Convergence of a Grassmannian Gradient Descent Algorithm for Subspace Estimation From Undersampled Data *

Dejiao Zhang and Laura Balzano
dejiao & girasole@umich.edu
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor

Abstract

Subspace learning and matrix factorization techniques have many applications in science and engineering, and efficient algorithms are critical as dataset sizes continue to grow. Many relevant problem formulations are non-convex, and in a variety of contexts it has been observed that solving the non-convex problem directly is not only efficient but reliably accurate. We discuss convergence theory for a particular method: first order incremental gradient descent constrained to the Grassmannian. The output of the algorithm is an orthonormal basis for a d-dimensional subspace spanned by an input streaming data matrix. We study two sampling cases: where each data vector of the streaming matrix is fully sampled, or where it is undersampled by a sampling matrix $A_t \in \mathbb{R}^{m\times n}$ with $m \ll n$. Our results cover two cases, where A_t is Gaussian or a subset of rows of the identity matrix. We propose an adaptive stepsize scheme that depends only on the sampled data and algorithm outputs. We prove that with fully sampled data, the stepsize scheme maximizes the improvement of our convergence metric at each iteration, and this method converges from any random initialization to the true subspace, despite the non-convex formulation and orthogonality constraints. For the case of undersampled data, we establish monotonic expected improvement on the defined convergence metric for each iteration with high probability.

This technical report was updated in February 2022 to match Dejiao Zhang's PhD dissertation [33], which corrected some errors. For the case with full observations (no compressed or missing data), the theoretical results herein have been superseded by several other results in the literature, including results for the GROUSE algorithm itself [6].

1 Introduction

Low-rank matrix factorization is an essential tool for high-dimensional inference with fewer measurements than variables of interest, where low-dimensional models are necessary to perform accurate and stable inference. Many modern problems fit this paradigm, where signals are undersampled because of sensor failure, resource constraints, or privacy concerns. Suppose we wish to factorize a matrix $M = UW^T$ when we only get a small number of linear measurements of M. Solving for the subspace basis U can be computationally burdensome in this undersampled problem and related regularized problems. Many algorithms that attempt to speed up computation are solving a non-convex optimization problem, and therefore come with few guarantees.

The Singular Value Decomposition (SVD) provides the solution to the non-convex matrix factorization problem formulation with full data, and there are several highly successful algorithms for solving it [18]. Unfortunately, these algorithms cannot easily be extended to problems with incomplete observations of the matrix. 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 [2, 11, 14, 16, 20, 21, 35]. These new algorithms, being gradient-based, are well-suited to extensions of the SVD where the matrix is not fully sampled and where we include different cost functions or regularizers. For example, with gradient methods to solve the SVD we may be able to solve Robust PCA [13, 19, 31], Sparse PCA [15], or even ℓ_1 PCA [12] with gradient methods as well. However, almost none of these results gives guarantees in *streaming problem*, where data can only be accessed one partial column vector at a time. This is a critical problem in the modern machine learning context with massive data and comparatively limited memory, or in applications where data are collected continuously and must be processed in realtime. The existing

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

theoretical results for the streaming problem significantly overestimate the number of samples needed for convergence for typical algorithms.

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 $\mathbb{R}^n$, denoted by $\mathcal{G}(n,d)$. 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, for fully sampled data without noise, this gradient algorithm *converges globally to the global minimizer* almost surely, *i.e.,* it converges from any random initialization to the global minimizer. For undersampled data, including compressively sampled data and missing data, we provide results showing monotonic improvement in expectation on the metric of convergence for each iteration.

This paper is organized as follows. The problem formulation and the GROUSE algorithm are described in Section 2. The global convergence result for fully sampled data is presented in Section 4, the convergence behavior of GROUSE with undersampled data is studied in Section 5, and the corresponding proofs are provided in Sections A.1, A.2 and A.3. Experiment results are in Section 6.

2 Problem Setting

In this paper, we consider the problem of learning a low dimensional subspace representation from streaming data. Specifically, we are given a sequence of observations $x_t = A_t v_t$ where $A_t \in \mathbb{R}^{m\times n}$ $(m \leq n)$ are sampling matrices that are given for each observation; $v_t \in \mathbb{R}^n$ are drawn from a continuous distribution with support on the true subspace, spanned by $\bar{U} \in \mathbb{R}^{n\times d}$ with orthonormal columns, *i.e.,* $v_t = \bar{U} s_t, s_t \in \mathbb{R}^d$. In this paper, we study three different sampling frameworks: the fully sampled case with A_t being the identity matrix, the compressively sampled case with $A_t \in \mathbb{R}^{m\times n}$ $(m \ll n)$ being random Gaussian matrices, and the missing data case where each row of A_t $(m \ll n)$ is uniformly sampled from the identity matrix.

We formulate subspace estimation as a non-convex optimization problem as follows. Let $U \in \mathbb{R}^{n\times d}$ be a matrix with orthonormal columns. Then we want to solve:

$$\begin{aligned} \underset{U\in\mathbb{R}^{n\times d}}{\text{minimize}} \quad & \sum_{t=1}^{T} \min_{w_t} \|A_t U w_t - x_t\|_2^2 \\ \text{subject to} \quad & \text{span}\,(U) \in \mathcal{G}(n,d) \end{aligned} \tag{1}$$

This problem is non-convex firstly because of the product of the two variables U and w_t and secondly because the optimization is over the Grassmannian $\mathcal{G}(n,d)$, the non-convex set of all d-dimensional subspaces in $\mathbb{R}^n$. We study an online algorithm 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.

We can see the relationship between our problem and the well studied low-rank matrix recovery problem. Let $W \in \mathbb{R}^{d\times T}$ and $M = [v_1, \ldots, v_T] \in \mathbb{R}^{n\times T}$, then (1) is equivalent to

$$\begin{aligned} \underset{U\in\mathbb{R}^{n\times d}, W\in\mathbb{R}^{d\times T}}{\text{minimize}} \quad & \|\mathcal{A}(UW) - \mathcal{A}(M)\|_2^2 \\ \text{subject to} \quad & \text{span}\,(U) \in \mathcal{G}(n,d) \end{aligned} \tag{2}$$

where $\mathcal{A} : \mathbb{R}^{n\times T} \to \mathbb{R}^{mT}$ is a linear operator. Our algorithm can be thought of as an incremental algorithm to solve this problem as well. Fueled by the great deal of recent success of directly solving non-convex factorization problems (as we discuss in related work below), we study the natural incremental gradient descent algorithm [10] applied to (1) directly. Since the optimization variable in our problem is a subspace, we constrain the gradient descent to the Grassmannian $\mathcal{G}(n,d)$. The resulting algorithm is called GROUSE (Grassmannian Rank-One Update Subspace Estimation) algorithm and is described in Algorithm 1. This description differs from its initial introduction in [7] in that it extends the missing data case to a more general sampling framework.

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 $A_t : \mathbb{R}^n \to \mathbb{R}^m$ and observation $x_t = A_t v_t$;
  Define $w_t := \arg\min_a \|A_t U_t a - x_t\|^2$;
  Define $p_t := U_t w_t$ and $\widetilde{r}_t := x_t - A_t p_t$, $r_t := A_t^T \widetilde{r}_t$;
  Using step size

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

  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

$$\frac{y_t}{\|y_t\|_2} = \frac{p_t}{\|p_t\|_2}\cos(\theta_t) + \frac{r_t}{\|r_t\|_2}\sin(\theta_t)$$

  $t := t + 1$;
until termination

2.1 Algorithm

At each step, the GROUSE algorithm receives a vector $x_t = A_t v_t$, and tries to minimize the inconsistency between $R(U)$ and the true subspace $R(\bar{U})$ with respect to the information revealed in the sampled vector x_t, *i.e.,*

$$\mathcal{F}(U; t) = \min_a \|A_t U a - x_t\|^2 \tag{5}$$

In order to do so, GROUSE forms the gradient of $\mathcal{F}$ with respect to U evaluated at the current estimate U_t, and takes a step in the direction of the negative gradient restricted to the Grassmannian. The derivation of the incremental gradient descent update rule on the Grassmannian is found in [7, 5], and we summarize it here.

To compute the gradient of $\mathcal{F}$ on the Grassmannian, we first need to compute the derivative of $\mathcal{F}$ with respect to U and evaluate it at U_t. As we will prove later, under mild conditions, $A_t U_t$ has full column rank with high probability. Therefore, the derivative is

$$\frac{d\mathcal{F}}{dU} = -2A_t^T \widetilde{r}_t w_t^T \tag{6}$$

where $\widetilde{r} := x_t - A_t U_t w_t$ denotes the residual vector with respect to the sampled vector x_t, and w_t is the least-squares solution of (5). Using Equation (2.70) in [17], the gradient of $\mathcal{F}$ on the Grassmannian then follows as

$$\begin{aligned}\nabla\mathcal{F} = \left(I - U_t U_t^T\right)\frac{d\mathcal{F}}{dU} &= -2\left(I - U_t U_t^T\right)A_t^T \widetilde{r}_t w_t^T \\ &= -2A_t^T \widetilde{r}_t w_t^T \ .\end{aligned} \tag{7}$$

The final equality follows by $\widetilde{r}_t \perp A_t U_t$, which can be verified using the definitions of w_t and $\widetilde{r}_t$. According to Eq (2.65) in [17], a gradient step along the geodesic with tangent vector $-\nabla\mathcal{F}$ can be then formed as a function of the singular values and singular vectors of $\nabla\mathcal{F}$. For this specific case of our rank one $\nabla\mathcal{F}$ given in (7), the update rule follows as

$$U(\eta) = U_t + \left[\left(\cos(\eta_t\sigma_t) - 1\right)\frac{U_t w_t}{\|w_t\|} + \sin(\eta_t\sigma_t)\frac{A_t^T \widetilde{r}_t}{\|A_t^T \widetilde{r}_t\|}\right]\frac{w_t^T}{\|w_t\|} \tag{8}$$

where $\eta_t > 0$ is the chosen step size at iteration t, $p_t := U_t w_t$ is the predicted value of the projection of the vector v_t onto $R(U_t)$ and $\sigma_t = \|A_t^T \widetilde{r}_t\|\|p_t\|$. By leveraging the fact that $\widetilde{r}_t \perp A_t U_t$ and $p_t \in R(U_t)$, it's easy to verify that the rank-one update (8) maintains orthogonality $U(\eta)^T U(\eta) = \mathbb{I}_d$, and tilts $R(U_t)$ to a new point on Grassmannian.

In summary, for each observation the GROUSE algorithm works as follows: it projects the data vector onto the current estimate of the true subspace with respect to the sampling matrix A_t, to get either the exact (when $A_t = \mathbb{I}_n$) or approximated projection p_t and residual $r_t = A_t^T \widetilde{r}_t$. Then GROUSE updates the current estimate with a rank-one

step as described by (4). In the present work, we propose an adaptive stepsize framework that sets the stepsize only based on the sampled data and the algorithm outputs. More specifically, at each iteration a stepsize η_t is chosen such that $\eta_t \sigma_t = \arctan\left(\frac{\|r_t\|}{\|p_t\|}\right)$. As shown in Section 4, the proposed stepsize scheme is greedy for the fully sampled data, *i.e.,* it maximizes the improvement of our defined convergence metric at each iteration. For the undersampled data, we establish a local convergence result by showing that, with the proposed stepsize, GROUSE moves the current estimated subspace towards the true subspace with high probability despite the nonconvex nature of the problem and undersampled data.

2.2 Related Work

Many recent results have shown theoretical support for directly solving non-convex matrix factorization problems with gradient or alternating minimization methods. Among the incremental methods [16] is the one closest to ours, where the authors consider recovering a positive semidefinite matrix with undersampled data. They propose a step size scheme with which they prove global convergence results from a randomly generated initialization. However, their convergence results contain a obscure term, and their choice of step size depends on the knowledge of some parameters that are likely to be unknown in practical problems. Without this knowledge, the results only hold with sufficiently small step size that implies significantly slower convergence.

In contrast, while our work applies more narrowly to the subspace estimation problem, we provide an explicit expression for the expected improvement at each iteration, using a step size that only depends on the observations and outputs of the algorithms. Based on that, we prove that with fully sampled data, the proposed stepsize scheme maximizes the improvement of our convergence metric at each iteration, and GROUSE converges from any random initialization to the true subspace, despite the non-convex formulation and orthogonality constraint global convergence. We further posit a conjecture on the global convergence rate that better matches the practical observations for fully sampled data. Although we have not yet established a complete proof of this conjecture, we present our current approach in Appendix A.2.

Other work that has looked at incremental methods has focused only on fully sampled vectors. For example, [4] invokes a martingale-based argument to derive the global convergence rate of the proposed incremental PCA method to the single top eigenvector in the fully sampled case. In contrast, [3] 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.

The results in this paper are very closely related to our previous work [9]. In [9], we prove that, within a local region of the true subspace, an expected improvement of their defined convergence metric for each iteration of GROUSE can be obtained. In contrast, we establish global convergence results to a global minimizer from any random initialization for fully sampled data, and extend the local convergence results to compressively sampled data. We also expand the local convergence results in [9] to a much less conservative region, and we provide a much simpler analysis framework that can be applied to different sampling strategies. Moreover, for each iteration of the GROUSE algorithm, the expected improvement on the convergence metric defined in [9] only holds locally in both theory and practice, while our theoretical result provides a tighter bound for the global convergence behavior of GROUSE over a variety of simulations. This suggests that our result has more promise to be extended to a global result for both missing data and compressively sampled data.

Turning to batch methods, [27, 21] provided the first theoretical guarantee for an alternating minimization algorithm for low-rank matrix recovery in the undersampled case. Under typical assumptions required for the matrix recovery problems [26], they established geometric convergence to the global optimal solution. Earlier work [22, 24] 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. [14, 11, 35] 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 [1]. 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 (1) and provide global convergence rates to the global minimum.

Before we present the main results, we first call out the following notation which we use throughout this chapter. For notational convenience, we will drop the iteration subscript except our convergence metric ζ_t defined in Definition 1 hereafter.

Notation We use $R(M)$ to denote the column space of a matrix M and $\mathcal{P}_M$ to denote the orthogonal projection onto $R(M)$. $\mathbb{I}_n$ denotes the identity matrix in $\mathbb{R}^{n\times n}$ and M_i denotes the i^{th} row of matrix M. In this paper, without specification, $\|\cdot\|$ denotes the ℓ_2 norm. $R(\bar{U})$ and $R(U)$ denote the true subspace and our estimated subspace respectively, here both $\bar{U}$ and U are matrices in $\mathbb{R}^{n\times d}$ with orthonormal columns. Also we use $v_\parallel$ and $v_\perp$ to denote the projection and residual of the underlying full vector $v \in \mathbb{R}^n$ onto the estimated subspace $R(U)$, *i.e.,* $v_\parallel = UU^T v, v_\perp = v - v_\parallel$. Note that these two quantities are in general unknown for the undersampled data case. We define them so as to relate the intermediate quantities, determined by the algorithm and sampled data, to the improvement on our defined convergence metric.

3 Preliminaries

In this section, we first define our convergence metric and describe an assumption on the streaming data needed to establish our results. Subsequently, we state a fundamental result that is essential to quantify the improvement on the convergence metric over GROUSE iterates.

Definition 1 (Determinant similarity)**.** *Our measure of similarity between $R(U)$ and $R(\bar{U})$ is $\zeta \in [0,1]$, defined as*

$$\zeta := \det(\bar{U}^T U U^T \bar{U}) = \prod_{k=1}^{d} \cos^2 \phi_k \ .$$

where ϕ_k denotes the k^{th} principal angle between $R(\bar{U})$ and $R(U)$, where $0 \le \phi_1 \le \cdots \le \phi_d \le \pi/2$ are defined by $\cos\phi_k = \sigma_k(\bar{U}^T U)$ with σ_k denoting the k^{th} singular value of $\bar{U}^T U$ (See [18, Section 6.4.3]).

The convergence metric ζ increases to one when our estimate $R(U)$ converges to $R(\bar{U})$, *i.e.,* all principal angles between the two subspaces equal zero. Compared to other convergence metrics defined either as $\|(I - \bar{U}\bar{U}^T)U\|_F^2 = d - \|\bar{U}^T U\|_F^2 = \sum_{k=1}^{d} \sin^2\phi_k$ or $1 - \|\bar{U}^T U\|_2^2 = \sin^2\phi_1$, our convergence metric ζ measures the similarity instead of the discrepancy between $R(U)$ and $R(\bar{U})$. In other words, ζ achieves its maximum value one when $R(U)$ converges to $R(\bar{U})$, while the typical subspace distance is zero when the subspaces are equal. Also note that $\zeta = 0$ iff at least one of the principal angles is a right angle. That is, all stationary points U_{stat} of the full data problem except the true subspace have $\det\left(\bar{U}^T U_{stat} U_{stat}^T \bar{U}\right) = 0$ [32, 5].

