.

3.3 Convergence With Noise

In this section, we study the convergence behavior of GROUSE with noise in each observation. Unlike the noise-free case, here we only provide *expected monotonic* improvements of our convergence metrics. As we prove in the appendix, the results we present here also imply the corresponding ones for the noiseless data.

Theorem 4 (Expected convergence rate of the determinant similarity ζ_t)**.** *Given Condition 1 is satisfied, after one iteration of GROUSE we have the following improvement of the determinant similarity in expectation:*

$$\mathbb{E}\left[\zeta_{t+1}\middle|U_t\right] \geq \left(1 + \beta_0\frac{1-\zeta_t}{d}\left(1 - \frac{\sigma^2}{\frac{1-\zeta_t}{d} + \sigma^2}\right)\right)\zeta_t$$

where $\beta_0 = \frac{1}{1+\frac{d}{n}\sigma^2}$.

This theorem implies that the expected convergence rate of determinant similarity is damped by the presence of noise. To be more specific, rewrite the expected improvement as $\mathbb{E}\left[\zeta_{t+1}\middle|U_t\right] \geq \left(1 + \frac{\beta_0}{(1-\zeta_t)/d+\sigma^2}\left(\frac{1-\zeta_t}{d}\right)^2\right)\zeta_t$. We can see that, comparing with the noiseless case, for small SNR (large σ^2), the expected increment on ζ_t is approximately scaled by $\frac{1-\zeta_t}{d} < \frac{1}{d}$. Hence the theoretical bound on the iterations necessary to achieve given accuracy ζ^* in the small SNR case should roughly be at least d times that required by the noiseless case. For large SNR (small σ^2), the expected convergence rate is close to that of the noise-free case, as long as ζ_t is not too close to 1. Therefore, the required iterations to arrive at the local region of the true subspace should be close to that in the noiseless case. We show the corresponding numerical illustrations in Figure 1 and Figure 3.

Theorem 5 (Expected convergence rate of the Frobenius norm discrepancy ϵ_t)**.** *Under Condition 1, we obtain the following upper bound on the decrease of Frobenius norm discrepancy ϵ_t in expectation:*

$$\mathbb{E}\left[\epsilon_{t+1}\middle|U_t\right] \leq \left(1 - \frac{\beta_0}{d}\left(\cos^2\phi_{t,d} - \frac{\beta_1\sigma^2}{\frac{\epsilon_t}{d} + \beta_1\sigma^2}\right)\right)\epsilon_t$$

where $\beta_0 = \frac{1}{1+\frac{d}{n}\sigma^2}$, $\beta_1 = 1 - \frac{d}{n}$, and $\phi_{t,d}$ is the largest principal angle between $R(U_t)$ and $R(\bar{U})$.

As indicated by Theorem 4, the expected convergence rate will slow down as ζ_t increases. However, the above theorem implies that for large SNR (small σ^2), once we enter the local region of the true subspace, the convergence rate of the Frobenius discrepancy will take over. Specifically, when $\cos^2\phi_{t,d} > 1/2$, the convergence rate of ϵ_t can be bounded from below by $1 - \left(\frac{1}{2} - \frac{1}{d}\right)\frac{1}{d}$ as long as $\epsilon_t \geq d^2\sigma^2$. Therefore, an implication of Theorem 5 is that GROUSE will converge to a ball centered on the true subspace whose radius is determined by the noise level and subspace dimension. The convergence rate will slow as GROUSE approaches this ball. On the other hand, since $1 - \epsilon_t \leq \cos^2\phi_{t,d} \leq 1 - \epsilon_t/d$, by a simple calculation we can see that for small SNR (large σ^2), the fast local convergence never kicks in. In that case, we only study the convergence behavior of GROUSE in terms of the determinant similarity ζ_t.

As we mentioned previously, with noise the improvement is not monotonic for either determinant similarity (ζ_t) or Frobenius norm discrepancy (ϵ_t). This is a hurdle to pass before we can provide similar global convergence results as we obtained for the noise-free case (Theorem 1). However, by leveraging techniques in stochastic process theory, it might be possible to establish asymptotic convergence results or even non-asymptotic convergence results in terms of the number of iterations required before GROUSE first achieves a given accuracy. We leave this as future work.

4 Supporting Theory

We first call out the following lemma to quantify the expectation of the determinant similarity between our random initialization and the true subspace. For convenience, we will drop the subscript of all terms except ϵ_t, ζ_t and U_t hereafter.