Assumption 1. *For the underlying data $v = \bar{U}s$, we assume the entries of s are independent, and identically distributed symmetrically about zero, and each entry has zero-mean and unit variance.*

Given this assumption, we have the following lemma which relates the projection $v_\parallel$ and the projection residual $v_\perp$ to the improvement on our convergence metric ζ_t. As we will show in the following sections, this lemma is crucial for us to establish the expected improvement on our defined convergence metric ζ_t for all the sampling frameworks considered in this work. The proof is provided in Section A.1.

Lemma 1. *Let $v_\parallel$ and $v_\perp$ denote the projection and residual of the full data sample v onto the current estimate $R(U)$. Then given Assumption 1, for each iteration of GROUSE we have*

$$\mathbb{E}\left[\frac{\|v_\perp\|^2}{\|v_\parallel\|^2}\middle| U\right] \ge \mathbb{E}\left[\frac{\|v_\perp\|^2}{\|v\|^2}\middle| U\right] \ge \frac{1-\zeta_t}{d} \ . \tag{9}$$

Although both projection ($v_\parallel$) and projection residual ($v_\perp$) are in general unknown for the undersampled data, we can relate the approximated projection residual $A^T\widetilde{r}$ to the true one $v_\perp$ by leveraging either random matrix theory or the incoherence property of the underlying subspace $R(\bar{U})$. Therefore, the above lemma provides a unifying step to quantify the improvement on the convergence metric for all cases considered in the present work.

4 Fully Sampled Data

In this section, we consider fully sampled data, *i.e.,* $A = \mathbb{I}_n$. The corresponding proofs for these results can be found in Section A.2. We start by deriving a greedy step size scheme for each iteration t that maximizes the improvement on our convergence metric ζ_t. For each update we prove the following:

$$\frac{\zeta_{t+1}}{\zeta_t} = \left(\cos\theta + \frac{\|v_\perp\|}{\|v_\|\|}\sin\theta\right)^2 . \tag{10}$$

It then follows that

$$\theta^* = \arg\max_\theta \frac{\zeta_{t+1}}{\zeta_t} = \arctan\left(\frac{\|v_\perp\|}{\|v_\|\|}\right). \tag{11}$$

This is equivalent to (3) in the fully sampled setting $A_t = \mathbb{I}_n$. Using θ^*, we obtain monotonic improvement on the determinant similarity that can be quantified by the following lemma.

Lemma 2 (Monotonicity for the fully sampled noiseless case)**.** *For fully sampled data, choosing step size* $\theta^* = \arctan\left(\frac{\|v_\perp\|}{\|v_\|\|}\right)$*, after one iteration of GROUSE we obtain*

$$\frac{\zeta_{t+1}}{\zeta_t} = 1 + \frac{\|v_\perp\|^2}{\|v_\|\|^2} \geq 1 .$$

To gain more insight into the improvement on ζ_t for each iteration of GROUSE, we call out the following lemma, which is a natural result of Lemma 1 and Lemma 2.

Lemma 3 (Expected improvement on ζ_t)**.** *When fully sampled data satisfying Assumption 1 are input to the GROUSE (Algorithm 1), the expected improvement after one update step is given as:*

$$\mathbb{E}\left[\zeta_{t+1}\middle|U\right] \geq \left(1 + \frac{1-\zeta_t}{d}\right)\zeta_t .$$

Under the mild assumption that each data vector is randomly sampled from the underlying subspace, we obtain strict improvement on ζ_t for each iteration provided $\|v_\perp\| > 0$ and $\|v_\|\| > 0$. Therefore, Lemma 2 provides insight into how the GROUSE algorithm converges to the global minimum of a non-convex problem formulation: GROUSE is not attracted to stationary points that are not the global minimum. As we mentioned previously, 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 $\bar{U}$ [5]. Therefore, if the initial point U_0 has determinant similarity with $\bar{U}$ strictly greater than zero, then we are guaranteed to stay away from other stationary points, since GROUSE increases the determinant similarity monotonically, according to Lemma 2. This together with Lemma 3 yields the following convergence result of GROUSE.

Theorem 4 (Convergence of GROUSE)**.** *Initialize the starting point* U_0 *of GROUSE such that* $\zeta_0 > 0$*. Let* $1 \geq \zeta^* \geq \zeta_0$ *be the desired accuracy of our estimated subspace. Then for any* $\rho > 0$*, after*

$$K \geq \left(\frac{d}{\zeta_0} + 1\right)\log\left(\frac{1}{\rho(1-\zeta^*)}\right)$$

iterations of GROUSE Algorithm 1,

$$\mathbb{P}\left(\zeta_K \geq \zeta^*\right) \geq 1 - \rho .$$

.

Notice that if we initialize GROUSE with U_0 drawn uniformly from the Grassmannian, *e.g.,* as the orthonormal basis of a random matrix $V \in R^{n\times d}$ with entries being independent standard Gaussian variables, this guarantees $\zeta_0 > 0$ with probability one. Therefore, Theorem 4 provides a global convergence result of GROUSE despite the non-convexity of our objective. However, with this randomly initialized U_0, the value of the associated determinant similarity ζ_0 is $\mathcal{O}\left(\left(\frac{d}{n}\right)^d\right)$. Thereby, GROUSE requires $\mathcal{O}\left(d\left(\frac{n}{d}\right)^d\right)$ iterations to converge to the required precision, which is quite pessimistic compared to the actual number of iterations required by GROUSE in numerical simulations. To narrow this gap, we call out the following conjecture on the global convergence rate for GROUSE.

Conjecture 1 (Global Convergence of GROUSE)**.** *Let $1 \geq \zeta^* > 0$ be the desired accuracy of our estimated subspace. With the initialization (U_0) of GROUSE as the range of an $n \times d$ matrix with entries being i.i.d standard normal random variables, then for any $\rho > 0$, after*

$$\begin{aligned} K &\geq K_1 + K_2 \\ &= \left(\frac{2d^2}{\rho} + 1\right) \tau_0 \log(n) + 2d \log\left(\frac{1}{2\rho(1-\zeta^*)}\right) \end{aligned}$$

iterations of GROUSE Algorithm 1,

$$\mathbb{P}\left(\zeta_K \geq \zeta^*\right) \geq 1 - 2\rho \ ,$$

where $\tau_0 = 1 + \frac{\log \frac{(1-\rho/2)}{C} + d\log(e/d)}{d \log n}$ with C be a constant approximately equal to 1.

This conjecture matches what we see in experimental results. We present a related theorem with additional assumptions in Section A.2. We show that the iteration complexity can potentially be a combination of iterations required by two phases: $K_1 = \left(\frac{2d^2}{\rho} + 1\right) \tau_0 \log(n)$ is the number of iterations required by GROUSE to achieve $\zeta_t \geq 1/2$ from a random initialization U_0; and $K_2 = 2d \log\left(\frac{1}{2\rho(1-\zeta^*)}\right)$ is the number of additional iterations required by GROUSE to converge to the given accuracy ζ^* from $\zeta_{K_1} = 1/2$.

We want to comment that conjecture 1 requires fully observed noiseless data, which is not very practical in many cases. However, it would potentially be the first convergence guarantee for the Grassmannian gradient descent based method for subspace estimation with streaming data. It is a very important initial step for further studies on more general cases, including undersampled data and noisy data with outliers. In the following section, we will analyze the convergence behavior of GROUSE for undersampled data. We leave the corrupted data case as future work.

5 Undersampled Data

In this section, we consider undersampled data where each vector v is subsampled by a sampling matrix $A \in \mathbb{R}^{m \times n}$ with the number of measurements being much smaller than the ambient dimension ($m \ll n$). We study two typical cases, the compressively sampled data where A are random Gaussian matrices, and the missing data where each row of A is uniformly sampled from the identity matrix, $\mathbb{I}_n \in \mathbb{R}^{n \times n}$.

We first outline several elementary facts that can help us understand how the GROUSE algorithm navigates on the Grassmannian with undersampled data. The proofs can be found in Section A.3.

Suppose AU has full column rank, then the projection coefficients w are found by the squares solution of $w =: \arg\min_a \|AUa - x\|^2$, *i.e.,* $w = (U^T A^T AU)^{-1} U^T A^T x$. Note that $x = Av$, therefore we can further decompose the projection coefficients w as $w = w_\| + w_\perp$ where

$$w_\| = \left(U^T A^T AU\right)^{-1} U^T A^T A v_\| \ , \qquad w_\perp = \left(U^T A^T AU\right)^{-1} U^T A^T A v_\perp \ . \tag{12}$$

This decomposition explicitly shows the perturbation induced by the undersampling framework, *i.e.,* $Av_\perp$ is not perpendicular to AU in general, though $v_\perp$ is orthogonal to $R(U)$. Now we are going to use this perturbation to show how the approximated projection p and residual r deviate from the exact ones obtained by projecting the full data sample v onto the current estimate $R(U)$.

Lemma 5. *Given Eq* (12)*, let $p = p_\| + p_\perp$ with $p_\| = Uw_\|$ and $p_\perp = Uw_\perp$, then*

$$p_\| = v_\| \quad \textit{and} \quad r = A^T A v_\perp - A^T \mathcal{P}_{AU}(Av_\perp) \ . \tag{13}$$

Proof. Let $a = U^T v_\|$, then a is the unique solution to $Uw = v_\|$ given that U has full column rank. Since AU also has full column rank, $b = \left(U^T A^T AU\right)^{-1} U^T A^T A v_\|$ is also the unique solution to $AUw = Av_\|$. It then follows that $AUa = Av_\| = AUb$. Therefore, $a = b$. As for the second statement, it simply follows due to the fact that $Av_\| = AUw_\| \in R(AU)$. Hence $\widetilde{r} = \left(\mathbb{I}_m - \mathcal{P}_{AU}\right) Av = \left(\mathbb{I}_m - \mathcal{P}_{AU}\right) Av_\perp$, recall that $\mathcal{P}_{AU}$ denotes the orthogonal projection operator onto the column space of AU. This together with $r = A^T \widetilde{r}$ completes the proof. □

Below we lower bound the improvement on ζ_t as a function of the key quantities $r, \widetilde{r}$ and p. Compared to Lemma 2, Lemma 5 and Lemma 6 highlight the how the perturbations induced by the undersampling framework influence the improvement on ζ_t for each iteration. Being able to analyze and bound the quantities that include the perturbations is the key to establish the expected improvement on ζ_t for undersampled data.

Lemma 6. *Suppose AU has full column rank, then for each iteration of GROUSE we have*

$$\frac{\zeta_{t+1}}{\zeta_t} \geq 1 + \frac{2\left\|\widetilde{r}\right\|^2 - \|r\|^2}{\|p\|^2} + 2\frac{\Delta}{\|p\|^2} \tag{14}$$

where $\Delta = w_\perp^T \left(\bar{U}^T U\right)^{-1} \bar{U}^T r$ with $w_\perp = \left(U^T A^T A U\right)^{-1} U^T A^T A v_\perp$.

The above lemma highlights the main hurdle in establishing global convergence for undersampled data. As is indicated by (14), there is no guarantee on monotonicity of the improvement on ζ_t. Indeed, the uncertainty and perturbations introduced by the undersampling framework can even prevent us from establishing monotonically expected improvement on ζ_t. However, we are still able to bound the key quantities in Lemma 6 and provide more insights on the convergence behavior of GROUSE for both compressively sampled data and missing data.

5.1 Compressively Sampled Data

This section presents convergence results for compressively sampled data. We use an approach that merges linear algebra with random matrix theory to establish an expected rate of improvement on the determinant similarity ζ_t at each iteration. We show that, under mild conditions, the determinant similarity increases in expectation with a rate similar to that of the fully sampled case, roughly scaled by $\frac{m}{n}$. Detailed proofs for this section are provided in Section A.3.

Theorem 7. *Suppose each sampling matrix A has i.i.d Gaussian entries distributed as $\mathcal{N}(0, 1/n)$. Let $\delta > 0$ and let ϕ_d denote the largest principal angle between $R(U)$ and $R(\bar{U})$. Then with probability exceeding $1 - \exp\left(-\frac{d\delta^2}{8}\right) - \exp\left(-\frac{m\delta^2}{32} + d\log\left(\frac{24}{\delta}\right)\right) - (4d+2)\exp\left(-\frac{m\delta^2}{8}\right)$ we obtain*

$$\mathbb{E}_v\left[\zeta_{t+1}\middle|U\right] \geq \left(1 + \gamma_1\left(1 - \gamma_2\frac{d}{m}\right)\frac{m}{n}\frac{1-\zeta_t}{d}\right)\zeta_t\ ,$$

where $\gamma_1 = \frac{(1-\delta)\left(1-2\delta\sqrt{\frac{m}{n}}\right)}{\left(1+\sqrt{\frac{1+\delta}{1-\delta}}\frac{d}{m}\right)^2}$ and $\gamma_2 = \left(1 + \frac{2\tan(\phi_d)+\delta\frac{d}{\cos(\phi_d)}}{\left(1-2\delta\sqrt{\frac{m}{n}}\right)\sqrt{(1+\delta)d/m}}\right)\frac{1+\delta}{1-\delta}$. Now let $\beta = \frac{8(1+\delta)}{(1-\delta)^2(1-2\delta)^2}$, further suppose

$$m \geq d\cdot\max\left\{\frac{32}{\delta^2}\log\left(\frac{24n^{2/d}}{\delta}\right), \beta\left(\tan\phi_d + \delta\cos\phi_d d\right)\left(\tan\phi_d + \delta\cos\phi_d d + \frac{1}{2}\right)\right\}\ ,$$

then with probability at least $1 - 2/n^2 - \exp\left(-d\delta^2/8\right)$ we have

$$\mathbb{E}_v\left[\zeta_{t+1}\middle|U\right] \geq \left(1 + \frac{1}{2\gamma_1}\frac{m}{n}\frac{1-\zeta_t}{d}\right)\zeta_t\ .$$

This theorem implies that, for each iteration of GROUSE, expected improvement on ζ_t can be obtained with high probability as long as the number of samples is enough. As shown in Theorem 7, our theory for GROUSE requires more measurements when $R(U)$ is far away from $R(\bar{U})$, in which case $\cos\phi_d =: \varepsilon$ is very small. In the high dimensional setting where $m \ll n$, compared to the fully sampled data case, the expected improvement on ζ_t is approximately scaled down by $\frac{m}{n}$. As we will show, this scaling factor is mainly determined by the relative amount of effective information stored in the approximated projection residual. On the other hand, due to the perturbation and uncertainty induced by the compressed sampling framework, the improvement on the determinant similarity given by the lower bound in Lemma 6 is neither monotonic nor global. As mentioned before, this is the main hurdle to pass before we can provide a global convergence result for undersampled data. However, despite of these difficulties, we are still able to establish Theorem 7 which shows that, with reasonable number of measurements, the expected

improvement on the convergence metric is monotonic with high probability as long as our estimate $R(U)$ is not too far away from the true subspace $R(\bar{U})$.

To prove Theorem 7, we provide the following intermediate results to quantify the key quantities in Lemma 6 with high probability, where probability is taken with respect to the random Gaussian sampling matrix A.

Lemma 8. *Under the same conditions as Theorem 7, with probability at least* $1 - \exp\left(-\frac{m\delta_2^2}{2}\right) - \exp\left(-\frac{m\delta_1^2}{8}\right) - \exp\left(-\frac{d\delta_1^2}{8}\right)$ *we obtain*

$$\|\widetilde{r}\|_2^2 \geq (1-\delta_1)\left(1-\beta\frac{d}{m}\right)\frac{m}{n}\|v_\perp\|_2^2 \tag{15}$$

$$2\|\widetilde{r}\|_2^2 - \|r\|_2^2 \geq (1-\delta_1)\left(1-2\delta_2\sqrt{\frac{m}{n}}\right)\left(1-\beta\frac{d}{m}\right)\frac{m}{n}\|v_\perp\|_2^2 \tag{16}$$

where $\delta_1, \delta_2 \in (0,1)$*, and* $\beta = \frac{1+\delta_1}{1-\delta_1}$.

To interpret the above results, note that

$$\|\widetilde{r}\|_2^2 = \|(\mathbb{I}_m - \mathcal{P}_{AU})\,Av_\perp\|_2^2 = \|Av_\perp\|_2^2 - \|\mathcal{P}_{AU}(Av_\perp)\|_2^2\ . \tag{17}$$

where the first equality follows by the fact that $(\mathbb{I}_m - \mathcal{P}_{AU})\,Av_\parallel = 0$ as we argued before, and the second equality holds since $\mathcal{P}_{AU}$ is an orthogonal projection onto $R(AU)$. Then by leveraging the concentration property of random projection, we can prove that $\|\widetilde{r}\|_2^2$ concentrates around its expectation $\frac{m-d}{n}\|v_\perp\|_2^2$ with high probability. Also note that $\|r\|_2^2 \leq \|A\|_2^2\|\widetilde{r}\|_2^2$, hence the second statement (16) can be established by the concentration result of $\|\widetilde{r}\|_2^2$ and that of $\|A\|_2^2$ according to the random matrix theory.

Next we establish high probability bounds on $\|p\|_2^2$ and Δ. Then Theorem 7 follows naturally by first replacing the key quantities in Lemma 6 with their high probability bounds, and then taking the expectation over the uncertainty of the underlying full data v_t.

Lemma 9. *With the same conditions as Theorem 7, for any* $\delta_1 \in (0,1)$*, we have*

$$\|p\|^2 \leq \left(1+\sqrt{\frac{1+\delta_1}{1-\delta_1}\frac{d}{m}}\right)^2\|v\|^2$$

with probability at least $1 - \exp\left(-\frac{d\delta_1^2}{8}\right) - \exp\left(-\frac{m\delta_1^2}{32} + d\log\left(\frac{24}{\delta_1}\right)\right)$.

Lemma 10. *With the same conditions as Theorem 7, let* $\delta_1, \delta_3 \in (0,1)$*, then*

$$\Delta \leq \sqrt{\frac{1+\delta_1}{1-\delta_1}\frac{d}{m}}\left(\tan(\phi_d) + \delta_3\frac{d}{\cos(\phi_d)}\right)\frac{m}{n}\|v_\perp\|^2$$

holds with probability at least $1 - \exp\left(-\frac{d\delta_1^2}{8}\right) - \exp\left(-\frac{m\delta_1^2}{32} + d\log\left(\frac{24}{\delta_1}\right)\right) - 4d\exp\left(-\frac{m\delta_3^2}{8}\right)$.

Lemma 9 shows that $\|p\|_2^2$ doesn't diverge significantly from $\|v\|_2^2$ as long as $m \geq d$. This together with Lemma 6 and Lemma 8 imply that the required number of measurements in Theorem 7 is mainly determined by that required by Lemma 10 so as to prevent Δ diverging too far from $\frac{m}{n}\|v_\perp\|_2^2$. As a result, the improvement on the determinant similarity is still dominated by the magnitude of the projection residual over that of the projection, which is proportional to that of the full data case scaled by the sampling density. On the other hand, Lemma 10 implies that, in order to guarantee Δ to be much smaller than $\frac{m}{n}\|v_\perp\|_2^2$, the number of required measurements increases along with first principal angle between the estimated subspace $R(U)$ and the true subspace $R(\bar{U})$.

For the sake of completeness, we sketch the proof of Theorem 7 here, and the detailed proof is provided in Section A.3.

Proof sketch of Theorem 7. Let $\eta_1 = \frac{1+\delta}{1-\delta}\frac{d}{m}$, $\eta_2 = (1-\delta)\left(1-2\delta\sqrt{\frac{m}{n}}\right)$ and $\eta_3 = \tan(\phi_d) + \delta\frac{d}{\cos(\phi_d)}$, then plugging in the results in Lemmas 8, 9 and 10 into Lemma 6 with $\delta_1 = \delta_2 = \delta_3 = \delta$ yields,

$$\frac{\zeta_{t+1}}{\zeta_t} \geq 1 + \gamma_1\left(1-\gamma_2\frac{d}{m}\right)\frac{m}{n}\frac{\|v_\perp\|^2}{\|v\|^2} \geq 1 + \gamma_1\left(1-\gamma_2\frac{d}{m}\right)\frac{m}{n}\frac{1-\zeta_t}{d} \tag{18}$$

where $\gamma_1 = \frac{(1-\delta)\left(1-2\delta\sqrt{\frac{m}{n}}\right)}{\left(1+\sqrt{\frac{1+\delta}{1-\delta}\frac{d}{m}}\right)^2}$ and $\gamma_2 = \left(1+2\frac{\tan(\phi_d)+\delta_3\frac{d}{\cos(\phi_d)}}{\left(1-2\delta\sqrt{\frac{m}{n}}\right)\sqrt{(1-\delta^2)d/m}}\right)\frac{1+\delta}{1-\delta}$.
The first probability bound is obtained by taking the union bound of those quantities used to generate Lemma 8 to Lemma 10, which can be lower bounded by

$$1-\exp\left(-\frac{d\delta^2}{8}\right)-\exp\left(-\frac{m\delta^2}{32}+d\log\left(\frac{24}{\delta}\right)\right)-(4d+2)\exp\left(-\frac{m\delta^2}{8}\right) \tag{19}$$

Next we establish the complexity bound on m. As we will prove in Section A.3, $\gamma_2\frac{d}{m} < \frac{1}{2}$ is equivalent to the following,

$$m \geq \frac{8(1+\delta)}{(1-\delta)^2(1-2\delta)^2}\left(\varepsilon+\delta\sqrt{1+\varepsilon^2}d\right)\left(\varepsilon+\delta\sqrt{1+\varepsilon^2}d+\frac{1}{2}\right)d \tag{20}$$

To establish another bound on m, $m \geq \frac{32}{\delta^2}\log\left(\frac{24n^{2/d}}{\delta}\right)d$ implies the following,

$$\exp\left(-\frac{m\delta^2}{32}+d\log\left(\frac{24}{\delta}\right)\right) \leq \exp(-\log n^2) = \frac{1}{n^2} \tag{21}$$

$$(4d+2)\exp\left(-\frac{m\delta^2}{8}\right) \leq \frac{(4d+2)}{n^8}\left(\frac{\delta}{24}\right)^{4d} \ll \frac{1}{n^2} \tag{22}$$

(21) and (22) complete the proof for the bound on m and justify the simplification of the probability bound in (19). □

5.2 Missing Data

In this section, we study the convergence of GROUSE for the missing data case. We show that within the local region of the true subspace, we obtain an expected monotonic improvement on our defined convergence metric with high probability. We use Ω to denote the indices of observed entries for each data vector, and we assume Ω is uniformly sampled over $\{1,2,\ldots,n\}$ with replacement. In other words, we assume each row of the sampling matrices A is uniformly sampled from the rows of identity matrix $\mathbb{I}_n$ with replacement. We use the notation $Av =: v_\Omega, AU =: U_\Omega$. Again our results are with high probability with respect to A, in this case with respect to the random draw of rows of $\mathbb{I}_n$, and in expectation with respect to the random data v. Please refer to Section A.3 for the proofs of this section.

Before we present our main results, we first call out the typical incoherence assumption on the underlying data.

Definition 2. *A subspace $R(U)$ is incoherent with parameter μ if*

$$\max_{i\in\{1,\ldots,n\}}\|\mathcal{P}_U e_i\|_2^2 \leq \frac{\mu d}{n}$$

where e_i is the i^{th} canonical basis vector and $\mathcal{P}_U$ is the projection operator onto the column space of U.

Note that $1 \leq \mu \leq \frac{n}{d}$. According to the above definition, the incoherence parameter of a vector $z \in \mathbb{R}^n$ is defined as:

$$\mu(z) = \frac{n\|z\|_\infty^2}{\|z\|_2^2} \tag{23}$$

In this section, we assume the true subspace $R(\bar{U})$ is incoherent with parameter μ_0, and use $\mu(U)$, $\mu(v_\perp)$ to denote the incoherence parameter of $R(U)$ and $v_\perp$ respectively. We now show the expected improvement of ζ_t in a local region of the true subspace.

Theorem 11. *Suppose $\sum_{k=1}^d \sin^2\phi_k \leq \frac{d\mu_0}{16n}$ and $|\Omega| = m$. If*

$$m > \max\left\{\frac{128d\mu_0}{3}\log\left(\sqrt{2d}n\right), 64\mu(v_\perp)^2\log(n), 52\left(1+2\sqrt{\mu(v_\perp)\log(n)}\right)^2 d\mu_0\right\}$$

then with probability at least $1-\frac{3}{n^2}$ we have

$$\mathbb{E}_v\left[\zeta_{t+1}\middle|U\right] \geq \left(1+\frac{1}{4}\frac{m}{n}\frac{1-\zeta_t}{d}\right)\zeta_t\ .$$

This theorem shows that, within the local region of the true subspace, expected improvement on ζ_t can be obtained with high probability. As is implied by the theorem, this local region gets enlarged if the true subspace is more coherent, which may seem at first counterintuitive. However, the required number of measurements also increases as we increase μ_0. In the extreme case, when m increases to n, the local convergence results can be extended to a global result, as we proved for the full data case in Section 4. On the other hand, compared to Theorem 7, the convergence result for the missing data case holds within a more conservative local region of the true subspace. This gap is induced by the challenge of maintaining the incoherence property of our estimates $R(U)$, for which we had to consider the worst case. We leave the extension of the local convergence results to global results as future work.

In order to compare our result to the local convergence result in [Corollary 2.15, [9]], consider the following corollary.

Corollary 12. *Define the determinant discrepancy as $\kappa_t = 1 - \zeta_t$, then under the same conditions as Theorem 11, we have*

$$\mathbb{E}_v\left[\kappa_{t+1}\middle|\kappa_t\right] \leq \left(1 - \frac{1}{4}\left(1 - \frac{d\mu_0}{16n}\right)\frac{m}{nd}\right)\kappa_t$$

with probability exceeding $1 - 3/n^2$.

Recall that $1 \leq \mu_0 \leq \frac{n}{d}$, therefore the expected linear decay rate of κ_t is at least $1 - \frac{9}{16}\frac{m}{nd}$. In [9] (Corollary 2.15), a similar linear convergence result is established in terms of the Frobenius norm discrepancy between $R(\bar{U})$ and $R(U)$, denoted as $\epsilon_t = \sum_{i=1}^{d} \sin^2\phi_d$. However, their result only holds when $\epsilon_t \leq (8 \times 10^{-6})\frac{m}{n^3 d^2}$ which is more conservative than our assumption in Theorem 11. Moreover, as we mentioned previously, empirical evidence shows the lower bound in Theorem 11 holds for every iteration from any random initialization. In contrast, in [9], even for numerical results expected linear improvements only hold within the local region of the true subspace.

Now we present the following intermediate results for the proof of Theorem 11. Note that in this missing data case, the projection residual r_Ω of v_Ω onto U_Ω is mapped back to $\mathbb{R}^n$ by zero padding the entries at the indices that are not in Ω. Therefore, unlike Lemma 10 of the compressively sampled data case, here $\|\widetilde{r}\| = \|r\| = \|r_\Omega\|$. Therefore, (14) becomes

$$\frac{\zeta_{t+1}}{\zeta_t} \geq 1 + \frac{\|r_\Omega\|^2}{\|p\|^2} + 2\frac{\Delta}{\|p\|^2}\,. \tag{24}$$

Now similarly to the compressively sampled data case, we proceed by establishing concentration results for the key quantities $\|r\|_2^2$, $\|p\|_2^2$ and Δ respectively.

Lemma 13 ([8], Theorem 1)**.** *Let $\delta > 0$, and suppose $m \geq \frac{8}{3} d\mu(U)\log(2d/\delta)$. Then, with probability exceeding $1 - 3\delta$,*

$$\|r_\Omega\|^2 \geq (1 - \alpha_0)\frac{m}{n}\|v_\perp\|^2$$

where $\alpha_0 = \sqrt{\frac{2\mu(v_\perp)^2}{m}\log\left(\frac{1}{\delta}\right)} + \frac{(\beta_1+1)^2}{1-\gamma_1}\frac{d\mu(U)}{m}$, $\beta_1 = \sqrt{2\mu(v_\perp)\log\left(\frac{1}{\delta}\right)}$, and $\gamma_1 = \sqrt{\frac{8d\mu(U)}{3m}\log(2d/\delta)}$.

Lemma 14. *Let $\delta > 0$. Under the same condition on m as Lemma 13, with probability at least $1 - 2\delta$ we have*

$$\|p\|^2 \leq \left(1 + \frac{\beta_1 + 1}{1 - \gamma_1}\sqrt{\frac{d\mu(U)}{m}}\right)^2 \|v\|^2$$

where β_1 and γ_1 equal to those defined in Lemma 13.

Lemma 15. *Let $\delta > 0$. Under the same condition on m as Lemma 13, with probability at least $1 - 3\delta$ we have*

$$|\Delta| \leq \frac{\eta_3}{\cos\phi_d}\sqrt{\sin^2\phi_d + \frac{d\mu_0}{m}}\sqrt{\frac{d\mu(U)}{m}}\frac{m}{n}\|v_\perp\|^2$$

where $\eta_3 = \frac{(1+\beta_1)(1+\beta_2)}{1-\gamma_1}$, $\beta_2 = \sqrt{2\mu(v_\perp)\log\left(\frac{1}{\delta}\right)\frac{d\mu_0}{d\mu_0 + m\sin^2\phi_d}}$, and β_1 and γ_1 equal to those defined in Lemma 13.

Lemma 13 shows that the concentration of $\|r\|_2^2 = \|r_\Omega\|_2^2$ does not only depend on the sampling framework, but also on the incoherence property of the current estimate and the true projection residual, *i.e.,* $\mu(U)$ and $\mu(v_\perp)$. To see this clearly, recall that $\|r_\Omega\|_2^2 = \|v_{\perp,\Omega}\|_2^2 - \|\mathcal{P}_{U_\Omega}(v_{\perp,\Omega})\|_2^2$, hence the incoherence property of $v_\perp$ and $R(U)$ directly influences the concentration of $\|r_\Omega\|_2^2$. On the other hand, for compressive data, the Gaussian distributed sampling matrices yield tight concentration results for $\|p\|_2^2$, $\|r_\Omega\|_2^2$ and Δ. Therefore, the upper bounds of the key quantities established in Lemmas 13, 14 and 15 are not as tight as those for the compressive data except the extreme case where $\mu(U) = \mu(v_\perp) = 1$, *i.e.,* both $R(U)$ and $v_\perp$ are incoherent.

As shown in the above lemmas, in order to establish concentration of the key quantities in (24), it is essential for the subspaces generated by GROUSE to be incoherent over iterates. It has been proven in [9] that within the local region of $R(\bar{U})$, the incoherence of $R(U)$ can be bounded by that of $R(\bar{U})$.

Lemma 16 ([9], Lemma 2.5)**.** *Suppose* $\sum_{k=1}^{d} \sin^2\phi_k \leq \frac{d}{16n}\mu_0$*, then* $\mu(U) \leq 2\mu_0$.

Now we are ready to prove Theorem 11. We sketch the proof here, and a detailed proof is provided in Section A.3.

Proof sketch of Theorem 11. Given the condition required by Theorem 11, we have $\sin\phi_d \leq \sqrt{d\mu_0/16n}$ and $\cos\phi_d \geq \sqrt{1 - d\mu_0/16n}$. This together with Lemma 16 and Lemma 15 yield $|\Delta| \leq \frac{11}{5}\eta_3\frac{d\mu_0}{n}\|v_\perp\|^2$. Also for β_2 in Lemma 15, $\beta_2 \leq \sqrt{2\mu(v_\perp)\log(1/\delta)} = \beta_1$. Hence,

$$|\Delta| \leq \frac{11}{5}\frac{(1+\beta_1)^2}{1-\gamma_1}\frac{d\mu_0}{n}\|v_\perp\|^2 \,. \tag{25}$$

Letting $\eta_2 = \frac{(1+\beta_1)^2}{1-\gamma_1}\frac{d\mu_0}{m}$ and $\alpha_1 = \sqrt{\frac{2\mu(v_\perp)^2}{m}\log\left(\frac{1}{\delta}\right)}$, then applying this definition together with Lemma 16 to Lemma 14 and Lemma 13 yields

$$\|p\|^2 \leq \left(1 + \sqrt{\frac{2\eta_2}{1-\gamma_1}}\right)^2 \|v\|^2 \tag{26}$$

$$\|r_\Omega\|^2 \geq (1 - \alpha_1 - 2\eta_2)\frac{m}{n}\|v_\perp\|^2 \tag{27}$$

Now applying (25), (26) and (27) to (24) we have

$$\frac{\zeta_{t+1}}{\zeta_t} \geq 1 + \frac{(1 - \alpha_1 - \frac{32}{5}\eta_2)}{(1 + \sqrt{2\eta_2/(1-\gamma_1)})^2}\frac{m}{n}\frac{\|v_\perp\|^2}{\|v\|^2} \tag{28}$$

with probability at least $1 - 3\delta$. The probability bound is obtained by taking the union bound of those generating Lemmas 13, 14 and 15, as we can see in the proofs in Section A.3 this union bound is at least $1 - 3\delta$.