Lemma 1. *[Nguyen et al., 2014] Initialize the starting point U_0 of GROUSE as the orthonormalization of an $n \times d$ matrix with entries being standard normal variables. Then*

$$\mathbb{E}[\zeta_0] = \mathbb{E}\left[\det(U_0^T\bar{U}\bar{U}^TU_0)\right] = C\left(\frac{d}{ne}\right)^d$$

where $C > 0$ is a constant approximately equal to 1.

As we mentioned in Section 3.1, both the determinant similarity ζ_t and the Frobenius discrepancy ϵ_t improve monotonically in the noiseless case. We formally present this in the following lemma.

Lemma 2 (Monotonic results for the noiseless case)**.** *When there is no noise, given the step size in Eq (3), after one update of GROUSE we obtain*

$$\frac{\zeta_{t+1}}{\zeta_t} = 1 + \frac{\|v_\perp\|^2}{\|v_\|\|^2}, \quad \text{and} \quad \epsilon_t - \epsilon_{t+1} = 1 - \frac{\|\bar{U}\bar{U}^Tv_\|\|^2}{\|v_\|\|^2}$$

where $v_\|$ and $v_\perp$ denote the projection and residual of v onto $R(U_t)$.

For the noisy case, we provide the following lemmas, which are the intermediate results that allow us to establish the expected improvements on both ζ_t and ϵ_t in Section 3.3.

Lemma 3. *Given Condition 1 is satisfied, after one update of GROUSE we obtain the following*

$$\mathbb{E}\left[\zeta_{t+1}\middle|U_t\right] \geq \left(1 + \mathbb{E}\left[(1-\alpha)^2\frac{\|r\|^2}{\|p\|^2}\middle|U_t\right]\right)\zeta_t$$

Lemma 4. *After one iteration of the GROUSE algorithm, we have the following*

$$\mathbb{E}\left[\epsilon_t - \epsilon_{t+1}\middle|U_t\right] = \mathbb{E}\left[1 - \mathcal{R} - \frac{\|\bar{U}\bar{U}^Tp\|^2}{\|p\|^2}\middle|U_t\right]$$

where $\mathcal{R} = \frac{\|(I-\bar{U}\bar{U}^T)(\xi-\alpha r)\|^2}{\|v+\xi-\alpha r\|^2}$.

According to our definition of α in Section 3.1, we can see that when $R(U_t)$ is not close to $R(\bar{U})$, $1-\alpha$ is large, as is $\|r\|^2/\|p\|^2$. Therefore, Lemma 3 implies that the expected convergence rate of the determinant similarity (ζ_t) is faster in the first phase. For the Frobenius norm discrepancy (ϵ_t), comparing to the noiseless case where $p = v_\|$, Lemma 4 implies that we obtain monotonic expected decrease in Frobenius norm discrepancy as long as we are outside a ball centered on the true subspace. This ball shrinks as $\sigma^2 \to 0$, with no such constraint for $\sigma^2 = 0$. As we approach this ball, the expected convergence rate slows.

5 Numerical Results

With our plots we illustrate why the two analysis approaches allow us to prove rates in both phases of GROUSE. For each numerical result in this section, we initialize GROUSE with orthonormalized Gaussian matrices with entries iid $\mathcal{N}(0,1)$. The underlying subspace of each trial is set to be a sparse subspace, as

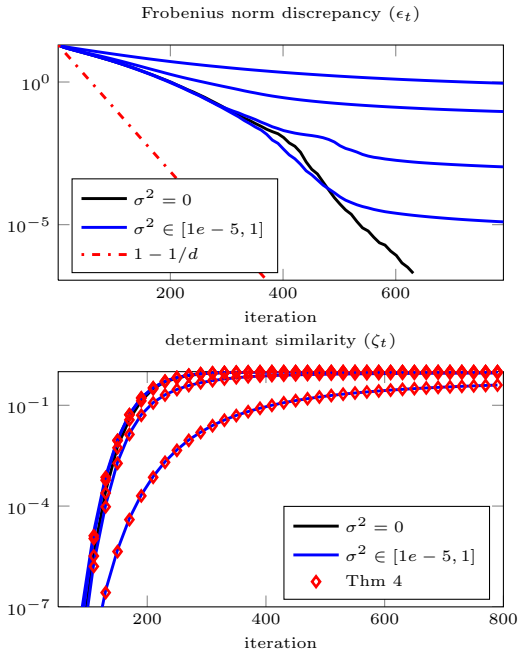

Figure 1: Illustration of expected convergence bounds given by Theorem 5 (top) and Theorem 4 (bottom) over 100 trials. In this simulation, $n = 2000$, $d = 20$ and $\sigma^2 \in \{0, 1e-5, 1e-3, 1e-1, 1\}$. The red dashed line indicates the linear convergence rate, while the diamonds denote the lower bound on expected convergence rate in Theorem 4. We can see that the convergence rate of each phase slows down in the noisy case. However, when σ^2 is small, the convergence behavior of GROUSE similar to that of the noise-free case. We get a faster convergence rate of ζ_t in the initial phase an almost linear convergence of ϵ_t in the local region of $R(\bar{U})$.

the range of an $n \times d$ matrix $\bar{U}$ with sparsity on the order of $\frac{\log(n)}{n}$. We generate the coefficient matrix W with entries i.i.d $\mathcal{N}(0, 1)$. For the noisy case, we then normalize the columns of the underlying matrix $\bar{U}W$ and add a noise matrix N, with $N_{ij} \sim \mathcal{N}(0, \sigma^2/n)$. In the noisy case, we run GROUSE with the step size described in Equation (8), where we set c to its expected value 1.

As is demonstrated in Figure 1, when there is no noise in the observations or the SNR is large enough, the determinant similarity (ζ_t) increases quickly in the initial phase, while the Frobenius norm discrepancy (ϵ_t) decreases slowly. Then in a local region of the true subspace, our accurate bound on the fast convergence of the Frobenius norm discrepancy takes over. However, if the SNR is small, the convergence rate of the Frobenius norm discrepancy slows down; in this scenario we only study the convergence of GROUSE in terms of the determinant similarity. In Figure 1, we show that the convergence rate of determinant similarity will also slow down as we increase the magnitude of σ^2, however, the convergence rate described in Theorem 4 is still tight. This allows us to obtain a good enough approximation of the number of iterations required to reach a ball around $R(\bar{U})$, which is captured by K_1 alone in this case.

We next examine the tightness of our theoretical values of K_1 and K_2 for noiseless convergence in Figure 2. 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 $\epsilon_t < \epsilon^*$. We show the ratio of K_1 to the bound $d^3 \log(n)$ in the initial phase (top plot) and the ratio of K_2 to the bound $d \log(1/\epsilon^*)$ in the local phase (bottom plot). We run 50 trials and show the mean and variance. We can see that the value for K_1 is very loose. On the other hand, the value for K_2 is very accurate; $O(d \log(1/\epsilon^*))$ iterations are required to get to accuracy ϵ^*.

Finally, we examine the tightness of approximated K_1 and K_2 for the noisy case in Fig 3. As we mentioned in Section 3.3, for small SNR (large σ^2), the necessary number of iterations to achieve the given accuracy should be roughly d times that required by the noise-free case, while for large SNR (small σ^2), this ratio would be less. For large SNR, we first run GROUSE to reach the local region of the true subspace, *i.e.,* $\zeta_{K_1} \geq \frac{1}{2}$, and record K_1; from this point we run GROUSE to converge to $\epsilon^* = \tau_1 \frac{d^2}{n} \sigma^2$ and then record K_2 and compare it with that required by the noise-free case. For small SNR (large σ^2), we only numerically examine the convergence rate of the first phase, *i.e.,* necessary iterations to achieve the given accuracy $\zeta_{K_1} \geq \left(1 - \tau_2 \frac{d}{n} \sigma^2\right)^d \approx e^{-\tau_2 d^2 \sigma^2 / n}$. As we can see in Figure 3, we test K_1 versus $O(d^3 \log(n))$, and as in noiseless case the bound on K_1 is loose. For small noise, the bound on K_2 is tight and stable.