Letting $\eta_1 = \frac{(1-\alpha_1-\frac{32}{5}\eta_2)}{(1+\sqrt{2\eta_2/(1-\gamma_1)})^2}$, then $\eta_1 > 0$ is equivalent to $1 - \alpha_1 - \frac{32}{5}\eta_2 > 0$. This further gives that if m satisfies the condition in Theorem 11, then $\eta_1 > \frac{1}{4}$. Now taking expectation with respect to v yields,

$$\mathbb{E}_v\left[\zeta_{t+1}\middle|U\right] \geq \left(1 + \frac{1}{4}\frac{m}{n}\mathbb{E}\left[\frac{\|v_\perp\|^2}{\|v\|^2}\middle|U\right]\right)\zeta_t \geq \left(1 + \frac{1}{4}\frac{m}{n}\frac{1-\zeta_t}{d}\right)\zeta_t \tag{29}$$

where the last inequality follows from Lemma 1. Finally choosing δ to be $1/n^2$completes the proof. □

6 Numerical Results

In this section, we demonstrate that our theoretical results match the empirical convergence behavior of GROUSE. We generate the underlying data matrix $M = \begin{bmatrix} v_1 & v_2 & \dots & v_T \end{bmatrix}$ as $M = \bar{U}W$. For both the fully sampled data case and compressively sampled data case, the underlying signals are generated from a sparse subspace, demonstrating that incoherence assumptions are not required by our results for these two cases. Specifically, the underlying subspace of each trial is set to be a sparse subspace, as the range of an $n \times d$ matrix $\bar{U}$ with sparsity on the order of $\frac{\log(n)}{n}$. For the missing data case, we generate the underlying subspace as the range of an $n \times d$ matrix with i.i.d standard normal distribution. The entries of the coefficient matrix W for all three cases are generated as i.i.d $\mathcal{N}(0,1)$ satisfying

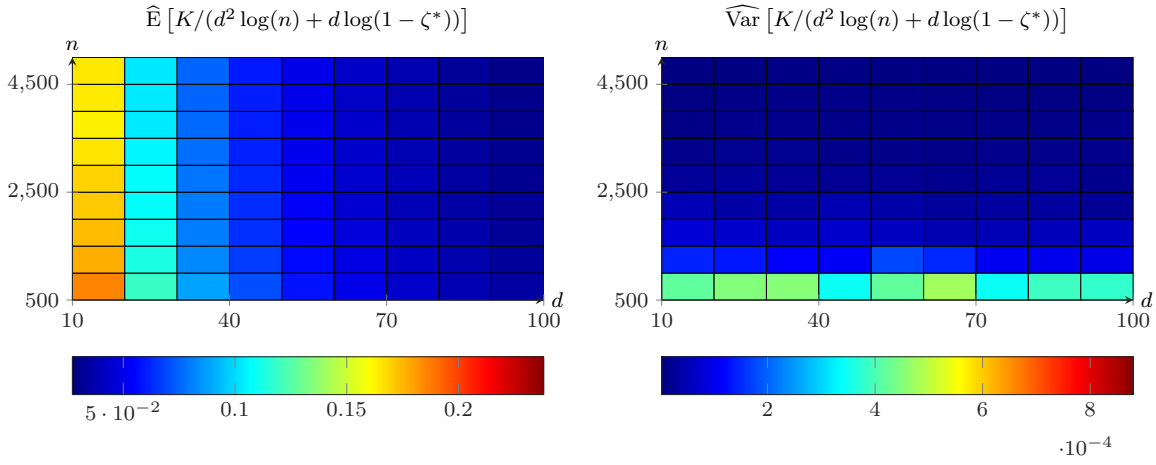

Figure 1: Illustration of the bounds on K in Conjecture 1 compared to their values in practice, averaged over 50 trials with different n and d. We show the ratio of K to the bound $d^2\log(n)+d\log(1-\zeta^*)$.

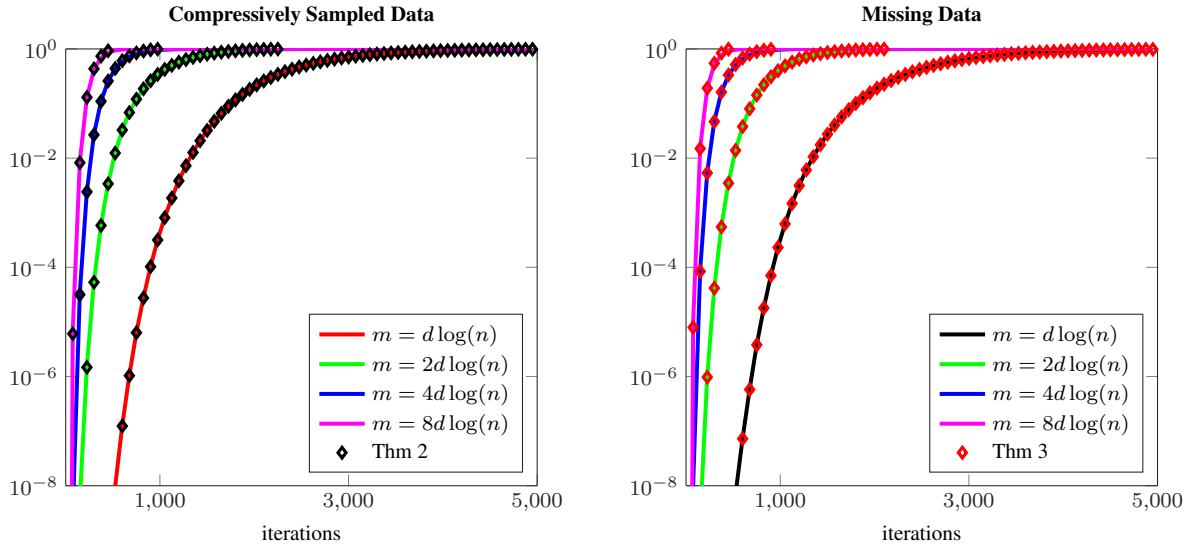

Figure 2: Illustration of expected improvement on ζ given by Theorem 7 (left) and Theorem 11 (right) over 50 trials. We set $n = 5000$, $d = 10$. The diamonds denote the lower bound on expected convergence rates described in Theorem 7 and Theorem 11.

Assumption 1. We also want to mention that we run GROUSE with random initialization for all of the plots in this section.

We first examine our global convergence result, *i.e.,* Theorem 4 and Conjecture 1, for the fully sampled data in Figure 1. We run GROUSE to convergence for a required accuracy $\zeta^* = 1 - 1e\text{-}4$ and show the ratio of K to the simplified bound of Conjecture 1, $d^2\log(n) + d\log\frac{1}{1-\zeta^*}$. We run GROUSE over 50 trials and show the mean and variance. We can see that, for fixed n, despite the conjecture's tighter convergence rate than the theorem's, it becomes 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 conjectured upper bounds and the actual iterations required by GROUSE is stable.

Next we examine our theoretical results (Theorem 7 and Theorem 11) for the expected improvement on ζ_t for the undersampled case in Figure 2. We set $n = 5000$ and $d = 10$. We run GROUSE over different sampling numbers m. The plots are obtained by averaging over 50 trials. We can see that our theoretical bounds on the expected improvement

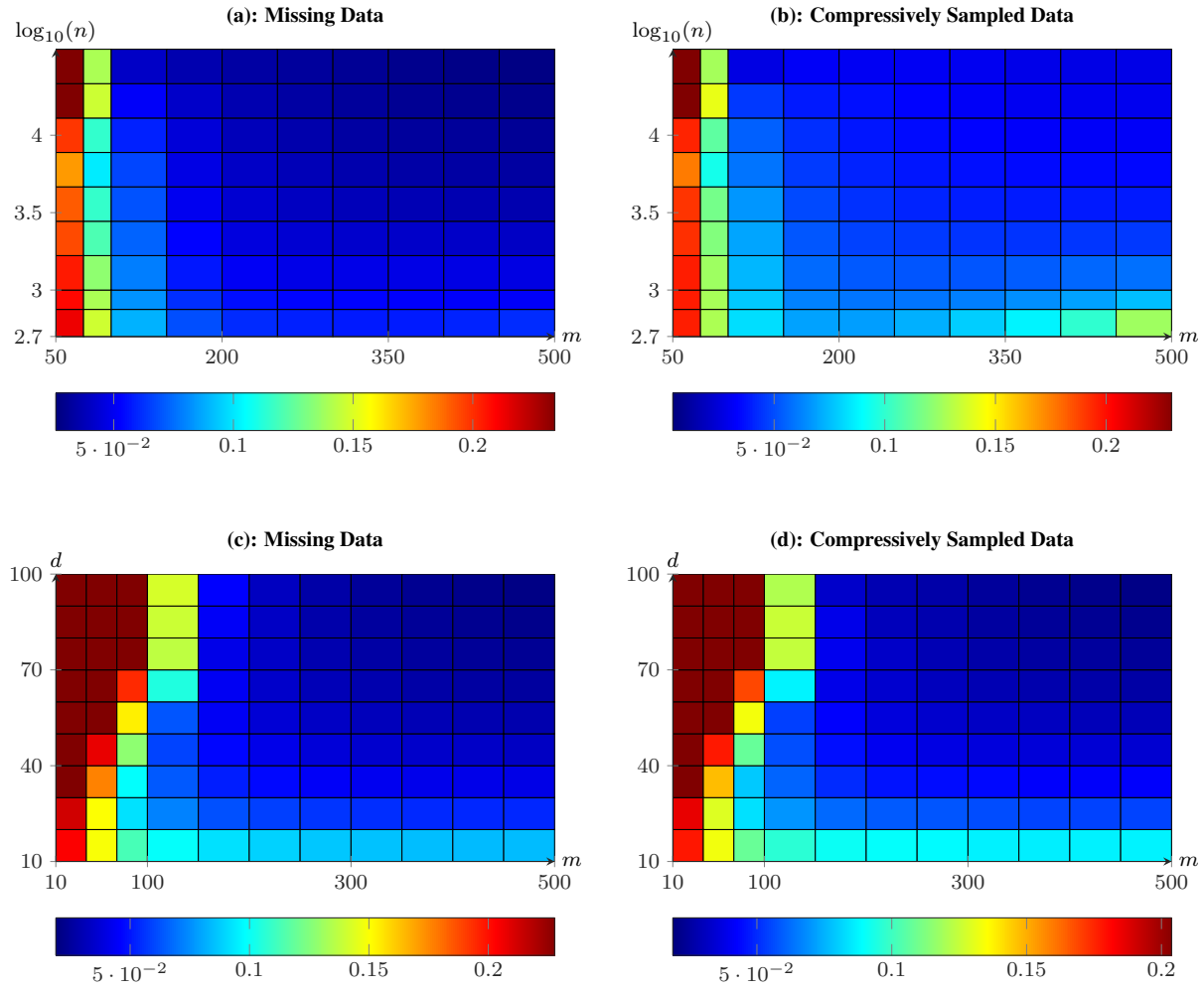

Figure 3: Illustration of our heuristic bounds on K (the actual iterations required by GROUSE to converge to the given accuracy) over different d, m and n, averaged over 20 trials. In this simulation, we run GROUSE from a random initialization to convergence for a required accuracy $\zeta^* = 1 - 1e\text{-}3$. We show the ratio of K to the heuristic bound $\frac{n}{m}\left(d^2 \log(n) + d \log(1-\zeta^*)\right)$. In (a) and (b), we set $d = 50$ and examine K over m and n for both missing data (a) and compressively sampled data (b). In (c) and (d), we set $n = 10000$ and examine K over m and d for both missing data (c) and compressively sampled data (d). In these plots, we use the dark red to indicate the failure of convergence.

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 both cases. Also note that Theorem 7 and Theorem 11 indicate that the expected improvement on the determinant similarity has a similar form to that of the fully sampled case roughly scaled by the sampling density (m/n). These together motivate us to approximate the required iterations to achieve a given accuracy as that required by the fully sampled case times the reciprocal of sampling density, n/m:

$$(n/m) \cdot \left(d^2 \log(n) + d \log(1-\zeta^*)\right) \ .$$

As we see in Figure 3, when m is slightly larger than d, the empirical mean of the ratio of the actual iterations required by GROUSE to our heuristic bound is similar to that of the full data case. We leave the rigorous proof of this heuristic as future work.

7 Conclusion

In this paper, we analyze a manifold incremental gradient descent algorithm applied to a particular non-convex optimization formulation for recovering a low-dimensional subspace from streaming data sampled from that subspace. We provide a simplified analysis as compared to [34], showing global convergence of the algorithm to the global minimizer for fully sampled data. However, the convergence rate we have established in theory is loose compared to what we observed in practice. A future direction is to narrow the gap between our theory and the actual performance of GROUSE, for which Conjecture 1 shows great promise.

With undersampled data, we show that expected improvement on our defined convergence metric can be obtained with high probability for each iteration. We prove that, comparing with fully sampled data, the expected improvement on determinant similarity is roughly proportional to the sampling density. With compressively sampled data this expected improvement holds from any random initialization, while it only holds within the local region of the true subspace for the missing data case. The limitation on the convergence of missing data arises due to the challenge of maintaining the incoherence property of our estimates in theory. Crossing this fundamental hurdle and extending the local convergence with missing data to a global result would be an interesting and valuable future direction.

A Supplementary material

A.1 Preliminaries

We start by providing the following lemma that we will use regularly in the manipulation of the matrix $\bar{U}^T U$. It also provides us with more insight into our metric of determinant similarity between the subspaces. The proof can be found in [28].

Lemma 17 ([28], Theorem 5.2)**.** *Let $U, \bar{U} \in \mathbb{R}^{n\times d}$ with orthonormal columns, then there are unitary matrices Q, $\bar{Y}$, and Y such that*

$$Q\bar{U}\bar{Y} := \begin{matrix} \\ d \\ d \\ n-2d \end{matrix} \begin{matrix} \\ \\ \\ \end{matrix} \begin{matrix} d \\ I \\ 0 \\ 0 \end{matrix} \quad \textit{and} \quad QUY := \begin{matrix} \\ d \\ d \\ n-2d \end{matrix} \begin{matrix} \\ \\ \\ \end{matrix} \begin{matrix} d \\ \Gamma \\ \Sigma \\ 0 \end{matrix}$$

where $\Gamma = \operatorname{diag}(\cos\phi_1, \ldots, \cos\phi_d), \Sigma = \operatorname{diag}(\sin\phi_1, \ldots, \sin\phi_d)$ with ϕ_i being the i^{th} principal angle between $R(U)$ and $R(\bar{U})$ defined in Definition 1.

Now we are going to prove Lemma 1, which is essential for us to establish expected improvement on the determinant similarity for each iteration in the various sampling cases we consider. Before that, we present the following lemmas that are requried for the proof.

Lemma 18. *Given any matrix $Q \in \mathbb{R}^{d\times d}$, suppose that $w \in \mathbb{R}^d$ is a random vector whose components w_i, $i = 1, \ldots, d$ are zero-mean, independent, and identically distributed symmetrically about zero (*i.e., *the distribution of w_i is an even function). Then*

$$E\left[\frac{w^T Q w}{w^T w}\right] = \frac{1}{d}\,\textbf{\textit{tr}}(Q)\ .$$

Proof of Lemma 18.

$$\begin{aligned} E\left[\frac{w^T Q w}{w^T w}\right] &= \sum_{i\neq j} E\left[\frac{w_i w_j Q_{ij}}{w^T w}\right] + \sum_{i=1}^{d} E\left[\frac{w_i^2 Q_{ii}}{w^T w}\right] \\ &= \sum_{i=1}^{d} Q_{ii} E\left[\frac{w_i^2}{w^T w}\right] \qquad (30) \\ &= \frac{1}{d}\,\mathbf{tr}\,Q\ , \qquad (31) \end{aligned}$$

where Eqs (30) and (31) hold by the following two arguments. For Eq (30), let $f(w_1, \ldots, w_d)$ be the joint distribution among the coordinates, and without loss of generality let $i = 1$ and $j \neq 1$, then

$$\begin{aligned}
&E\left[\frac{w_1 w_j Q_{1j}}{w^T w}\right] \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{w_1 w_j Q_{ij}}{w^T w} f(w_1, \ldots, w_d) dw_1 dw_2 \cdots dw_d \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{w_1 w_j Q_{1j}}{w_1^2 + \sum_{k \neq i} w_k^2} f(w_1) f(w_2) \cdots f(w_d) dw_1 dw_2 \cdots dw_d \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \frac{w_1}{w_1^2 + \sum_{k \neq i} w_k^2} f(w_1) dw_1 \right) w_j Q_{1j} f(w_2) \cdots f(w_d) dw_2 \cdots dw_d \\
&= 0
\end{aligned}$$

where the last inequality holds since $\frac{w_1}{w_1^2 + \sum_{k \neq i} w_k^2}$ is an odd function of w_1 and $f(w_1)$ is an even function of w_1, thereby the term in parentheses will integrate to zero. We note that if w_i is a discrete random variable, the argument would be similar.