6 Conclusion

This paper has provided the first global convergence result for an incremental gradient descent method on the Grassmannian for noise-free data. For optimizing a particular cost function (2) in the noiseless case, we showed that the gradient algorithm converges from any random initialization to the global minimizer. Our novel analysis shows the convergence happens in two phases: the initial convergence and the local convergence. In the initial phase, we provided a very loose bound on the number of iterations $K_1 = O(d^3 \log(n)/\rho')$ required to get to a local region of the global minimizer from the random initialization

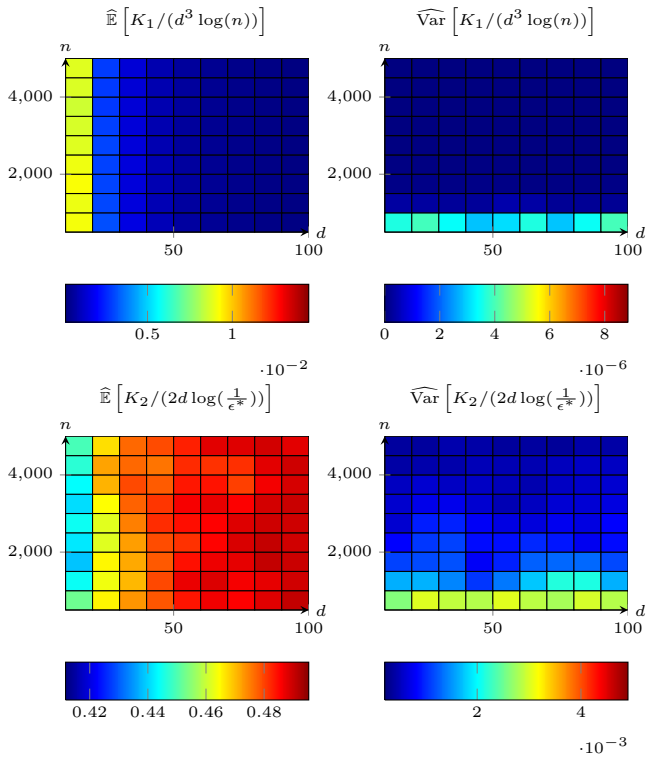

Figure 2: Illustration in the noise-free case of the bounds on K_1 and K_2 compared to their values in simulation, averaged over 50 trials with different n and d. We show the ratio of K_1 to the bound $d^3\log(n)$ in the initial phase (left) and the ratio of K_2 to the bound $d\log(1/\epsilon^*)$ in the local phase (right).

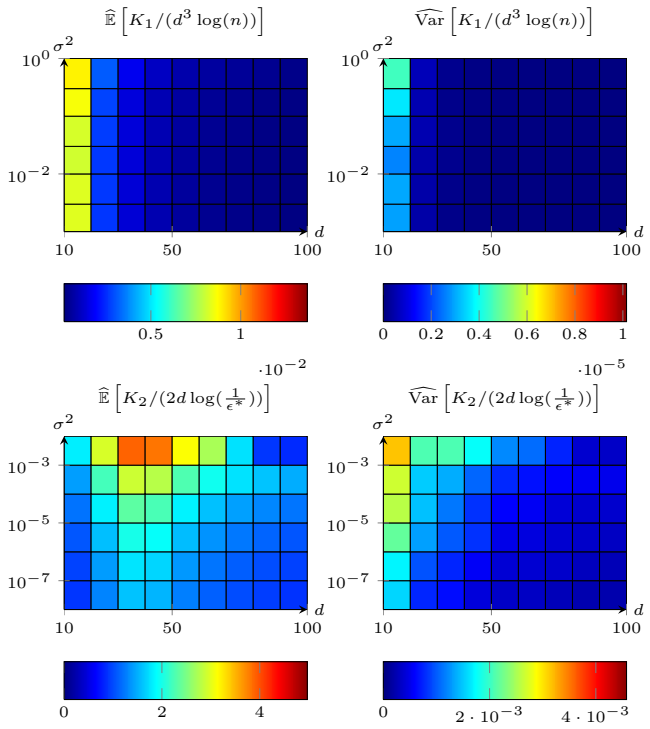

Figure 3: Illustration in the noisy case of the bounds on K_1, K_2 over d and σ with fixed n, averaged over 50 trials. In this simulation, $n = 5000$ with d ranging from 10 to 100 and σ^2 chosen from 1e-3 to 1 (above) and 1e-8 to $1e-2$ (below). We run GROUSE until it converges to $\min\left\{1/2, e^{-\tau_2 d^2/n}\right\}$ and record the corresponding K_1. For large SNR (small σ^2) we first run GROUSE until it converges to 1/2 and record K_1, then let GROUSE further converge to $\epsilon^* = \max\left\{\sigma^2, \tau_1\frac{d^2}{n}\sigma^2\right\}$ and record K that gives $K_2 = K - K_1$. For this plot we set both τ_1, τ_2 be $\log(d)$.

with probability $1-\rho'$. In fact, this phase usually takes many fewer iterations and reaches the local region in all empirical trials. In the local phase for the noiseless case, we provided a very tight bound for the required iterations $K_2 = O(d\log(1/\epsilon^*)$ to achieve a final desired accuracy of ϵ^*.

When the observations contain noise, we establish a rate of expected improvement of both of our metrics ζ_t and ϵ_t for all iterations t. Establishing the global convergence result remains as future work.

7 Acknowledgment

The work of both authors in this publication was supported by the U.S. Army Research Office under grant number W911NF1410634.

References

[Absil et al., 2009] Absil, P.-A., Mahony, R., and Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

[Armentano et al., 2014] Armentano, D., Beltrán, C., and Shub, M. (2014). Average polynomial time for eigenvector computations. *arXiv preprint arXiv:1410.2179*.

[Arora et al., 2013] Arora, R., Cotter, A., and Srebro, N. (2013). Stochastic optimization of PCA with capped MSG. In *Advances in Neural Information Processing Systems*, pages 1815–1823.

[Balsubramani et al., 2013] Balsubramani, A., Dasgupta, S., and Freund, Y. (2013). The fast convergence of incremental PCA. In *Advances in Neural Information Processing Systems*, pages 3174–3182.

[Balzano et al., 2010] Balzano, L., Nowak, R., and Recht, B. (2010). Online identification and tracking of subspaces from highly incomplete information. In *Communication, Control, and Computing*

(Allerton), 2010 48th Annual Allerton Conference on, pages 704–711. IEEE.

[Balzano and Wright, 2014] Balzano, L. and Wright, S. J. (2014). Local convergence of an algorithm for subspace identification from partial data. *Foundations of Computational Mathematics*, pages 1–36.

[Balzano, 2012] Balzano, L. K. (2012). *Handling missing data in high-dimensional subspace modeling*. PhD thesis, University of Wisconsin – Madison.

[Bhojanapalli et al., 2015] Bhojanapalli, S., Kyrillidis, A., and Sanghavi, S. (2015). Dropping convexity for faster semi-definite optimization. *arXiv preprint arXiv:1509.03917*.

[Brooks et al., 2013] Brooks, J. P., Dulá, J., and Boone, E. L. (2013). A pure l1-norm principal component analysis. *Computational statistics & data analysis*, 61:83–98.

[Candès et al., 2011] Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11.

[Chen and Wainwright, 2015] Chen, Y. and Wainwright, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*.

[d'Aspremont et al., 2008] d'Aspremont, A., Bach, F., and Ghaoui, L. E. (2008). Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294.

[De Sa et al., 2014] De Sa, C., Olukotun, K., and Ré, C. (2014). Global convergence of stochastic gradient descent for some nonconvex matrix problems. *arXiv preprint arXiv:1411.1134*.

[Edelman et al., 1998] Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353.

[Golub and Van Loan, 2012] Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*. JHU Press, 4 edition.

[He et al., 2012] He, J., Balzano, L., and Szlam, A. (2012). Incremental gradient on the grassmannian for online foreground and background separation in subsampled video. In *IEEE CVPR*.

[Jain et al., 2013] Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM Symposium on Theory of computing*, pages 665–674. ACM.

[Keshavan, 2012] Keshavan, R. H. (2012). *Efficient algorithms for collaborative filtering*. PhD thesis, Stanford University.

[Keshavan et al., 2010] Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998.

[Ngo and Saad, 2012] Ngo, T. and Saad, Y. (2012). Scaled gradients on grassmann manifolds for matrix completion. In *Advances in Neural Information Processing Systems*, pages 1412–1420.

[Nguyen et al., 2014] Nguyen, H. H., Vu, V., et al. (2014). Random matrices: Law of the determinant. *The Annals of Probability*, 42(1):146–167.

[Recht et al., 2010] Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.

[Richtárik and Takáč, 2014] Richtárik, P. and Takáč, M. (2014). Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38.

[Stewart and Sun, 1990] Stewart, G. W. and Sun, J.-g. (1990). *Matrix perturbation theory*. Academic press.

[Xu et al., 2010] Xu, H., Caramanis, C., and Sanghavi, S. (2010). Robust PCA via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504.

[Zheng and Lafferty, 2015] Zheng, Q. and Lafferty, J. (2015). A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117.