To get Eq (31) we note that

$$1 = E\left[\frac{\sum_i w_i^2}{\sum_j w_j^2}\right] = \sum_i E\left[\frac{w_i^2}{w^T w}\right] = dE\left[\frac{w_i^2}{w^T w}\right] , i = 1, \ldots, d ,$$

where the last step holds because each w_i is identically distributed. □

Lemma 19 ([16], Lemma 16)**.** *Let $X = [X_1, \cdots, X_d]$ with $X_i \in [0, 1], i = 1, \ldots, d$, then*

$$d - \sum_{i=1}^{d} X_i \geq 1 - \Pi_{i=1}^{d} X_i$$

Proof of Lemma 1. According to Lemma 18 and Lemma 19 we have the following

$$\begin{aligned}
\mathbb{E}\left[\frac{\|v_\perp\|^2}{\|v\|^2}\middle| U\right] = \mathbb{E}\left[\frac{\|\bar{U}s\|^2 - \|UU^T\bar{U}s\|^2}{\|\bar{U}s\|^2}\middle| U\right] &\overset{\vartheta_1}{=} \mathbb{E}\left[\frac{s^T\bar{Y}(I - \Gamma^2)\bar{Y}^T s}{s^T s}\middle| U\right] \\
&\overset{\vartheta_2}{=} \frac{1}{d}\,\mathbf{tr}\left(I - \Gamma^2\right) \overset{\vartheta_3}{\geq} \frac{1 - \zeta_t}{d}
\end{aligned} \tag{32}$$

where ϑ_1 follows by Lemma 17 and $\|\bar{U}s\|^2 = \|s\|^2$, ϑ_2 from Lemma 18, and ϑ_3 from Lemma 19 with $X_i = \cos^2 \phi_i$. □

A.2 Proof of Fully Sampled Data

In this section we prove the results of Section 4. We start by proving Eq 10, the deterministic expression for the change in determinant similarity from one step of the GROUSE algorithm to the next. Using this expression, we prove the GROUSE monotonic improvement of Lemma 2, expected improvement of Lemma 3, and finally the global convergence conjecture 1.

Recall that $\frac{y}{\|y\|} = \cos(\theta)\frac{v_\|}{\|v_\|\|} + \sin(\theta)\frac{v_\perp}{\|v_\perp\|}$ in Algorithm 1. Then according to the GROUSE update in 4 we have

$$\begin{aligned}
\det\left(\bar{U}^T U_{t+1}\right) &= \det\left(\bar{U}^T U + \left(\frac{\bar{U}^T y}{\|y\|} - \frac{\bar{U}^T v_\|}{\|v_\|\|}\right)\frac{w^T}{\|w\|}\right) \\
&\overset{\vartheta_1}{=} \det\left(\bar{U}^T U\right)\left(1 + \frac{w^T(\bar{U}^T U)^{-1}}{\|w\|}\left(\frac{\bar{U}^T y}{\|y\|} - \frac{\bar{U}^T v_\|}{\|v_\|\|}\right)\right) \\
&\overset{\vartheta_2}{=} \det\left(\bar{U}^T U\right)\frac{w^T(\bar{U}^T U)^{-1}\bar{U}^T y}{\|y\|\|w\|} \\
&\overset{\vartheta_3}{=} \det\left(\bar{U}^T U\right)\left(\cos\theta + \frac{\|v_\perp\|}{\|v_\|\|}\sin\theta\right) \qquad (33)
\end{aligned}$$

where ϑ_1 follows from the Schur complement, *i.e.,* that for any invertible matrix M we have $\det\left(M + ab^T\right) = \det(M)\left(1 + b^T M^{-1} a\right)$; ϑ_2 and ϑ_3 hold since $\|v_\|\|^2 = \|Uw\|^2 = \|w\|^2$ and the following

$$w^T(\bar{U}^T U)^{-1}\bar{U}^T v_\| \overset{w=U^T\bar{U}s}{=} v^T v_\| = \|v_\|\|^2 \qquad (34a)$$

$$w^T(\bar{U}^T U)^{-1}\bar{U}^T v_\perp \overset{w=U^T\bar{U}s}{=} v^T v_\perp = \|v_\perp\|^2. \qquad (34b)$$

Given this, the proof of Lemma 2 follows directly from the above proof and the greedy step size derived in Eq. 11.

Proof of Lemma 2. By using $\theta = \arctan\left(\frac{\|v_\perp\|}{\|v_\|\|}\right)$, we have $\cos\theta = \frac{\|v_\|\|}{\|v\|}$ and $\sin\theta = \frac{\|v_\perp\|}{\|v\|}$. This together with 33 gives $\det\left(\bar{U}^T U_{t+1}\right) = \det\left(\bar{U}^T U\right)\frac{\|v\|}{\|v_\|\|}$. Therefore, $\frac{\zeta_{t+1}}{\zeta_t} = \frac{\det\left(\bar{U}^T U_{t+1}\right)^2}{\det\left(\bar{U}^T U\right)^2} = \frac{\|v\|^2}{\|v_\|\|^2} = 1 + \frac{\|v_\perp\|^2}{\|v_\|\|^2}$. □

Proof of Lemma 3. Lemma 3 follows directly from 1 and 2, *i.e.,*

$$\begin{aligned}
\mathbb{E}\left[\frac{\zeta_{t+1}}{\zeta_t}\middle| U\right] = 1 + \mathbb{E}\left[\frac{\|v_\perp\|^2}{\|v_\|\|^2}\middle| U\right] &\geq 1 + \mathbb{E}\left[\frac{\|v_\perp\|^2}{\|v\|^2}\middle| U\right] \\
&\geq 1 + \frac{1-\zeta_t}{d} \qquad (35)
\end{aligned}$$

Note that, given U, ζ_t is a constant, hence completes the proof. □

With the above results, we are ready to prove Theorem 4.

Proof of Theorem 4. Let $\kappa_t = 1 - \zeta_t$ denote the determinant *discrepancy* between $R(\bar{U})$ and $R(U)$. According to Lemma 3 we have the following:

$$\mathbb{E}\left[\frac{\kappa_{t+1}}{\kappa_t}\middle| U\right] \leq 1 - \frac{1-\kappa_t}{d} \qquad (36)$$

Now according to Lemma 2, $\kappa_t \leq 1 - \zeta_0$ for all $t \geq 0$. So using Eq (36) we have the following:

$$\mathbb{E}\left[\kappa_{t+1}\middle| U\right] \leq \left(1 - \frac{1-\kappa_t}{d}\right)\kappa_t \leq \left(1 - \frac{\zeta_0}{d}\right)\kappa_t \ .$$

Taking expectation of both sides, we have

$$\mathbb{E}\left[\kappa_{t+1}\right] \leq \left(1 - \frac{\zeta_0}{d}\right)\mathbb{E}\left[\kappa_t\right] \ .$$

After $K \geq \frac{d}{\zeta_0}\log\frac{1}{\rho(1-\zeta^*)} \geq \frac{d}{\zeta_0}\log\frac{\mathbb{E}[\eta_{K_1}]}{\rho(1-\zeta^*)}$ iterations of GROUSE we obtain

$$\mathbb{E}\left[\kappa_{t+K_1}\right] \leq \left(1 - \frac{\zeta_0}{d}\right)^K \mathbb{E}[\kappa_0] \leq \left(1 - \frac{\zeta_0}{d}\right)^{\frac{d}{\zeta_0}\log\frac{\mathbb{E}[\kappa_0]}{\rho(1-\zeta^*)}} \mathbb{E}[\kappa_0] \leq \rho(1-\zeta^*) \ .$$

Therefore

$$\mathbb{P}\left(\zeta_K \geq \zeta^*\right) = 1 - \mathbb{P}\left(\kappa_K \geq 1 - \zeta^*\right) \geq 1 - \frac{\mathbb{E}\left[\kappa_K\right]}{1-\zeta^*} \geq 1 - \rho\ . \tag{37}$$

□

To get full convergence results, we need the following lemma, which gives us guarantees for a random initial point.

Lemma 20. *[25] Initialize the starting point U_0 of GROUSE as the orthonormalization of an $n \times d$ matrix with entries being standard normal random variables. Then*

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

where $C > 0$ is a constant.

Now we will show a result that gives evidence for Conjecture 1.

Theorem 21 (Global Convergence of GROUSE: Evidence for Conjecture 1)**.** *Let $1 \geq \zeta^* > 0$ be the desired accuracy of our estimated subspace. Let ρ be any number within the range $(0, 1]$. Let $\bar{\zeta}_t$ be a non-decreasing sequence with $\mathbb{E}[\bar{\zeta}_0] = \mathbb{E}[\zeta_0]$ such that*

$$\mathbb{E}\left[\bar{\zeta}_{t+1} \middle| U\right] \geq \left(1 + \frac{\rho}{2d}\right)\bar{\zeta}_t\ .$$

Assume the ζ_t produced by GROUSE converges faster than $\bar{\zeta}_t$, i.e.,

$$\mathbb{E}\left[\zeta_{K_1}\right] \geq \mathbb{E}\left[\bar{\zeta}_{K_1}\right] \geq 1 - \frac{\rho}{2} \tag{38}$$

Suppose the initialization for GROUSE (U_0) is the range of an $n \times d$ matrix with entries being i.i.d standard normal random variables. Then after

$$\begin{aligned} K &\geq K_1 + K_2 \\ &= \left(\frac{2d^2}{\rho} + 1\right)\tau_0 \log(n) + 2d \log\left(\frac{1}{2\rho(1-\zeta^*)}\right) \end{aligned}$$

iterations of GROUSE Algorithm 1,

$$\mathbb{P}\left(\zeta_K \geq \zeta^*\right) \geq 1 - 2\rho\ ,$$

where $\tau_0 = 1 + \frac{\log \frac{(1-\rho/2)}{C} + d\log(e/d)}{d \log n}$ with C a constant approximately equal to 1.

Proof of Theorem 21. Let $\kappa_t = 1 - \zeta_t$ denote the determinant *discrepancy* between $R(\bar{U})$ and $R(U)$. According to Lemma 3 we have the following:

$$\mathbb{E}\left[\frac{\zeta_{t+1}}{\zeta_t}\middle| U\right] \geq 1 + \frac{1-\zeta_t}{d} \tag{39a}$$

$$\mathbb{E}\left[\frac{\kappa_{t+1}}{\kappa_t}\middle| U\right] \leq 1 - \frac{1-\kappa_t}{d} \tag{39b}$$

Therefore, the expected convergence rate of ζ_t is faster when $R(U)$ is far away from $R(\bar{U})$, while that of κ_t is faster when $R(U)$ is close to $R(\bar{U})$. This motivates us to split the analysis into two phases, bounding the number of iterations in each phase. We first use Eq (39a) to get the necessary K_1 iterations for GROUSE to converge to a local region of global optimal point from a random initialization. From there, we obtain the necessary K_2 iterations for GROUSE to converge to the required accuracy by leveraging Eq (39b).

As in the assumptions, let ρ be any number within the range $(0, 1]$. Let $\bar{\zeta}_t$ be a non-decreasing sequence with $\mathbb{E}[\bar{\zeta}_0] = \mathbb{E}[\zeta_0]$ and the expected increase rate being lower bounded as

$$\mathbb{E}\left[\bar{\zeta}_{t+1}\middle| U\right] \geq \left(1 + \frac{\rho}{2d}\right)\bar{\zeta}_t\ .$$

Taking expectation of both sides, we obtain the following:

$$\mathbb{E}\left[\bar{\zeta}_{t+1}\right] \geq \left(1+\frac{\rho}{2d}\right)\mathbb{E}[\bar{\zeta}_t]$$

Therefore after $K_1 \geq (2d/\rho+1)\log\frac{1-\frac{\rho}{2}}{\mathbb{E}[\zeta_0]}$ steps we have

$$\begin{aligned}\mathbb{E}\left[\bar{\zeta}_{K_1}\right] \geq \left(1+\frac{\rho}{2d}\right)^{K_1}\mathbb{E}[\zeta_0] &\geq \left(\left(1+\frac{\rho}{2d}\right)^{\frac{2d}{\rho}+1}\right)^{\log\frac{1-\frac{\rho}{2}}{\mathbb{E}[\zeta_0]}}\mathbb{E}[\zeta_0] \\ &\geq \mathbb{E}[\zeta_0]e^{\log\frac{1-\frac{\rho}{2}}{\mathbb{E}[\zeta_0]}} = 1-\frac{\rho}{2} \end{aligned} \tag{40}$$

Now we apply the assumption in (38), that the ζ_t produced by GROUSE converges faster than $\bar{\zeta}_t$. Therefore,

$$\mathbb{P}\left(\zeta_{K_1} \geq \frac{1}{2}\right) = 1-\mathbb{P}\left(1-\zeta_{K_1} \geq \frac{1}{2}\right) \overset{\vartheta_1}{\geq} 1-\frac{\mathbb{E}[1-\zeta_{K_1}]}{1/2} \geq 1-\rho \tag{41}$$

where ϑ_1 follows by applying Markov inequality to the nonnegative random variable $1-\bar{\zeta}_{K_1}$.

Now with probability at least $1-\rho$, $\zeta_t \geq \frac{1}{2}$ for all $t \geq K_1$, *i.e.,* $\kappa_t \leq \frac{1}{2}$ for all $t \geq K_1$. So using Eq (39b) we have the following:

$$\mathbb{E}\left[\kappa_{t+1}\middle|U\right] \leq \left(1-\frac{1-\kappa_t}{d}\right)\kappa_t \leq \left(1-\frac{1}{2d}\right)\kappa_t \ .$$

Taking expectation of both sides, we have

$$\mathbb{E}\left[\kappa_{t+1}\right] \leq \left(1-\frac{1}{2d}\right)\mathbb{E}\left[\kappa_t\right] \ .$$

After $K_2 \geq 2d\log\frac{1/2}{\rho(1-\zeta^*)} \geq 2d\log\frac{\mathbb{E}[\eta_{K_1}]}{\rho(1-\zeta^*)}$ additional iterations of GROUSE we obtain

$$\mathbb{E}\left[\kappa_{t+K_1}\right] \leq \left(1-\frac{1}{2d}\right)^{K_2}\mathbb{E}[\kappa_{K_1}] \leq \left(1-\frac{1}{2d}\right)^{2d\log\frac{\mathbb{E}[\kappa_{K_1}]}{\rho(1-\zeta^*)}}\mathbb{E}[\kappa_{K_1}] \leq \rho(1-\zeta^*) \ .$$

Hence following a similar argument as before we have

$$\mathbb{P}\left(\zeta_{K_1+K_2} \geq \zeta^*\right) = 1-\mathbb{P}\left(\kappa_{K_1+K_2} \geq 1-\zeta^*\right) \geq 1-\frac{\mathbb{E}\left[\kappa_{K_1+K_2}\right]}{1-\zeta^*} \geq 1-\rho \ . \tag{42}$$

(41) and (42) together complete the proof. □

Although we still need more rigorous analysis to justify our assumption, this proof provides the form of the convergence rate we can expect. We also want to emphasize that the above proof provides the local convergence rate for GROUSE. Specifically, as is indicated by the proof of the second phase, GROUSE requires at most $2d\log\frac{1/2}{\rho(1-\zeta^*)}$ iterations to converge from $\zeta_t = 1/2$ to any required accuracy $\zeta^* \in (1/2, 1)$.

A.3 Proof of Undersampled Data

In this section, we prove our main results for undersampled data. We again start by proving a result for the deterministic expression for the change in determinant similarity from one step of the GROUSE algorithm to the next, in this case a lower bound given by Lemma 6.

Proof of Lemma 6. Note that,

$$w^T(\bar{U}^T U)^{-1}\bar{U}^T p = w^T(\bar{U}^T U)^{-1}\bar{U}^T U w = \|p\|^2 \tag{43a}$$

$$w_1^T(\bar{U}^T U)^{-1}\bar{U}^T r \overset{\vartheta_1}{=} s^T\bar{U}^T U(\bar{U}^T U)^{-1}\bar{U}^T r = v^T A^T\widetilde{r} \overset{\vartheta_2}{=} \|\widetilde{r}\|^2 \tag{43b}$$

where ϑ_1 follows by Lemma 5 and ϑ_2 holds since $v^T A^T \widetilde{r} = v^T A^T \left(\mathbb{I}_m - \mathcal{P}_{AU}\right) \widetilde{r} = \|\widetilde{r}\|^2$. As a consequence, we have the following

$$\begin{aligned}
\det\left(\bar{U}^T U_{t+1}\right) &= \det\left(\bar{U}^T U + \bar{U}^T \left(\frac{p+r}{\|p+r\|} - \frac{p}{\|p\|}\right) \frac{w^T}{\|w\|}\right) \\
&\overset{\vartheta_3}{=} \det(\bar{U}^T U) \frac{w^T (\bar{U}^T U)^{-1} \bar{U}^T (p+r)}{\|p\| \sqrt{\|p\|^2 + \|r\|^2}} \\
&= \det(\bar{U}^T U) \frac{\|p\|^2 + \|r\|^2 + \|\widetilde{r}\|^2 - \|r\|^2 + \Delta}{\|p\| \sqrt{\|p\|^2 + \|r\|^2}}
\end{aligned}$$

where $\Delta = w_2^T \left(\bar{U}^T U\right)^{-1} \bar{U}^T r$; and ϑ_3 follows by the Schur complement $\det\left(M + ab^T\right) = \det(M)\left(1 + b^T M^{-1} a\right)$ for any invertible $M \in \mathbb{R}^{n \times n}$ and $a, b \in \mathbb{R}^n$. Hence

$$\frac{\bar{\zeta}_{t+1}}{\zeta_t} = \left(\frac{\det\left(\bar{U}^T U_{t+1}\right)}{\det\left(\bar{U}^T U\right)}\right)^2 \overset{\vartheta_4}{\geq} 1 + \frac{\|r\|^2}{\|p\|^2} + 2\frac{\|\widetilde{r}\| - \|r\|^2}{\|p\|^2} + 2\frac{\Delta}{\|p\|^2}$$

where ϑ_4 holds since $(c+d)^2 \geq c^2 + 2cd$ with $c = \frac{\|p\|^2 + \|r\|^2}{\|p\| \sqrt{\|p\|^2 + \|r\|^2}}$, $d = \frac{\|\widetilde{r}\|^2 - \|r\|^2 + \Delta}{\|p\| \sqrt{\|p\|^2 + \|r\|^2}}$. □

In the following sections, we proceed by establishing the convergence results of missing data and compressively sampled data by bounding the key quantities in Lemma 6.

Proof for Compressively Sampled Data We start by showing how the results on the key quantities in Lemmas 8, 9 and 10 lead to the main result of the compressively sampled data case.

Proof of Theorem 7. Let $\eta_1 = \frac{1+\delta}{1-\delta}\frac{d}{m}$, $\eta_2 = (1-\delta)\left(1 - 2\delta\sqrt{\frac{m}{n}}\right)$ and $\eta_3 = \tan(\phi_d) + \delta\frac{d}{\cos(\phi_d)}$, then plugging in the results in Lemma 8 to Lemma 10 into Lemma 6 with $\delta_1 = \delta_2 = \delta_3 = \delta$ yields,

$$\begin{aligned}
\frac{\zeta_{t+1}}{\zeta_t} &\geq 1 + \frac{2\|\widetilde{r}\|^2 - \|r\|^2}{\|p\|^2} + 2\frac{\Delta}{\|p\|^2} \\
&\geq 1 + \frac{1}{\left(1 + \sqrt{\eta_1}\right)^2} \left(\eta_2(1 - \eta_1) - 2\sqrt{\eta_1}\eta_3\right) \frac{m}{n} \frac{\|v_\perp\|^2}{\|v\|^2} \\
&= 1 + \gamma_1 \left(1 - \gamma_2 \frac{d}{m}\right) \frac{m}{n} \frac{\|v_\perp\|^2}{\|v\|^2} \\
&= \left(1 + \gamma_1 \left(1 - \gamma_2 \frac{d}{m}\right) \frac{m}{n}\right) \frac{1 - \zeta_t}{d}
\end{aligned} \tag{44}$$

where $\gamma_2 = \left(1 + 2\frac{\eta_3}{\eta_2\sqrt{\eta_1}}\right)\frac{1+\delta}{1-\delta} = \left(1 + 2\frac{\tan(\phi_d) + \delta_3\frac{d}{\cos(\phi_d)}}{\left(1 - 2\delta\sqrt{\frac{m}{n}}\right)\sqrt{(1-\delta^2)d/m}}\right)\frac{1+\delta}{1-\delta}$, $\gamma_1 = \frac{\eta_2}{\left(1+\sqrt{\eta_1}\right)^2} = \frac{(1-\delta)\left(1-2\delta\sqrt{\frac{m}{n}}\right)}{\left(1+\sqrt{\frac{1+\delta}{1-\delta}\frac{d}{m}}\right)^2}$, and the last equality follows from Lemma 1.

The probability bound is obtained by taking the union bound of those quantities (in Lemma 22, Lemma 25, Lemma 24, Corollary 27, Lemma 35) used to generate Lemma 8 to Lemma 10. As we can see, this union bound is

$$\begin{aligned}
&1 - \exp\left(-\frac{m\delta^2}{2}\right) - \exp\left(-\frac{d\delta^2}{8}\right) - \exp\left(-\frac{m\delta^2}{32} + d\log\left(\frac{24}{\delta}\right)\right) - (4d+1)\exp\left(-\frac{m\delta^2}{8}\right) \\
&> 1 - \exp\left(-\frac{d\delta^2}{8}\right) - \exp\left(-\frac{m\delta^2}{32} + d\log\left(\frac{24}{\delta}\right)\right) - (4d+2)\exp\left(-\frac{m\delta^2}{8}\right)
\end{aligned} \tag{45}$$

To get the complexity bound on m, let $\varepsilon = \tan(\phi_d)$, $\alpha_1 = \varepsilon + \delta\sqrt{1+\varepsilon^2}d$, $\alpha_2 = \frac{1+\delta}{1-\delta}$ and $\alpha_3 = \left(1 - 2\delta\sqrt{\frac{m}{n}}\right)\sqrt{1+\delta}$,

then according to 54 we have $\gamma_2 \frac{d}{m} < \frac{1}{2}$ is equivalent to the following,

$$\begin{aligned}
&\alpha_2 d + \frac{2\alpha_1\alpha_2\sqrt{d}}{\alpha_3}\sqrt{m} < \frac{m}{2} \\
&\Leftrightarrow \left(\sqrt{\frac{m}{2}} - \frac{\alpha_1\alpha_2\sqrt{d}}{\alpha_3}\right)^2 > \left(\alpha_2 + \frac{\alpha_1^2\alpha_2^2}{\alpha_3^2}\right) d \\
&\overset{\vartheta_1}{\Leftarrow} m \geq 8\frac{\alpha_1^2\alpha_2^2}{\alpha_3^2} d + 4\sqrt{\alpha_2}\frac{\alpha_1\alpha_2}{\alpha_3} d \\
&\overset{\vartheta_2}{\Leftarrow} m \geq \beta\left(\varepsilon + \delta\sqrt{1+\varepsilon^2 d}\right)\left(\varepsilon + \delta\sqrt{1+\varepsilon^2} d + \frac{1}{2}\right) d
\end{aligned} \tag{46}$$

where ϑ_1 follows from $\sqrt{\left(\alpha_2 + \frac{\alpha_1^2\alpha_2^2}{\alpha_3^2}\right) d} < \sqrt{\alpha_2 d} + \frac{\alpha_1\alpha_2}{\alpha_3}\sqrt{d}$; and ϑ_2 follows by $\beta = \frac{8(1+\delta)}{(1-\delta)^2(1-2\delta)^2}$.

To establish another bound on m we can see that $m \geq \frac{32}{\delta^2}\log\left(\frac{24n^{2/d}}{\delta}\right) d$ implies the following,

$$\exp\left(-\frac{m\delta^2}{32} + d\log\left(\frac{24}{\delta}\right)\right) \leq \exp(-\log n^2) = \frac{1}{n^2} \tag{47}$$

$$(4d+2)\exp\left(-\frac{m\delta^2}{8}\right) \leq \frac{(4d+2)}{n^8}\left(\frac{\delta}{24}\right)^{4d} \to 0 \tag{48}$$

Eqs (47) and (48) complete the proof for the bound on m and justify the simplification of the probability bound in Eq (45). □

Next we are going to prove the intermediate lemmas in Section 5.1, *i.e.,* bound the key quantities in Lemma 6, for which we need the following concentration results.

Lemma 22. *Let $A \in \mathbb{R}^{m\times n}$ with entries being i.i.d Gaussian random variables distributed as $\mathcal{N}(0, 1/n)$, $v \in \mathbb{R}^n$ is an vector. Then for any $\delta \in (0,1)$, with probability at least $1 - 2\exp^{-m\delta^2/8}$, we have*

$$\mathbb{P}\left(\|Av\|_2^2 > (1+\delta)\frac{m}{n}\|v\|_2^2\right) < \exp\left(-\frac{m\delta^2}{8}\right),$$
$$\mathbb{P}\left(\|Av\|_2^2 < (1-\delta)\frac{m}{n}\|v\|_2^2\right) < \exp\left(-\frac{m\delta^2}{8}\right).$$

Proof. Note that Av is a random vector with i.i.d entries distributing as $\mathcal{N}\left(0, \|v\|_2^2/n\right)$. Therefore, $\frac{n\|Av\|_2^2}{\|v\|_2^2}$ is a chi-squared distribution with m degrees of freedom, which yields,

$$\mathbb{P}\left[\frac{n\|Av\|_2^2}{m\|v\|_2^2} - 1 > \delta\right] < \exp\left(-m\delta^2/8\right)$$
$$\mathbb{P}\left[\frac{n\|Av\|_2^2}{m\|v\|_2^2} - 1 < -\delta\right] < \exp\left(-m\delta^2/8\right)$$

□

Lemma 23. *Let $A \in \mathbb{R}^{m\times n}$ be a random matrix whose entries are independent and identically distributed Gaussian random variables with mean zero, and variance γ . Let $z_1, z_2 \in \mathbb{R}^n$ such that $z_1 \perp z_2$, then Az_1 and Az_2 are independent of each other.*

Proof. Let a_i^T denote the i^{th} row of A and $M = Az_1z_2^TA^T$. Then we have

$$\mathbb{E}[M]_{ii} = \mathbb{E}\left[a_i^T z_1 z_1^T a_i\right] = z_1^T\mathbb{E}[a_i a_i^T]z_2 = \gamma z_1^T z_2 = 0$$
$$\mathbb{E}[M]_{ij} = \mathbb{E}\left[a_i^T z_1 z_1^T a_j\right] = z_1^T\mathbb{E}[a_i a_j^T]z_2 = 0$$

Therefore Az_1 and Az_2 are uncorrelated. This together with the fact that both Az_1 and Az_2 are Gaussian distributed random vectors imply that Az_1 and Az_2 are independent. □

Lemma 24 ([30], Corollary 5.35)**.** *Let A be an $n \times m$ matrix ($n \geq m$) whose entries are independent standard normal random variables. Then for every $h \geq 0$, with probability at least $1 - 2\exp\left(-h^2/2\right)$ one has*

$$\sqrt{n} - \sqrt{m} - h \leq \sigma_{min}(A) \leq \sigma_{max}(A) \leq \sqrt{n} + \sqrt{m} + h \tag{49}$$

where $\sigma_{min}, \sigma_{max}$ denote the smallest and largest singular values of A.

With the above results, we are able to call out the following intermediate result to quantify $\|\mathcal{P}_{AU}(Av_\perp)\|_2^2$, which is a key quantity that will be used for proving Lemmas 8, 9 and 10.

Lemma 25. *Let $A \in \mathbb{R}^{m\times n}$ with entries being i.i.d Gaussian random variables distributed as $\mathcal{N}(0, 1/n)$, then for any $\delta \in (0, 1)$ we have*

$$\|\mathcal{P}_{AU} A v_\perp\|_2^2 \leq (1+\delta)\frac{d}{n}\|v_\perp\|_2^2$$

hold with probability at least $1 - \exp\left(-\frac{d\delta^2}{8}\right)$.

Proof. Note that $Av_\perp$ is a Gaussian random vector with i.i.d entries distributed as $\mathcal{N}\left(0, \|v_\perp\|_2^2/n\right)$, and AU is a Gaussian random matrix with i.i.d entries distributed as $\mathcal{N}(0, 1/n)$. Then according to Lemma 23, AU and $Av_\perp$ are independent of each other. Therefore, $y = \mathcal{P}_{AU}(Av_\perp)$ is the projection of $Av_\perp$ onto a independent random d-dimensional subspace. According to the rotation invariance property of $Av_\perp$, $\|\mathcal{P}_{AU}(Av_\perp)\|$ is equivalent to the length of projecting $Av_\perp$ onto its first d coordinates. Hence,

$$\mathbb{P}\left(\|\mathcal{P}_{AU}(Av_\perp)\|_2^2 = \sum_{k=1}^{d} y_k^2 \leq (1+\delta)\frac{d}{n}\|v_\perp\|_2^2\right) \geq 1 - \exp\left(-\frac{d\delta^2}{8}\right) \tag{50}$$

Similar to the proof for Lemma 22, here the probability bound is followed from the concentration bound for Chi-squared distribution with degree d. □

Now we start by proving that Lemma 8 follows directly from Lemma 22 and Lemma 24.

Proof of Lemma 8. According to Lemmas 22 and 25, we have

$$\begin{aligned}
\|\widetilde{r}\|_2^2 = \|\left(\mathbb{I}_m - \mathcal{P}_{AU}\right) A v_\perp\|_2^2 &= \|Av_\perp\|_2^2 - \|\mathcal{P}_{AU}(Av_\perp)\|_2^2 \\
&\geq (1-\delta_1)\frac{m}{n}\|v_\perp\|_2^2 - (1+\delta_1)\frac{d}{n}\|v_\perp\|_2^2 \\
&= (1-\delta_1)\left(1 - \frac{1+\delta_1}{1-\delta_1}\frac{d}{m}\right)\frac{m}{n}\|v_\perp\|_2^2
\end{aligned} \tag{51}$$

hold with probability at least $1 - \exp\left(-\frac{m\delta_1^2}{8}\right) - \exp\left(-\frac{d\delta_1^2}{8}\right)$. As for the second part of Lemma 8, we have

$$\begin{aligned}
2\|\widetilde{r}\|_2^2 - \|r\|_2^2 = 2\|\widetilde{r}\|_2^2 - \|A^T\widetilde{r}\|_2^2 &\geq (2 - \sigma_{\max}^2(A^T))\|\widetilde{r}\|_2^2 \\
&\overset{\vartheta_1}{\geq} \left(1 - 2\delta_2\sqrt{\frac{m}{n}}\right)\|\widetilde{r}\|_2^2 \\
&\geq \left(1 - 2\delta_2\sqrt{\frac{m}{n}}\right)(1-\delta_1)\left(1 - \frac{1+\delta_1}{1-\delta_1}\frac{d}{m}\right)\frac{m}{n}\|v_\perp\|_2^2
\end{aligned} \tag{52}$$

here ϑ_1 follows from Lemma 24 with $A_{ij} \sim \mathcal{N}(0, 1/n)$ and $h = \delta\sqrt{m/n}$. The probability bound $1 - \exp\left(-\frac{m\delta_1^2}{8}\right) - \exp\left(-\frac{d\delta_1^2}{8}\right) - \exp\left(-\frac{m\delta_2^2}{2}\right)$ is obtained by taking the union bound over 51 and ϑ_1. □

To prove Lemma 9 and Lemma 10, we need the following extra results which are implied by Lemma 22. The corresponding proofs are provided at the end of this section.

Corollary 26. *Under the conditions of Lemma 22, for $x, y \in \mathbb{R}^n$ and δ, with probability exceeding $1 - 4e^{-m\delta^2/8}$ we have*

$$\frac{m}{n}\left(x^T y - \delta\|x\|\|y\|\right) \leq x^T A^T A y \leq \frac{m}{n}\left(x^T y + \delta\|x\|\|y\|\right)$$

Corollary 27. *Under the condition of Lemma 22, for any vector $v \in R(U)$ we have*

$$\mathbb{P}\left(\|Av\|_2^2 > (1+\delta)\frac{m}{n}\|v\|_2^2\right) < \exp\left(-\frac{m\delta^2}{32} - d\log(\delta) + d\log(24)\right),$$
$$\mathbb{P}\left(\|Av\|_2^2 < (1-\delta)\frac{m}{n}\|v\|_2^2\right) < \exp\left(-\frac{m\delta^2}{32} - d\log(\delta) + d\log(24)\right).$$

Given Lemma 26 and Corollary 27, we prove Lemma 9 and Lemma 10 by first proving the following intermediate results to bound the key components of p and Δ.

Lemma 28. *Let $w_2 = \left(U^T A^T A U\right)^{-1} U^T A^T A v_\perp$, then*

$$\mathbb{P}\left(\|w_2\| \leq \sqrt{\frac{1+\delta_1}{1-\delta_2}\frac{d}{m}}\|v_\perp\|\right)$$
$$\geq 1 - \exp\left(-\frac{d\delta_1^2}{8}\right) - \exp\left(-\frac{m\delta_2^2}{8} - d\log(\delta_2) + d\log(24)\right)$$

Proof. Given the fact that $U \in \mathbb{R}^{n\times d}$ with columns being orthonormal, we have $\|w_2\| = \|Uw_2\|$. It then follows that,

$$\|Uw_2\| \overset{\vartheta_1}{\leq} \frac{\|AUw_2\|}{\sqrt{(1-\delta_2)m/n}} \overset{\vartheta_2}{\leq} \sqrt{\frac{1+\delta_1}{1-\delta_2}\frac{d}{m}}\|v_\perp\|$$

where ϑ_1 follows from Corollary 27, and ϑ_2 followed by Lemma 25, *i.e.,*

$$\|AUw_2\| = \|\mathcal{P}_{AU}(Av_\perp)\| \leq \sqrt{(1+\delta_1)\frac{d}{n}\|v_\perp\|^2}$$

The probability bound is obtained by applying the union bound over ϑ_1 and ϑ_2. □

Lemma 29. *Let ϕ_d denote the largest principal angle between $R(U)$ and $R(\bar{U})$, then*

$$\mathbb{P}\left(\left\|\bar{U}^T A^T A v_\perp\right\| \leq (\sin\phi_d + d\delta)\frac{m}{n}\|v_\perp\|\right) \geq 1 - 4d\exp\left(-\frac{m\delta^2}{8}\right)$$

Proof of Lemma 29. Let $\bar{u}_k$ denote the k^{th} column of $\bar{U}$, and $\delta \in (0,1)$. Then

$$\begin{aligned}
\left\|\bar{U}^T A^T A v_\perp\right\| &= \left\|\bar{U}^T\left(A^T A - \frac{m}{n}\mathbb{I}_n\right)v_\perp + \frac{m}{n}\bar{U}^T v_\perp\right\| \\
&\leq \frac{m}{n}\left\|\bar{U}^T v_\perp\right\| + \left\|\bar{U}^T\left(A^T A - \frac{m}{n}\mathbb{I}_n\right)v_\perp\right\| \\
&= \frac{m}{n}\left\|\bar{U}^T v_\perp\right\| + \sqrt{\sum_{k=1}^{d}\left(\bar{u}_k^T A^T A v_\perp - \frac{m}{n}\bar{u}_k^T v_\perp\right)^2} \\
&\overset{\vartheta_1}{\leq} \frac{m}{n}\left\|\bar{U}^T v_\perp\right\| + \sqrt{\sum_{k=1}^{d}\left(\delta\frac{m}{n}\|\bar{u}_k\|\|v_\perp\|\right)^2} \\
&\overset{\vartheta_2}{\leq} \sin\phi_d\frac{m}{n}\|v_\perp\| + \frac{m}{n}d\delta\|v_\perp\| \qquad (53)
\end{aligned}$$

where ϑ_1 follows from Lemma 26; ϑ_2 holds from Lemma 35 and the fact that $\sqrt{\sum_{k=1}^{d}\left(\delta\frac{m}{n}\|\bar{u}_k\|\|v_\perp\|\right)^2} \leq d\delta\frac{m}{n}\|\bar{u}_k\|\|v_\perp\|$; and the probability bound is obtained by taking the union bound of that in Lemma 26. □

We are ready to prove Lemma 9 and Lemma 10.

Proof of Lemma 9. Let $\eta = \sqrt{\frac{1+\delta_1}{1-\delta_1}\frac{d}{m}}$, then according to Lemma 28 we have

$$\begin{aligned}\|p\|^2 = \|Uw_1 + Uw_2\|^2 &\leq \left(\|v_\parallel\| + \|Uw_2\|\right)^2 \\ &\leq \left(\|v_\parallel\| + \eta\|v_\perp\|\right)^2 \\ &\leq (1+\eta)^2\|v\|^2\end{aligned}$$

with probability at least

$$1 - \exp\left(-\frac{m\delta_1^2}{32} - d\log(\delta_1) + d\log(24)\right) - \exp\left(-\frac{d\delta_1^2}{8}\right) .$$

Here the probability bound is obtained by choosing $\delta_1 = \delta_2$ in Lemma 28, hence completes the proof. □

Proof of Lemma 10. According to the definition of Δ, we can see Lemma 10 is a direct results of Lemma 28 and Lemma 35, that is

$$\begin{aligned}|\Delta| &= w_2^T\left(\bar{U}^TU\right)^{-1}\bar{U}^TA^T\left(\mathbb{I}_m - \mathcal{P}_{AU}\right)Av_\perp \\ &\leq \left\|w_2^T\right\|\left\|\left(\bar{U}^TU\right)^{-1}\right\|\left\|\bar{U}^TA^T\left(\mathbb{I}_m - \mathcal{P}_{AU}\right)Av_\perp\right\| \\ &\overset{\vartheta_1}{\leq} \|w_2\|\left\|\left(\bar{U}^TU\right)^{-1}\right\|\left\|\bar{U}^TA^TAv_\perp\right\| \\ &\overset{\vartheta_2}{\leq} \frac{1}{\cos(\phi_d)}\sqrt{\frac{1+\delta_1}{1-\delta_1}\frac{d}{m}}\|v_\perp\|\left(\sin\phi_d\frac{m}{n} + \frac{m}{n}d\delta_3\right)\|v_\perp\| \\ &= \frac{1}{\cos(\phi_d)}\sqrt{\frac{1+\delta_1}{1-\delta_1}\frac{d}{m}}\left(\sin(\phi_d) + d\delta_3\right)\frac{m}{n}\|v_\perp\|^2 \end{aligned} \tag{54}$$

where ϑ_1 holds since $\left\|\bar{U}^TA^T\left(\mathbb{I}_m - \mathcal{P}_{AU}\right)Av_\perp\right\| \leq \left\|\bar{U}^TA^TAv_\perp\right\|$; ϑ_2 followed by Lemma 28 and Lemma 29; and the probability bound is obtained by taking the union bound that in Lemma 28 and Lemma 29. □

Finally, we are going to prove the auxiliary results Corollary 27 and Lemma 26. The key idea for proving Corollary 27 is using the covering numbers argument and applying Lemma 8 to a given d-dimensional subspace $R(U)$. This is a common strategy used for compress sensing.

Proof of Corollary 27. Without loss of generality we restrict $\|v\| = 1$. From covering numbers [29], there exists a finite set Q with at most $\left(\frac{24}{\delta}\right)^d$ points such that $Q \subset \mathbb{R}(U)$, $\|q\| = 1, \forall q \in Q$, and for all $x \in R(U)$ with $\|v\| = 1$ we can find a $q \in Q$ such that

$$\|v - q\| \leq \delta/8$$

Now applying Lemma 22 to the points in Q with $\varepsilon = \delta/2$ and using the standard union bound, then with probability at least $1 - 2\left(\frac{24}{\delta}\right)^d\exp\left(-\frac{\delta^2}{32}m\right)$ we have

$$(1-\delta/2)\frac{m}{n}\|v\|^2 \leq \|Ax\|^2 \leq (1+\delta/2)\frac{m}{n}\|v\|^2$$

which gives

$$\sqrt{1-\delta/2}\sqrt{\frac{m}{n}}\|v\| \leq \|Ax\| \leq \sqrt{1+\delta/2}\sqrt{\frac{m}{n}}\|v\| \tag{55}$$

Since $\|v\| = 1$, we define γ as the smallest number such that

$$\|Ax\| \leq \sqrt{1+\gamma}\sqrt{\frac{m}{n}} \quad \forall x \in R(U) \tag{56}$$

Since for any $x \in R(U)$ with $\|v\| = 1$ we can find a $q \in Q$ such that $\|x - q\| \leq \delta/8$, we have the following

$$\|Ax\| \leq \|Aq\| + \|A(x-q)\| \leq \sqrt{1+\delta/2}\sqrt{\frac{m}{n}} + \sqrt{1+H}\sqrt{\frac{m}{n}}\delta/8$$

Since γ is the smallest number (56) holds, we have $\sqrt{1+\gamma} \leq \sqrt{1+\delta/2} + \sqrt{1+\gamma}\delta/8$.

$$\sqrt{1+\gamma} \leq \frac{\sqrt{1+\delta/2}}{1-\delta/8} \leq \sqrt{1+\delta} \tag{57}$$

Similarly, the lower bound follows by

$$\begin{aligned}
\|Ax\| \geq \|Aq\| - \|A(x-q)\| &\geq \sqrt{1-\delta/2}\sqrt{\frac{m}{n}} - \sqrt{1+\gamma}\frac{\delta}{8}\sqrt{\frac{m}{n}} \\
&\geq \left(\sqrt{1-\delta/2} - \sqrt{1+\delta}\frac{\delta}{8}\right)\sqrt{\frac{m}{n}} \\
&\geq \sqrt{1-\delta}\sqrt{\frac{m}{n}}
\end{aligned}$$

This completes the proof. □

Proof of Lemma 26. Note that,

$$\frac{x^T A^T A y}{\|x\|\|y\|} = \frac{1}{4}\left(\left\|A\left(\frac{x}{\|x\|} + \frac{y}{\|y\|}\right)\right\|^2 - \left\|A\left(\frac{x}{\|x\|} - \frac{y}{\|y\|}\right)\right\|^2\right)$$

Applying Lemma 22 on both terms separately and applying the union bound we have

$$\begin{aligned}
&\mathbb{P}\left[\frac{x^T A^T A y}{\|x\|\|y\|} \leq \frac{m}{n}\left(\frac{x^T y}{\|x\|\|y\|} - \delta\right)\right] \\
&\quad = \mathbb{P}\left[\frac{x^T A^T A y}{\|x\|\|y\|} \leq \frac{1}{4}\left((1-\delta)\frac{m}{n}\left\|\frac{x}{\|x\|} + \frac{y}{\|y\|}\right\|^2 - (1+\delta)\frac{m}{n}\left\|\frac{x}{\|x\|} - \frac{y}{\|y\|}\right\|^2\right)\right] \\
&\quad < 2\exp\left(-\frac{m\delta^2}{8}\right)
\end{aligned} \tag{58}$$

Similarly,

$$\begin{aligned}
&\mathbb{P}\left[\frac{x^T A^T A y}{\|x\|\|y\|} \geq \frac{m}{n}\left(\frac{x^T y}{\|x\|\|y\|} + \delta\right)\right] \\
&\quad = \mathbb{P}\left[\frac{x^T A^T A y}{\|x\|\|y\|} \geq \frac{1}{4}\left((1+\delta)\frac{m}{n}\left\|\frac{x}{\|x\|} + \frac{y}{\|y\|}\right\|^2 - (1-\delta)\frac{m}{n}\left\|\frac{x}{\|x\|} - \frac{y}{\|y\|}\right\|^2\right)\right] \\
&\quad < 2\exp\left(-\frac{m\delta^2}{8}\right)
\end{aligned} \tag{59}$$

holds with probability no more than 58 and 59 complete the proof. □

Proof of Missing Data Here we again bound the quantities in Lemma 6, Equation 14, this time assuming A represents an entry-wise observation operation and assuming incoherence on the signals of interest. As we show below, in the proof of Theorem 11, we put together bounds given by Lemmas 13, 14 and 15, which are all proved in this section too, along with Lemma 16 for completeness. We start by proving the main result for missing data.

Proof of Theorem 11. Given the condition required by Theorem 11, we have $\sin\phi_d \leq \sqrt{d\mu_0/16n}$ and $\cos\phi_d \geq \sqrt{1 - d\mu_0/16n}$. This together with Lemma 16 and Lemma 15 yield $|\Delta| \leq \frac{\eta_3\sqrt{1+\frac{m}{16n}}}{\sqrt{1-d\mu_0/16n}}\frac{2d\mu_0}{n}\|v_\perp\|^2 \leq$

$\frac{2\eta_3\sqrt{1+\frac{1}{16}}}{\sqrt{1-\frac{1}{16}}}\frac{d\mu_0}{n}\|v_\perp\|^2 \leq \frac{11}{5}\eta_3\frac{d\mu_0}{n}\|v_\perp\|^2$. Also for β_2 in Lemma 15 we have $\beta_2 \leq \sqrt{2\mu(v_\perp)\log(1/\delta)} = \beta_1$. Therefore,

$$|\Delta| \leq \frac{11}{5}\frac{(1+\beta_1)^2}{1-\gamma_1}\frac{d\mu_0}{n}\|v_\perp\|^2 . \tag{60}$$

Letting $\eta_2 = \frac{(1+\beta_1)^2}{1-\gamma_1}\frac{d\mu_0}{m}$ and $\alpha_1 = \sqrt{\frac{2\mu(v_\perp)^2}{m}\log\left(\frac{1}{\delta}\right)}$, then applying this definition together with Lemma 16 to Lemma 14 Lemma 13 yields

$$\|p\|^2 \leq \left(1+\sqrt{\frac{2\eta_2}{1-\gamma_1}}\right)^2\|v\|^2 \tag{61}$$

$$\|r_\Omega\|^2 \geq (1-\alpha_1-2\eta_2)\frac{m}{n}\|v_\perp\|^2 \tag{62}$$

Now applying 60, 61 and 62 to 24 we obtain

$$\begin{aligned}\frac{\zeta_{t+1}}{\zeta_t} &\geq 1+\frac{(1-\alpha_1-2\eta_2)}{(1+\sqrt{2\eta_2/(1-\gamma_1)})^2}\frac{m}{n}\frac{\|v_\perp\|^2}{\|v\|^2} - \frac{22}{5}\frac{\eta_2}{(1+\sqrt{2\eta_2/(1-\gamma_1)})^2}\frac{m}{n}\frac{\|v_\perp\|^2}{\|v\|^2} \\ &\geq 1+\frac{(1-\alpha_1-\frac{32}{5}\eta_2)}{(1+\sqrt{2\eta_2/(1-\gamma_1)})^2}\frac{m}{n}\frac{\|v_\perp\|^2}{\|v\|^2}\end{aligned} \tag{63}$$

which holds with probability at least $1-3\delta$. The probability bound is obtained by taking the union bound of those generating Lemmas 13, 14 and 15, as we can see in the proofs of them in this Section, this union bound is at least $1-3\delta$.

Letting $\eta_1 = \frac{(1-\alpha_1-\frac{32}{5}\eta_2)}{(1+\sqrt{2\eta_2/(1-\gamma_1)})^2}$, then $\eta_1 > 0$ is equivalent to $1-\alpha_1-\frac{32}{5}\eta_2 > 0$, for which we have the following: if

$$m > \max\left\{\frac{128d\mu_0}{3}\log\left(\frac{2d}{\delta}\right), 32\mu(v_\perp)^2\log\left(\frac{1}{\delta}\right), 52d\mu_0\left(1+\sqrt{2\mu(v_\perp)\log\left(\frac{1}{\delta}\right)}\right)^2\right\} \tag{64}$$

then $\eta_1 > \frac{1}{4}$.

Under this condition, taking expectation with respect to v yields,

$$\mathbb{E}_v\left[\frac{\zeta_{t+1}}{\zeta_t}\middle|U\right] \geq 1+\frac{1}{4}\frac{m}{n}\mathbb{E}\left[\frac{\|v_\perp\|^2}{\|v\|^2}\middle|U\right] \geq 1+\frac{1}{4}\frac{m}{n}\frac{1-\zeta_t}{d} \tag{65}$$

where the last inequality follows from Lemma 1. Finally choosing δ to be $1/n^2$completes the proof. □

We then prove Corollary 12, the result that allows comparison between our convergence rate and that in [9].

Proof of Corollary 12. Let $X = [X_1, \ldots, X_d]$ with $X_i = \sin^2\phi_i$. Let $f(X) = 1-\sum_{i=1}^d X_i - \Pi_{i=1}^d(1-X_i)$, then $\frac{\partial f(X)}{\partial X_i} = -1+\Pi_{j\neq i}(1-X_j) \leq 0$. That is, $f(X)$ is a decreasing function of each component. Therefore, $f(X) \leq f(0) = 0$. It follows that

$$\zeta_t = \Pi_{i=1}^d(1-X_i) \geq 1-\sum_{i=1}^d X_i \geq 1-\frac{d\mu_0}{16n} \tag{66}$$

With a slight modification of Theorem 11 we obtain

$$\mathbb{E}\left[\kappa_{t+1}\middle|\kappa_t\right] \leq \left(1-\frac{1}{4}\frac{m}{n}\frac{\zeta_t}{d}\right)\kappa_t . \tag{67}$$

(66) and (67) together complete the proof. □

We now focus on proving the key lemmas for establishing Theorem 11, for which we need the following lemmas (the proofs can be found in [8]).

Lemma 30. *[8] Let $\delta > 0$. Suppose $m \geq \frac{8}{3} d\mu(U) \log(2d/\delta)$, then*

$$\mathbb{P}\left(\left\| \left(U_\Omega^T U_\Omega \right)^{-1} \right\| \leq \frac{n}{(1-\gamma_1)m} \right) \geq 1 - \delta$$

where $\gamma_1 = \sqrt{\frac{8d\mu(U)}{3m} \log(2d/\delta)}$.

Lemma 31 ([8], Lemma 1)**.** *Let $\alpha = \sqrt{\frac{2\mu(v_\perp)^2}{m} \log(1/\delta)}$, then*

$$\mathbb{P}\left(\|v_{\perp,\Omega}\|^2 \geq (1-\alpha)\frac{m}{n}\|v_\perp\|^2 \right) \geq 1 - \delta$$

Lemma 32 ([8], Lemma 2)**.** *Let $\mu(U), \mu(v_\perp)$ denote the incoherence parameters of $R(U)$ and $v_\perp$, and let $\delta \in (0,1)$ and $\beta_1 = \sqrt{2\mu(v_\perp)\log(1/\delta)}$, then*

$$\mathbb{P}\left(\left\| U_\Omega^T v_{\perp,\Omega} \right\|^2 \leq (\beta_1+1)^2 \frac{m}{n}\frac{d\mu(U)}{n}\|v_\perp\|^2 \right) \geq 1 - \delta$$

Now we are ready for the proof of Lemmas 13, 14 and 15.

Proof of Lemma 13. According to Lemmas 31, 32 and 30, we have

$$\begin{aligned} \|r_\Omega\|^2 &= \|v_{\perp,\Omega}\|^2 - v_{\perp,\Omega}^T U_\Omega \left(U_\Omega^T U_\Omega \right)^{-1} U_\Omega^T v_{\perp,\Omega} \\ &\geq \|v_{\perp,\Omega}\|^2 - \left\| \left(U_\Omega^T U_\Omega \right)^{-1} \right\| \|U_\Omega^T v_{\perp,\Omega}\|^2 \\ &\overset{\vartheta_1}{\geq} \left(1 - \alpha - \frac{(\beta_1+1)^2}{1-\gamma_1}\frac{d\mu(U)}{m} \right) \frac{m}{n}\|v_\perp\|^2 \end{aligned}$$

with probability at least $1 - 3\delta$. □

Proof of Lemma 14. Lemma 32 and Lemma 30 together give the following

$$\begin{aligned} \|Uw_2\|^2 = \left\| \left(U_\Omega^T U_\Omega \right)^{-1} U_\Omega^T v_{\perp,\Omega} \right\|^2 &\leq \left\| \left(U_\Omega^T U_\Omega \right)^{-1} \right\|^2 \left\| U_\Omega^T v_{\perp,\Omega} \right\|^2 \\ &\leq \frac{(\beta_1+1)^2}{(1-\gamma_1)^2}\frac{d\mu(U)}{m}\|v_\perp\|^2 \end{aligned}$$

holds with probability exceeding $1 - 2\delta$. Therefore,

$$\|p\|^2 \leq \left(\|v_\|\| + \|Uw_2\| \right)^2 \leq \left(1 + \frac{\beta_1+1}{1-\gamma_1}\sqrt{\frac{d\mu(U)}{m}} \right)^2 \|v\|^2$$

□

We also need the following lemma for the proof of Lemma 15, the proof of which is provided at the end of this section.

Lemma 33. *Let $\beta_2 = \sqrt{2\mu(v_\perp)\log\left(\frac{1}{\delta}\right)\frac{d\mu_0}{d\mu_0 + m\sin^2\phi_d}}$, where again μ_0 denoting the incoherence parameter of $R(\bar{U})$. Then*

$$\mathbb{P}\left(\left\| \bar{U}_\Omega^T v_{\perp,\Omega} \right\| \leq (1+\beta_2)\sqrt{\frac{m}{n}\frac{d\mu_0}{n}}\sqrt{\frac{m\sin^2\phi_d}{d\mu_0} + 1}\|v_\perp\| \right) \geq 1 - \delta$$

Proof of Lemma 15. Note that $|\Delta| = \|\Delta\|$, for which we have the following,

$$\begin{aligned}
\|\Delta\| &= \left\|w_2^T(\bar{U}^T U)^{-1}\bar{U}^T r\right\| \\
&= \left\|v_{\perp,\Omega}^T U_\Omega \left(U_\Omega^T U_\Omega\right)^{-1}\left(\bar{U}^T U\right)^{-1}\bar{U}_\Omega^T\left(I - \mathcal{P}_{U_\Omega}\right)v_{\perp,\Omega}\right\| \\
&\leq \left\|v_{\perp,\Omega}^T U_\Omega\right\|\left\|\left(U_\Omega^T U_\Omega\right)^{-1}\right\|\left\|\left(\bar{U}^T U\right)^{-1}\right\|\left\|\bar{U}_\Omega^T\left(I - \mathcal{P}_{U_\Omega}\right)v_{\perp,\Omega}\right\| \\
&\overset{\vartheta_1}{\leq} \frac{1}{\cos\phi_d}\left\|v_{\perp,\Omega}^T U_\Omega\right\|\left\|\left(U_\Omega^T U_\Omega\right)^{-1}\right\|\left\|\bar{U}_\Omega^T v_{\perp,\Omega}\right\| \\
&\leq \frac{1}{\cos\phi_d}(\beta_1+1)\sqrt{\frac{m}{n}\frac{d\mu(U)}{n}}(1+\beta_2)\sqrt{\frac{m}{n}\frac{d\mu_0}{n}}\sqrt{\frac{m\sin^2\phi_d}{d\mu_0}+1}\frac{n}{m(1-\gamma_1)}\|v_\perp\|^2 \\
&\overset{\vartheta_2}{\leq} \frac{(1+\beta_1)(1+\beta_2)}{(1-\gamma_1)\cos\phi_d}\sqrt{\frac{m\sin^2\phi_d}{d\mu_0}+1}\sqrt{\frac{d\mu_0}{n}}\sqrt{\frac{d\mu(U)}{n}}\|v_\perp\|^2
\end{aligned}$$

where ϑ_1 holds since from the following:

$$\left\|\bar{U}_\Omega^T\left(I - \mathcal{P}_{U_\Omega}\right)v_{\perp,\Omega}\right\| \leq \left\|\bar{U}_\Omega^T v_{\perp,\Omega}\right\| \ , \qquad \left\|\left(U_\Omega^T U_\Omega\right)^{-1}\right\| \leq \frac{1}{\cos\phi_d}$$

and ϑ_2 follows by putting Lemmas 32, 30 and 33 together. □

We also prove Lemma 16 for completeness. Before that we first call out the following lemma, the proof of which can be found in [9].

Lemma 34. *[9] There exists an orthogonal matrix $V \in \mathbb{R}^{d\times d}$ such that*

$$\sum_{k=1}^{d}\sin^2\phi_k \leq \left\|\bar{U}V - U\right\|_F^2 \leq 2\sum_{k=1}^{d}\sin^2\phi_k$$

Proof of Lemma 16. According to Lemma 34 we have

$$\begin{aligned}
\|U_i\|_2 \leq \left\|\bar{U}_i\right\|_2 + \left\|\bar{U}_i V - U_i\right\|_2 &\leq \left\|\bar{U}_i\right\| + \sqrt{2\sum_{k=1}^{d}\sin^2\phi_k} \\
&\leq \left(1 + \frac{1}{2\sqrt{2}}\right)\sqrt{\frac{d\mu_0}{n}}
\end{aligned}$$

It hence follows that $\|U_i\|_2^2 \leq 2\frac{d\mu_0}{n}$. □

We need the following lemma and McDiarmid's inequality to prove Lemma 35.

Lemma 35. $\left\|\bar{U}^T v_\perp\right\|^2 \leq \sin^2(\phi_d)\|v_\perp\|^2$*, where ϕ_d denotes the largest principal angle between $R(\bar{U})$ and $R(U)$.*

Proof. According to the definition of $v_\perp$ and Lemma 17, we have

$$\begin{aligned}
\left\|\bar{U}^T y\right\|^2 = \left\|\bar{U}^T\left(\mathbb{I} - UU^T\right)\bar{U}s\right\|^2 &= s^T\bar{Y}\Sigma^4\bar{Y}^4 s \\
&\overset{\vartheta_3}{\leq} \sin^2\phi_d s^T\bar{Y}\Sigma^2\bar{Y}^T s = \sin^2\phi_d\|v_\perp\|^2
\end{aligned}$$

here $\bar{Y}$ and Σ are the same as those defined in Lemma 17, and the last equality holds since $\|v_\perp\|^2 = \|s\|^2 - v^T UU^T v = s^T\bar{Y}\Sigma^2\bar{Y}^T s$. □

Theorem 36. *(McDiarmid's Inequality [23]). Let $X_1, \ldots, X_n$ be independent random variables, and assume f is a function for which there exist t_i, $i = 1, \ldots, n$ satisfying*

$$\sup_{x_1,\ldots,x_n,\widehat{x}_i} |f(x_1, \ldots, x_n) - f(x_1, \ldots, \widehat{x}_i, \ldots, x_n)| \leq t_i$$

where $\widehat{x}_i$ indicates replacing the sample value x_i with any other of its possible values. Call $f(X_1, \ldots, X_n) := Y$. Then for any $\epsilon > 0$,

$$\mathbb{P}\left[Y \geq \mathbb{E}Y + \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n t_i^2}\right)$$
$$\mathbb{P}\left[Y \leq \mathbb{E}Y - \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n t_i^2}\right)$$

Proof of Lemma 33. We use McDiarmid's inequality to prove this. For the simplicity of notation denote $v_\perp$ as y. Let $X_i = \bar{U}_{\Omega(i)} y_{\Omega(i)} \in \mathbb{R}^d$, and $f(X_1, \ldots, X_m) = \|\sum_{i=1}^m X_i\|_2 = \left\|\bar{U}_\Omega^T v_{\perp,\Omega}\right\|_2$, then $|f(x_1, \ldots, x_n) - f(x_1, \ldots, \widehat{x}_i, \ldots x_n|$ can be bounded via

$$\left|\left\|\sum_{i=1}^m X_i\right\|_2 - \left\|\sum_{i\neq k}^m X_i + \widehat{X}_k\right\|_2\right| \leq \left\|X_k - \widehat{X}_k\right\|_2 \leq \|X_k\|_2 + \|\widehat{X}_k\|_2$$
$$\leq 2\|y\|_\infty \sqrt{d\mu_0/n} \tag{68}$$

We next calculate $\mathbb{E}\left[f(X_1, \ldots, X_m)\right] = \mathbb{E}\left[\|\sum_{i=1}^m X_i\|_2\right]$. Note that

$$\mathbb{E}\left[\left\|\sum_{i=1}^m X_i\right\|^2\right] = \mathbb{E}\left[\sum_{i=1}^m \|X_i\|^2 + \sum_{i=1}^m \sum_{j\neq i} X_i^T X_j\right] \tag{69}$$

Recall that we assume the samples are taken uniformly with replacement. This together with the fact that $\left\|\bar{U}_i\right\|_2 = \|\mathcal{P}_{R(\bar{U})}(e_i)\| \leq \sqrt{d\mu_0/n}$ yield the following

$$\mathbb{E}\left[\sum_{i=1}^m \|X_i\|^2\right] = \sum_{i=1}^m \mathbb{E}\left[\left\|U_{\Omega(i)} y_{\Omega_{(i)}}\right\|^2\right]$$
$$= \sum_{i=1}^m \sum_{k=1}^n \|\bar{U}_k\|^2 y_k^2 \mathbb{P}_{\{\Omega(i)=k\}} \leq \frac{m}{n}\frac{d\mu_0}{n}\|y\|^2 \tag{70}$$

$$\mathbb{E}\left[\sum_{i=1}^m \sum_{j\neq i} X_i^T X_j\right] = \sum_{i=1}^m \sum_{j\neq i} \sum_{k_1=1}^n \sum_{k_2=1}^n y_{k_1} \bar{U}_{k_1}^T \bar{U}_{k_2} y_{k_2} \mathbb{P}(\Omega_j = k_2)\mathbb{P}(\Omega_i = k_1)$$
$$= \frac{m^2 - m}{n^2}\|\bar{U}^T y\|^2 \leq \frac{m^2}{n^2}\sin^2\phi_d \|y\|^2 \tag{71}$$

where the last inequality holds by Lemma 35.

Eqs (69) (70) and (71) together with the Jensen's inequality imply

$$\mathbb{E}\left[\left\|\sum_{i=1}^m X_i\right\|\right] \leq \sqrt{\frac{m}{n}}\sqrt{\frac{m}{n}\sin^2\phi_d + \frac{d\mu_0}{n}}\|y\| = \sqrt{\frac{m}{n}\frac{d\mu_0}{n}}\sqrt{\frac{m\sin^2\phi_d}{d\mu_0} + 1}\|y\| \tag{72}$$

Let $\epsilon = \beta_2 \sqrt{\frac{m}{n}\frac{d\mu_0}{n}} \sqrt{\frac{m \sin^2 \phi_d}{d\mu_0} + 1} \|y\|$, then (68) and (72) together with Theorem 36 give

$$\begin{aligned}
\mathbb{P}\left[\|U_\Omega y_\Omega\| \geq (1+\beta_2) \sqrt{\frac{m}{n}\frac{d\mu_0}{n}} \sqrt{\frac{m \sin^2 \phi_d}{d\mu_0} + 1} \|y\| \right] & \\
&\leq \exp\left(\frac{-2\beta_2^2 \frac{m}{n}\frac{d\mu_0}{n} \left(\frac{m \sin^2 \phi_d}{d\mu_0} + 1 \right) \|y\|^2}{4m \|y\|_\infty^2 \frac{d\mu_0}{n}} \right) \\
&= \exp\left(\frac{-\beta_2^2 \left(\frac{m \sin^2 \phi_d}{d\mu_0} + 1 \right) \|y\|^2}{2n \|y\|_\infty^2} \right) = \delta
\end{aligned} \tag{73}$$

where the last inequality follows by submitting our definition of $\mu(y)$ Eq (23) and β_2. □

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Diego Armentano, Carlos Beltrán, and Michael Shub. Average polynomial time for eigenvector computations. *arXiv preprint arXiv:1410.2179*, 2014.

[3] Raman Arora, Andy Cotter, and Nati Srebro. Stochastic optimization of PCA with capped MSG. In *Advances in Neural Information Processing Systems*, pages 1815–1823, 2013.

[4] Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental PCA. In *Advances in Neural Information Processing Systems*, pages 3174–3182, 2013.

[5] Laura Balzano. *Handling missing data in high-dimensional subspace modeling*. PhD thesis, University of Wisconsin – Madison, 2012.

[6] Laura Balzano. On the equivalence of Oja's algorithm and GROUSE. In *Proceedings of AIStats*, 2022.

[7] Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *48th Annual Allerton Conference on Communication, Control, and Computing*, pages 704–711. IEEE, 2010.

[8] Laura Balzano, Benjamin Recht, and Robert Nowak. High-dimensional matched subspace detection when data are missing. In *2010 IEEE International Symposium on Information Theory*, pages 1638–1642. IEEE, 2010.

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

[10] Dimitri P Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38):3, 2011.

[11] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582. PMLR, 2016.

[12] J Paul Brooks, JH Dulá, and Edward L Boone. A pure l1-norm principal component analysis. *Computational statistics & data analysis*, 61:83–98, 2013.

[13] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[14] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[15] Alexandre d'Aspremont, Francis Bach, and Laurent El Ghaoui. Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294, 2008.

[16] Christopher D De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2332–2341, 2015.

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

[18] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU Press, 4 edition, 2012.

[19] Jun He, Laura Balzano, and Arthur Szlam. Incremental gradient on the grassmannian for online foreground and background separation in subsampled video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1568–1575. IEEE, 2012.

[20] Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near optimal finite sample guarantees for oja's algorithm. In *29th Annual Conference on Learning Theory*, pages 1147–1164, 2016.

[21] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[22] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.

[23] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

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

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

[26] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

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

[28] Gilbert W Stewart and Ji-guang Sun. *Matrix perturbation theory*. Academic press, 1990.

[29] Stanislaw J Szarek. Metric entropy of homogeneous spaces. *arXiv preprint math/9701213*, 1997.

[30] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[31] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010.

[32] Bin Yang. Projection approximation subspace tracking. *IEEE Transactions on Signal processing*, 43(1):95–107, 1995.

[33] Dejiao Zhang. *Extracting Compact Knowledge From Massive Data*. PhD thesis, University of Michigan, Ann Arbor, 2019.

[34] Dejiao Zhang and Laura Balzano. Global convergence of a grassmannian gradient descent algorithm for subspace estimation. In *AISTATS*, pages 1460–1468, 2016.

[35] Qinqing Zheng and John Lafferty. 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, 2015.